Nicholas J. Daras
Michael Th. Rassias *Editors*

Computation, Cryptography, and Network Security

Springer

Computation, Cryptography, and Network Security

Nicholas J. Daras • Michael Th. Rassias
Editors

Computation, Cryptography, and Network Security

Springer

Editors
Nicholas J. Daras
Department of Mathematics
and Engineering
Hellenic Military Academy
Vari Attikis, Greece

Michael Th. Rassias
Department of Mathematics
ETH Zürich
Zürich, Switzerland

ISBN 978-3-319-18274-2 ISBN 978-3-319-18275-9 (eBook)
DOI 10.1007/978-3-319-18275-9

Library of Congress Control Number: 2015945103

Mathematics Subject Classification (2010): 03D25, 11U05, 26D15, 31A10, 45P05, 47G10, 47A07, 44A10, 46E30, 68R10, 94C30

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This book entitled *Computation, Cryptography, and Network Security* brings together a broad variety of mathematical methods and theories with several applications from a number of disciplines. It discusses new directions for further inventions in computation, cryptography, and network security.

It is hoped to provide some good understanding of the subject of security in the broadest sense. It consists of papers written by eminent scientists from the international mathematical community, who present important research works in several theories and problems. These contributions focus on both old and new developments of pure and applied mathematics with emphasis to the geometry of the zeros of a polynomial, multivariate Birkhoff interpolation, variational principles in vector spaces, parameterized Yang-Hilbert-type integral inequalities and their operator expressions, operators preserving linear functions, integral estimates for the composition of Green's and bounded operators, asymptotic behavior of orthogonal polynomials on the unit circle, generalized Laplace transform inequalities in multiple weighted Orlicz spaces, and functional equations.

Furthermore, some survey papers are published in this volume, which are particularly useful for a broader audience of readers, particularly in credential technologies, cryptographic schemes, current challenges for IT security with focus on biometry, flaws in the initialization process of stream ciphers, entropy and information measures, information theory, quantum analogues of Hermite-Hadamard type inequalities for generalized convexity, producing fuzzy inclusion and entropy measures, as well as applications on the unstable equilibrium points and system separations in electric power systems, and a supply chain game theory for cybersecurity investments subject to network vulnerability.

We would like to express our deepest thanks to all the contributors of papers who, through their works, participated in this book. We would also wish to acknowledge the superb assistance that the staff of Springer has provided for the publication of this book.

Athens, Greece — Nicholas J. Daras
Princeton, NJ, USA — Michael Th. Rassias

Contents

Transformations of Cryptographic Schemes Through Interpolation Techniques

Stamatios-Aggelos N. Alexandropoulos, Gerasimos C. Meletiou, Dimitrios S. Triantafyllou, and Michael N. Vrahatis

Abstract The problem of transforming cryptographic schemes using interpolation techniques is studied. Firstly, explicit forms for the discrete logarithm and the Diffie–Hellman cryptographic functions are given. Subsequently, the inverse Aitken and Neville interpolation methods for the discrete logarithm and the Lucas logarithm problems are presented. Next, the representation of cryptographic functions through polynomials or algebraic functions as well as a special case of discrete logarithm problem is given. Finally, a study of cryptographic functions using factorization of matrices is analyzed.

Keywords: Public key cryptography • Discrete logarithm • Diffie Hellman mapping • Polynomial interpolation techniques • Matrix factorization

1 Introduction

A basic task of cryptography is the transformation or encryption, of a given message into another one which appears meaningful only to the intended recipient after the process of decryption. Messages and cryptograms are represented as elements of finite algebraic structures. Encryption and decryption processes are functions over finite structures especially over finite fields.

It is well known that, in a finite field GF(q), where q is a prime power, every function can be represented as a polynomial through the Lagrangian interpolation.

S.-A.N. Alexandropoulos (✉) • M.N. Vrahatis
Computational Intelligence Laboratory (CILab), Department of Mathematics, University of Patras, GR-26110 Patras, Greece
e-mail: alekst@master.math.upatras.gr; vrahatis@math.upatras.gr

G.C. Meletiou
A.T.E.I. of Epirus, P.O. 110, GR-47100 Arta, Greece
e-mail: gmelet@teiep.gr

D.S. Triantafyllou
Hellenic Army Academy (SSE), University of Military Education, GR-16673 Vari, Attica, Greece
e-mail: dtriant@sse.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_1

Also, for every function, $f : \mathrm{GF}(q) \to \mathrm{GF}(q)$, there exists a unique polynomial $p(x)$ of degree at most $(q-1)$ that coincides with f.

One of the most basic aspects of the numerical analysis, with diverse applications in the field of cryptography, is the interpolation techniques. It is worth noting that the past three decades have witnessed an increasing interest in the application of interpolation techniques of cryptographic functions. The Langrange's, Hermite's, Aitken's and Neville's interpolation methods are widely used for the interpolation process through which the encryption and decryption functions are approximated.

Interpolation is computationally attractive only in the case of a polynomial with small number of nonzero coefficients. Since encryption and decryption functions are defined as functions over finite fields, it is of great importance to attempt to express them as polynomials and perform cryptanalysis by polynomial computation.

In the work at hand we study the problem of transforming cryptographic schemes using interpolation techniques. In the second section we consider explicit forms of cryptographic functions, such as the discrete logarithm and the Diffie–Hellman functions. Subsequently, in the third section we present inverse interpolation methods, such as Aitken's and Neville's methods for the well-known discrete logarithm problem as well as the Lucas logarithm problem. Next, in the fourth section we present the representation of cryptographic functions through polynomials or algebraic functions, while in the fifth section we give a special case of discrete logarithm problem. Finally the chapter ends at the sixth section with a study of cryptographic functions using factorization of matrices.

2 Explicit Forms of Cryptographic Functions

Definition 1. Consider the case of a prime field $\mathbb{Z}_p$, where p is a prime. For a generator g of $\mathbb{Z}_p^*$, $\langle g \rangle = \mathbb{Z}_p^*$, the polynomial:

$$p(x) = \sum_{i=1}^{p-2} \frac{x^i}{1-g^i},$$

represents the ***discrete logarithm*** of x to the basis g, $\forall\ x \in \mathbb{Z}_p^*$.

Remark 1. Surprisingly enough the formulas of the coefficients are very simple [24].

Proposition 1 ([17]). *Using the discrete Fourier transform, we can also derive the following matrix representation:*

$$\log_g(x) = -(1\ 2\ \ldots\ p-1)\left(g^{-ij}\right)\begin{pmatrix} x \\ x^2 \\ \vdots \\ x^{p-1} \end{pmatrix},$$

where $(-g^{-ij})$, $1 \leqslant i,j \leqslant p-1$ *is an* $(p-1)\times(p-1)$ *matrix.*

It seems natural to generalize these results to logarithms where the base is not necessarily a primitive element in a field of prime power order. To this end, we recall the following result [16–20]:

Theorem 1. *Let $g \in \mathbb{F}_{p^n}^*$, g generator of the multiplicative group of the field, that is $\langle g \rangle = \mathbb{F}_{p^n}^*$, $g^z = x \in \mathbb{F}_{p^n}^*, 1 \leqslant z \leqslant p^n - 1$. Suppose that the numeral system with p as a basis is used:*

$$z = \sum_{s=0}^{n-1} d_s p^s, \qquad 0 \leqslant d_s \leqslant m.$$

Then, it holds that:

$$d_s = \sum_{i=1}^{p^n-2} \frac{x^i}{(1 - g^i)^{p^s}} \ .$$

Concerning the representations of the Diffie–Hellman key function Meidl and Winterhof in [15] gave the following result:

Theorem 2. *Assume that $g \in \mathbb{F}_{p^n}^*$, $|\langle g \rangle| = m$, m divides $p^n - 1$ and $1 \leqslant a, b \leqslant m$. Then the polynomial:*

$$f(x, y) = m^{-1} \sum_{i,j=1}^{m} g^{i \cdot j} x^i y^j,$$

satisfies the relation:

$$f\left(g^a, g^b\right) = g^{ab}.$$

Proposition 2. *Using the discrete Fourier transform, we can also derive the following matrix representation:*

$$f(x, y) = m^{-1} (y \ y^2 \ \ldots \ y^m) \left(g^{-ij}\right) \begin{pmatrix} x \\ x^2 \\ \vdots \\ x^m \end{pmatrix},$$

where (g^{-ij}) is an $m \times m$ matrix, $1 \leqslant i, j \leqslant m$.

3 Interpolation and Inverse Interpolation Methods

Aitken's and Neville's interpolation techniques, as well as the Lagrange interpolation method, are well known and they are considered as the state of the art for transforming of cryptographic functions over finite fields. In contrast to the Lagrange method, Aitken's and Neville's methods are constructive in a way that permits the addition of a new interpolation point directly and with low computational cost. Thus, the interpolation procedure is initially applied to a small number of points and unless the required polynomial is found, new interpolation points are added sequentially to the previously obtained polynomial with low cost. This advantage over the Lagrange interpolation method and the fact that Aitken's and Neville's interpolation formulae can be applied in any field, have motivated the investigation of their performance over finite fields. In this section, we study the inverse Aitken and the inverse Neville interpolation methods over finite fields for the discrete logarithm and the Lucas logarithm function.

3.1 The Aitken and Neville Interpolation and Inverse Interpolation Methods

We study the Aitken and Neville interpolation methods by considering a function $f(x)$ defined on a field $\mathbb{F}$ and $x_i \in \mathbb{F}$ be mutually different interpolation points. Also, we assume that $f_i = f(x_i)$, with $i = 0, 1, \ldots, n$. Then, the ***Aitken polynomial*** is defined as follows:

$$P_{0,1,\ldots,m,i}(x) = \frac{1}{(x_i - x_m)} \begin{vmatrix} P_{0,1,\ldots,m}(x) & x_m - x \\ P_{0,1,\ldots,m-1,i}(x) & x_i - x \end{vmatrix},$$

where $m = 0, 1, \ldots, n-1$, $i = m+1, \ldots, n$ and $x_0, x_1, \ldots, x_k$ are the interpolated points.

Similarly, the ***Neville interpolation*** formula is given by:

$$P_{i,1+i,\ldots,i+m}(x) = \frac{1}{(x_{i+m} - x_i)} \begin{vmatrix} P_{i,i+1,\ldots,i+m-1}(x) & x_i - x \\ P_{i+1,i+2,\ldots,i+m}(x) & x_{i+m} - x \end{vmatrix},$$

where $m = 1, 2, \ldots, n$, $i = 0, 1, \ldots, n-m$ and where $x_i, x_{i+1}, \ldots, x_{i+k}$ are the interpolated points.

The inverse interpolation problem [12] can be approached through Aitken's and Neville's interpolation techniques using the corresponding formulae [3]. Specifically, the corresponding formulae of the ***inverse Aitken interpolation method*** and the ***inverse Neville interpolation method*** are given as follows:

$$P_{0,1,\ldots,m,i}(y) = \frac{1}{(y_i - y_m)} \begin{vmatrix} P_{0,1,\ldots,m}(y) & y_m - y \\ P_{0,1,\ldots,m-1,i}(x) & y_i - y \end{vmatrix},$$

where $m = 0, 1, \ldots, n-1$, $i = m+1, \ldots, n$ and:

$$P_{i,1+i,\ldots,i+m}(y) = \frac{1}{(y_{i+m} - y_i)} \begin{vmatrix} P_{i,i+1,\ldots,i+m-1}(y) & y_i - y \\ P_{i+1,i+2,\ldots,i+m}(y) & y_{i+m} - y \end{vmatrix}.$$

An interesting point is the approach on the values of the ***shifted exponential function***:

$$f(x) = \alpha^x - b \,(\mathrm{mod}\, p), \quad \text{for } p \text{ prime and } \alpha \in \mathbb{Z}_p,$$

using the inverse Aitken and the inverse Neville interpolations method. Selected points of the function f are used to construct a polynomial that interpolates the value $f(x^*) = 0 \,(\mathrm{mod}\, p)$. The resulting polynomial is evaluated at zero by interpolating two random values of x in the beginning. Every new point becomes a new interpolate point, unless the value is the discrete logarithm of b over $\alpha \,(\mathrm{mod}\, p)$.

As it has been presented in [12] the computational cost for tackling the problem of discrete logarithm through both methods is high. Overall, Aitken's method proved slightly better than the Neville's method. The performance of two methods implies that the resulting polynomials were most often of low degree and in most cases there exists a low degree polynomial that interpolates the discrete logarithm.

3.2 Inverse Interpolation Methods for the Lucas Logarithm Problem

The Lucas function is a one-way function used in public key cryptography. The security of cryptosystems based on the Lucas function relies on the difficulty of solving the Lucas logarithm problem. In this subsection the Lucas logarithm problem is studied using the inverse Aitken and Neville interpolation methods. These methods are applied to values of the Lucas function to obtain a polynomial that interpolates the Lucas logarithm.

Definition 2. Suppose that p is an odd prime and let $\mathbb{F}_p$ be the finite field of order p. For a fixed element $m \in \mathbb{F}_p$ consider the following second-order linear recurrence relation:

$$\begin{cases} V_0(m) = 2, \\ V_1(m) = m, \\ V_t(m) = mV_{t-1}(m) - V_{t-2}(m), \quad t \geqslant 2. \end{cases}$$

Then the sequence $\{V_t(m)\}_{t=0}^{\infty}$ is called ***Lucas sequence*** generated by m and the mapping:

$$t \mapsto V_t(m), \quad t \geqslant 0,$$

is called ***Lucas function***. Furthermore, given a prime p any $m \in \mathbb{F}_p$ and $z \in \{V_t(m)\}$ then, the integer x which satisfies the relation $V_x(m) = z$ is called the ***Lucas logarithm*** of z.

Remark 2. The security of cryptosystems based on the Lucas function relies on the difficulty of addressing the Lucas logarithm problem.

Remark 3. It was shown in [20] that $V_t(m) = \mu^t + \mu^{-t}$, $t \geqslant 0$, where μ and μ^{-1} are the roots of the characteristic polynomial of the above second-order linear recurrence relation.

Remark 4. The roots of the following equation:

$$f(X) = X^2 - mX + 1,$$

are given by the expressions:

$$\mu = \frac{m + \sqrt{m^2 - 4}}{2}, \qquad \mu^{-1} = \frac{m - \sqrt{m^2 - 4}}{2},$$

and if $m^2 - 4$ is zero or a quadratic residue modulo p, then both μ and $-\mu$ are in $\mathbb{F}_p$, otherwise they are in the extension field $\mathbb{F}_{p^2}$.

Let us study the inverse Aitken and the inverse Neville interpolation methods over the ***shifted Lucas function***:

$$f(t) = V_t(m) - z, \quad t \geqslant 0,$$

with $z \in \mathbb{F}_p$, which is not a bijection. Specifically, a polynomial that interpolates the function value $f(t^*) = 0$ is required. Both methods are constructive, thus the interpolation procedure begins by interpolating two function values of the function $f(t)$ for two random values of t. The resulting polynomial is evaluated at zero and the obtained value t_0 is verified by computing $f(t_0)$. If $f(t_0) = 0$, then t_0 is the Lucas logarithm to the base m and the procedure is terminated, otherwise the value $f(t_0)$ becomes a new interpolation point.

As it has been presented in [13] through several experiments, both Aitken's and Neville's methods have similar behavior in finding the polynomial that interpolates the Lucas logarithm value and require about one third of the field cardinality for verifications to obtain the polynomial, which is not small.

In comparison with the results for the discrete logarithm problem [12], in the case of Lucas logarithm problem the number of verifications required to find the proper polynomial is smaller than the corresponding one for the discrete logarithm

problem. Concerning the polynomial degree, the degrees of the polynomials that interpolate the discrete logarithm value are higher [12] than that of the polynomials that interpolate the Lucas logarithm value.

4 Interpolation of Cryptographic Functions for a Given Set of Data

Another approach is to represent the cryptographic functions with polynomials or algebraic functions coinciding with the functions over proper subsets of the domain. However it has been shown that polynomials approximating cryptographic transformations on sufficiently large sets must be of sufficiently large degree and sparsity. To this end, lower bounds on the degrees and the sparsity (i.e., the number of the nonzero coefficients) of polynomials interpolating the cryptographic functions can be obtained.

It has been shown that even for polynomial representations of the discrete logarithm over quite thin sets, the degree is still required to be high. These results support the assumption of hardness of the aforementioned functions if the parameters are properly chosen. The term "approximation" has been used for polynomials which coincide with the cryptographic function over a subset of its domain.

Concerning the discrete logarithm we have the result given by Coppersmith and Shparlinski [7] and Shparlinski [21]:

Theorem 3. *Let p be a prime, $g \in \mathbb{Z}_p^*$. Consider the subset $S \subset \{1, 2, \ldots, p-1\}$, $|S| = p - 1 - s$, $F(X) \in \mathbb{Z}_p[X]$ a polynomial satisfying $F(g^x) = x$, $\forall\, x \in S$. Then it holds that:*

$$\deg(F) \geqslant p - 2 - 2s \qquad \textbf{(lower bound)}.$$

Similar results can be derived for the Diffie–Hellman mapping:

Theorem 4. *Let q be a prime power, $g \in \mathbb{F}_q^*$. Consider the subset $A \subset [N+1, N+h] \times [N+1, N+h]$, where $2 \leqslant h \leqslant q-1$ and $|A| \geqslant 10h^{8/5}$. Assume that $F(U, V) \in \mathbb{F}_q[X, Y]$ satisfies $F(g^x, g^y) = g^{xy}$ for all $(x, y) \in A$. Then it holds that:*

$$\deg(F) \geqslant \frac{|A|^2}{128h^3} \qquad \textbf{(lower bound)}.$$

El Mahassni and Shparlinski in [10] gave for the decision Diffie–Hellman key problem the following result:

Theorem 5. *Let q be a prime power, $g \in \mathbb{F}_q^* = \langle g \rangle$. Consider the subset $A \subset [N+1, N+h] \times [N+1, N+h]$, where $2 \leqslant h \leqslant q-1$. The three variable polynomial $F(U, V, T) \in \mathbb{F}_q[X, Y, Z]$ satisfies $F(g^x, g^y, g^{xy}) = 0$ for all $(x, y) \in A$. Then it holds that:*

$$\deg(F) \geqslant \frac{|A|}{3h^{8/5}} \qquad \textbf{(lower bound)}.$$

Furthermore, lower bounds have been computed for functions related to the integer factoring problem and the RSA cryptosystem [1] as well as the Lucas logarithm [2].

5 The Double Discrete Logarithm and the Root of the Discrete Logarithm

Definition 3. Let G be a cyclic group of order t, $|\langle g \rangle| = |G| = t$ and $h \in \mathbb{Z}_t^*$ be an element of order $|\langle h \rangle| = m$. The ***double discrete logarithm*** of an element $z = g^{h^x} \in G$ to the bases g and h is the unique $x : 0 \leqslant x < m$.

Remark 5. The parameters G, t, g, and h should be chosen such that computing discrete logarithms in G to the base g and in $\mathbb{Z}_t^*$ to the base h are infeasible.

Remark 6. The double discrete logarithm is used as one-way function in several cryptographic schemes, in particular in group signature schemes and publicity verifiable secret sharing schemes.

The verifiable encryption of discrete logarithms is a typical example. Specifically we have [22]:

1. Assume that $|\langle g \rangle| = |G| = p$, p is prime, $p = 2q + 1$, $h \in \mathbb{Z}_{p'}^*$, $|\langle h \rangle| = q$, q prime.
2. A private key $z \in \mathbb{Z}_q$ is randomly chosen and the public-key $y \equiv h^z (\mathrm{mod}\, p)$ is published.
3. A message v is encrypted as (A, B), $A \equiv h^\alpha (\mathrm{mod}\, p)$ and $B \equiv v^{-1} y^\alpha (\mathrm{mod} p)$ (El Gamal's public key cryptosystem [9]).
4. The element $w = g^v$ becomes public.
5. Verifying that a pair (A, B) encrypts the discrete logarithm of a public element $w = g^v$ of the group G is equivalent to verifying that the discrete logarithm of A to the base h is identical to the double discrete logarithm of w^B to the bases g and y.

Definition 4. Let G be a cyclic group of order t, $|\langle g \rangle| = |G| = t$, $Y \in G$ be an element of the group G. A kth ***root of the discrete logarithm*** of Y to the base g is an integer satisfying $x : 0 \leqslant x < t$ satisfying $Y = g^{x^k}$ if such an x exists.

Remark 7. Existence and uniqueness of the kth root of the discrete logarithm are not guaranteed. In the case $\left|\{x : g^{x^k} = y\}\right| \geqslant 2$ branches of the kth root of the discrete logarithm are defined.

Remark 8. Group G and parameters g and t can be chosen in such a way that computing discrete logarithms to the base g is infeasible. Also, it can be chosen such that obtaining kth roots modulo t is hard.

Remark 9. The kth root of the discrete logarithm is used as one-way function [4, 5, 14] in group signature schemes, publicly verifiable secret sharing schemes, electronic cash, offline electronic cash systems, anonymity control in multi-bank e-cash system, in history-based signatures, etc.

The following proposition gives an insight for the lower bounds of the polynomial representation of the double discrete logarithm:

Proposition 3. *Let $t \geqslant 3$ be an integer, p be a prime, $p \equiv 1 (\mathrm{mod}\ t)$, $g \in \mathbb{F}_p^*$ an element of order $m \geqslant 2$, $S \subseteq \{0, 1, \ldots, m-1\}$ a set of order $|S| = m - s$ and $f(x) \in \mathbb{F}_p[x]$ a polynomial satisfying the following relation:*

$$f(g^{h^n}) = n, \quad \forall\ n \in S.$$

Then it holds that:

$$\deg(f) \geqslant \frac{m - 2s}{2v} \qquad \textbf{(lower bound)},$$

where v is the smallest integer in the set $\{h^n (\mathrm{mod}\ t) : 1 \leqslant n \leqslant m\}$.

Similar results can be obtained in the case of the multiplicative group of fields of prime power order and groups derived from elliptic curves. Lower bounds can also be computed for the degree of the polynomial which represents the root of the discrete Logarithm:

Proposition 4. *Let p be a prime number, $g \in \mathbb{Z}_p^*$, $|\langle g \rangle| = t$ and let $k \geqslant 1$ be an integer s.t. $\gcd(k, \phi(t)) = 1$. Let $S \subset \mathbb{Z}_t^*$ be a subset of order $|S| = \phi(t) - s$. We assume the existence of a polynomial $F(X) \in \mathbb{Z}_p[X]$ s.t. $F(g^{x^k}) = x$, $\forall\, x \in S$. Then it holds that:*

$$\deg(F) \geqslant \frac{\phi(t) - 2s}{2} \qquad \textbf{(lower bound)}.$$

Remark 10. The exponent k is odd and relatively prime to $\phi(t)$ and the kth root function becomes a bijection.

Remark 11. The main motivation stems from RSA. In this case k is the encryption exponent e. In some applications the message m is encrypted as $c \equiv m^e (\mathrm{mod}\ \ N)$ and g^{m^e} becomes public. Recovering m from g^{m^e}, or verifying properties of m is the problem. For proofs of knowledge of roots of discrete logarithms, we refer the interested reader to [4].

6 Matrix Factorization in Cryptography

Before we proceed to methods for the matrix representation of cryptographic functions, we give some necessary definitions and theorems.

Definition 5. An $m \times n$ matrix whose row-entries are terms of a geometric progression is called ***Vandermonde matrix*** and has the following expression:

$$V = \begin{pmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_{m-1} & a_{m-1}^2 & \cdots & a_{m-1}^{n-1} \\ 1 & a_m & a_m^2 & \cdots & a_m^{n-1} \end{pmatrix}.$$

In order to extract useful pieces of information for a matrix, including the rank, the eigenvalues and eigenvectors as well as the determinant among others, its factorization can be used. In matrices with real or complex entries, the use of orthogonal transformations such as Householder's transformations for computing the QR factorization or the singular value decomposition (SVD) [8] improves the stability of the algorithms increasing simultaneously the floating point operations. Non-orthogonal techniques such as LU factorization with partial or complete pivoting [8] reduce the required computational complexity giving a higher, but acceptable bound, for the norm of the error.

In the case of finite fields there is no error, thus the use of non-orthogonal methods which are faster is more suitable. Since in cryptography the required storage capacity of a method should not be greater than that of the initial data, the QR factorization is not preferable. The LU factorization does not require extra storage capacity and has less computational complexity.

Below we present the LU factorization with/without partial/complete pivoting of a matrix.

Theorem 6 (LU Factorization without Pivoting [8]). *Let A be an $m \times n$ matrix. Then there are a lower triangular $m \times m$ matrix L with ones in its main diagonal and an upper triangular $m \times n$ matrix such that $A = L \cdot U$.*

Theorem 7 (LU Factorization with Partial Pivoting [8]). *Let A be an $m \times n$ matrix. Then there are an $m \times m$ row permutation matrix P, a lower triangular $m \times m$ matrix L with ones in its main diagonal and an upper triangular $m \times n$ matrix such that $P \cdot A = L \cdot U$.*

Theorem 8 (LU Factorization with Complete Pivoting [8]). *Let A be an $m \times n$ matrix. Then there are an $m \times m$ row permutation matrix P, an $n \times n$ column permutation matrix Q, a lower triangular $m \times m$ matrix L with ones in its main diagonal, and an upper triangular $m \times n$ matrix such that $P \cdot A \cdot Q = L \cdot U$.*

Proposition 5. *The required floating point operations of LU factorization of an* $m \times n$ *matrix is* $O(n^2(m - \frac{n}{3}))$.

Below, we present the error analysis for the LU factorization with partial pivoting.

Proposition 6. *The LU factorization is the exact factorization of the slightly disturbed initial matrix A:*

$$A + E = L \cdot U, \qquad \|E\|_\infty \leqslant n^2 \, \rho \, u \, \|A\|_\infty,$$

where ρ *is the growth factor (in case of row pivoting) and u the unit round off.*

Remark 12. The theoretical bound of the norm of the error matrix is unfortunately large due to the growth factor.

Remark 13. It has been proved that in the case of Gaussian elimination with partial pivoting holds that [8, 25]:

$$\rho \leqslant 2^{n-1},$$

while in the case of Gaussian elimination with complete pivoting holds that:

$$\rho \leqslant \left(n \cdot 2^1 \cdot 3^{1/2} \cdot 4^{1/3} \cdots n^{1/(n-1)}\right)^{1/2}.$$

Remark 14. Although the theoretical bound for the norm of the error matrix in LU factorization with partial pivoting is too high, in practice there are only a few examples for which the error is not satisfactory. Thus, the LU factorization with partial pivoting is one of the most popular matrix-factorization methods.

Next we present a high level description of the LU factorization with partial pivoting algorithm:

Algorithm LU factorization with partial pivoting [8]

for $k = 1 : \min\{m - 1, n\}$
  Find $r : |a_{r,k}| = \max_{k \leqslant i \leqslant m}\{|a_{i,k}|\}$
  Interchange rows k and r
  $m_{ik} = -a_{ik}/a_{kk}, \; i = k + 1 : m$
  $a_{ij} = a_{ij} + m_{ik} \, a_{kj}, \; i = k + 1 : m, j = k + 1 : n$
  Set $a_{i,j} = 0$ if $|a_{i,j}| \leqslant \epsilon_t, i = k : m + n, j = k : m + n$

Row interchanges can be saved in a vector p, where p_i is the number of row which is the maximum element in absolute value in column i for the rows $i, i + 1, \ldots, m$ in step i of the algorithm. Let P_i be the permutation matrix in step i and $P = P_{n-1} \cdots P_2 \cdot P_1$, then the LU factorization with partial pivoting is $P \cdot A = L \cdot U$.

6.1 Vandermonde Matrices

The Vandermonde matrices can be used for the representation of the discrete logarithm function as well as the Diffie–Hellman mapping. These matrices are derived from the interpolation process.

In [11, 18] LU-decomposition for Vandermonde matrices through Newton polynomial has been elaborated and new forms of both these problems have been provided. These new forms constitute an alternative approach to view and study the equivalence of the two problems and evidence new ideas for the generation of new cryptographic functions. The symmetric $(p-1)\times(p-1)$ Vandermonde matrix W is used:

$$W = \{W_{ij}\},\ \ i \leqslant i,j \leqslant p-1, \quad \text{with} \quad W_{ij} = w^{(i-1)(j-1)},$$

where $w = g^{-1}$. The matrix W is a ***discrete Fourier transform***, thus explicit forms for the cryptographic function of Sect. 2 can be written as follows:

$$\log_g(x) = -(p-1,1,2,\ldots,p-2)\ W\,(x^{\,p-1},x,\ldots,x^{\,p-2})^{\top}, \tag{1}$$

and

$$K(x,y) = -(x^{\,p-1},x,x^2,\ldots,x^{\,p-2})\ W\,(y^{\,p-1},y,\ldots,y^{\,p-2})^{\top}, \tag{2}$$

respectively. Then, using LU-decomposition, the matrix W can be factorized to $W = L\cdot U$, which equals to

$$U = \begin{pmatrix} 1 & 1 & 1 & 1 & \ldots & 1 \\ 0 & w-1 & w^2-1 & w^3-1 & \ldots & w^{p-2}-1 \\ 0 & 0 & (w^2-1)(w^2-w) & (w^3-1)(w^3-w) & \ldots & (w^{p-2}-1)(w^{p-2}-w) \\ 0 & 0 & 0 & \Pi_{j=0}^{2}(w^3-w^j) & \ldots & \Pi_{j=0}^{2}(w^{p-2}-w^j) \\ \vdots & \vdots & \vdots & \ldots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 & \Pi_{j=0}^{p-3}(w^{p-2}-w^j) \end{pmatrix}.$$

Since the matrix W is symmetric, the upper triangular matrix U can also be factorized to $U = D\cdot L^{\top}$, where $D = \mathrm{diag}(U)$.

Thus, the matrix L assumes the form:

$$U=\begin{pmatrix} 1 & 0 & 0 & 0 \ldots 0 \\ 1 & 1 & 0 & 0 \ldots 0 \\ 1 & (w^2-1)(w-1)^{-1} & 1 & 0 \ldots 0 \\ 1 & (w^2-1)(w-1)^{-1} & (w^3-1)(w^3-w)(w^2-1)^{-1}(w^2-w)^{-1} & 1 \ldots 0 \\ \vdots & \vdots & \vdots & \vdots \ddots \vdots \\ 1 & (w^{p-2}-1)(w-1)^{-1} & \ldots & \ldots\ldots 1 \end{pmatrix}.$$

By setting $F(x) = L^\top x$, with $x^\top = (x^{p-1}, x, \ldots, x^{p-2})$ and by using the previous factorization of the matrix W and taking into consideration the Eqs. (1) and (2), then the ***discrete logarithm function*** can be written as follows:

$$-\eta^\top LDL^\top x = -\eta^\top LDF(x),$$

where $\eta^\top = (p-1, 1, 2, \ldots, p-2)$. Also, the ***Diffie–Hellman key function*** can be written as follows:

$$-y^\top LDL^\top x = -F^\top(y)LDF(x),$$

where $y^\top = (y^{p-1}, y, y^2, \ldots, y^{p-2})$. In the case of the ***Diffie–Hellman mapping*** (where $x = y$), we obtain the following quadratic form:

$$-x^\top LDL^\top x = -F^\top(x)DF(x),$$

which is computationally equivalent to the Diffie–Hellman function. The Diffie–Hellman mapping can also be written as follows:

$$-c^\top LDL^\top y, \quad \text{where} \quad c^\top = \left(g^0, g^{1^2}, g^{2^2}, \ldots, g^{(p-2)^2}\right).$$

6.2 LU Factorization in Cryptography

The LU factorization with partial pivoting can be applied in order to encrypt a message [6, 23]. Let $A \in \mathbb{R}^{m\times n}$ (or $A \in \mathbb{C}^{m\times n}$) be a matrix containing the initial message. If L and U are lower and upper triangular matrices, respectively, and P is a row permutation matrix as described previously, such that $P \cdot A = L \cdot U$, then the initial message is efficiently encrypted in L and U. It has been proved that the problem of restoring the initial message even though the matrix L or the matrix U is known constitutes an NP-hard problem, i.e., it cannot be solved in a practical amount of time [6]. If L is known from one person and U is known from another one, then

the two persons have to meet together and multiply their matrices in order to decrypt the initial message. Alternatively, the LU factorization with complete pivoting can be applied in order to enforce the stability of the algorithm.

Below, we present an example implementing the LU factorization with complete pivoting in order to encrypt an initial message. Then, the matrix multiplications is used in order to restore the message.

Example 1. Let us assume the following matrix:

$$A = \begin{pmatrix} 0.5688 & 0.1622 & 0.1656 & 0.6892 \\ 0.4694 & 0.7943 & 0.6020 & 0.7482 \\ 0.0119 & 0.3112 & 0.2630 & 0.4505 \\ 0.3371 & 0.5285 & 0.6541 & 0.0838 \end{pmatrix}.$$

We apply the LU factorization with complete pivoting to A.

Step 1:

The maximum element in absolute value in A is 0.7943 in the second row and second column.
Interchange rows 1 and 2 and columns 1 and 2 of A.
Compute the multipliers $A_{i,1} \equiv L_{i,1} \equiv m_{i,1} = \frac{A_{i,1}}{A_{1,1}}, i = 2, 3, 4$
Update the elements of A: $A_{i,j} = A_{i,j} - A_{i,1} \cdot A_{1,j},\ i = 2, 3, 4,\ j = 1, 2, 3, 4$

$$A^{(1)} = \begin{pmatrix} 0.7943 & 0.4694 & 0.6020 & 0.7482 \\ 0 & 0.4730 & 0.0427 & 0.5365 \\ 0 & -0.1720 & 0.0271 & 0.1574 \\ 0 & 0.0248 & 0.2535 & -0.4140 \end{pmatrix}.$$

$$L = \begin{pmatrix} 1.0000 & 0 & 0 & 0 \\ 0.2042 & 1.0000 & 0 & 0 \\ 0.3918 & 0 & 1.0000 & 0 \\ 0.6654 & 0 & 0 & 1.0000 \end{pmatrix}.$$

Step 2:

The maximum element in absolute value in $A^{(1)}_{i,j}, i = 2, 3, 4,\ j = 2, 3, 4$ is 0.5365 in the second row and fourth column.
Interchange columns 2 and 4 of A.
Compute the multipliers $A_{i,2} \equiv L_{i,2} \equiv m_{i,2} = \frac{A_{i,2}}{A_{2,2}}, i = 3, 4$
Update the elements of A: $A_{i,j} = A_{i,j} - A_{i,1} \cdot A_{1,j},\ i = 3, 4,\ j = 2, 3, 4$

$$A = \begin{pmatrix} A^{(2)} = 0.7943 & 0.7482 & 0.6020 & 0.4694 \\ 0 & 0.5365 & 0.0427 & 0.4730 \\ 0 & 0 & 0.0146 & -0.3108 \\ 0 & 0 & 0.2865 & 0.3898 \end{pmatrix}.$$

$$L = \begin{pmatrix} L = 1.0000 & 0 & 0 & 0 \\ 0.2042 & 1.0000 & 0 & 0 \\ 0.3918 & 0.2934 & 1.0000 & 0 \\ 0.6654 & -0.7718 & 0 & 1.0000 \end{pmatrix}.$$

Step 3:

The maximum element in absolute value in $A_{i,j}^{(2)}, i = 3, 4,\ j = 3, 4$ is 0.3898 in the fourth row and fourth column.
Interchange rows 3 and 4 and columns 3 and 4 of A.
Interchange rows 3 and 4 of L except the diagonal entries.
Compute the multipliers $A_{i,3} \equiv L_{i,2} \equiv m_{i,3} = \frac{A_{i,3}}{A_{3,3}}, i = 4$
Update the elements of A: $A_{i,j} = A_{i,j} - A_{i,1} \cdot A_{1,j},\ i = 4,\ j = 3, 4$

$$A = \begin{pmatrix} U \equiv A^{(3)} = 0.7943 & 0.7482 & 0.4694 & 0.6020 \\ 0 & 0.5365 & 0.4730 & 0.0427 \\ 0 & 0 & 0.3898 & 0.2865 \\ 0 & 0 & 0 & 0.2430 \end{pmatrix}.$$

$$L = \begin{pmatrix} L = 1.0000 & 0 & 0 & 0 \\ 0.2042 & 1.0000 & 0 & 0 \\ 0.6654 & -0.7718 & 1.0000 & 0 \\ 0.3918 & 0.2934 & -0.7973 & 1.0000 \end{pmatrix}.$$

$U \equiv A$

In order to reduce the required storage capacity we save the matrix U in the upper triangular part of the initial matrix A, the matrix L (except the 1's of the main diagonal) to the lower triangular part of A, the row permutation matrix P as a vector $p = [2\ 1\ 4\ 3]$, and the column permutation matrix Q as a vector $q = [2\ 4\ 1\ 3]$ (matrices P and Q are the identity matrix with interchanged their rows and columns, respectively). Thus, $P \cdot A \cdot Q = L \cdot U$. The use of $A,\ p$, and q instead of $L,\ U,\ P,\ Q$ keeps the storage capacity to $O(n^2)$ which is the order of the storage capacity of the initial data. Even knowing either U or L it is an NP-hard problem to obtain the initial

data A. In order to restore the initial matrix A the following product $P^{-1} \cdot L \cdot U \cdot Q^{-1}$ must be computed. Due to the triangular form of L and U, only the required floating point operations have to be computed reducing the computational complexity of the multiplication. P and Q are permutation matrices, thus their inverses and their product do not increase the complexity.

7 Synopsis

In the work at hand we studied the problem of transforming cryptographic schemes using interpolation techniques.

We gave explicit forms for the discrete logarithm and Diffie–Hellman cryptographic functions. Also, we presented inverse interpolation methods, such as Aitken's and Neville's methods, for the well-known discrete logarithm problem as well as the Lucas logarithm problem.

Furthermore, we gave the representation of cryptographic functions through polynomials or algebraic functions and a special case of discrete logarithm problem. Finally, we analyzed a study of cryptographic functions using factorization of matrices.

References

1. Adelmann, C., Winterhof, A.: Interpolation of functions related to the integer factoring problem. Lect. Notes Comput. Sci. **3969**, 144–154 (2006)
2. Aly, H., Winterhof, A.: Polynomial representations of the Lucas logarithm. Finite Fields Appl. **12**(3), 413–424 (2006)
3. Burden, R.L., Faires, J.D.: Numerical Analysis, 6th edn. Brooks/Cole Publishing Company, Pacific Grove (1997)
4. Camenisch, J.L.: Group signature schemes and payment systems based on the discrete logarithm problem. Doctoral Dissertation, Zurich (1998)
5. Camenisch, J.L., Stadler, M.A.: Efficient group signature schemes for large groups. Lect. Notes Comput. Sci. **1294**, 410–424 (1997)
6. Choi, S.J., Youn, H.Y.: A novel data encryption and distribution approach for high security and availability using LU decomposition. Lect. Notes Comput. Sci. **3046**, 637–646 (2004)
7. Coppersmith, D., Shparlinski, I.: On polynomial approximation of the discrete logarithm and the Diffie-Hellman mapping. J. Cryptol. **13**(3), 339–360 (2000)
8. Datta, B.N.: Numerical Linear Algebra and Applications, 2nd edn. SIAM, Philadelphia (2010)
9. El Gamal, T.: A public-key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inf. Theory **31**(4), 469–472 (1985)
10. El Mahassni, E., Shparlinski, I.E.: Polynomial representations of the Diffie-Hellman mapping. Bull. Aust. Math. Soc. **63**, 467–473 (2001)
11. Laskari, E.C., Meletiou, G.C., Tasoulis, D.K., Vrahatis, M.N.: Transformations of two cryptographic problems in terms of matrices. ACM SIGSAM Bull. **39**(4), 127–130 (2005)
12. Laskari, E.C., Meletiou, G.C., Vrahatis, M.N.: Aitken and Neville inverse interpolation methods over finite fields. Appl. Numer. Anal. Comput. Math. **2**(1), 100–107 (2005)

13. Laskari, E.C., Meletiou, G.C., Vrahatis, M.N.: Aitken and Neville inverse interpolation methods for the Lucas logarithm problem. Appl. Math. Comput. **209**, 52–56 (2009)
14. Lysyanskaya, A., Ramzan, Z.: Group blind digital signatures: a scalable solution to electronic cash. Lect. Notes Comput. Sci. **1465**, 184–197 (1998)
15. Meidl, W., Winterhof, A.: A polynomial representation of the Diffie-Hellman mapping. Appl. Algebra Eng. Commun. Comput. **13**, 313–318 (2002)
16. Meletiou, G.C.: Explicit form for the discrete logarithm over the field GF(p, k). Arch. Math. (Brno) **29**, 25–28 (1993)
17. Meletiou, G.C., Mullen, G.L.: A note on discrete logarithms in finite fields. Appl. Algebra Eng. Commun. Comput. **3**(1), 75–78 (1992)
18. Meletiou, G.C., Laskari, E.C., Tasoulis, D.K., Vrahatis, M.N.: Matrix representations of cryptographic functions. J. Appl. Math. Bioinformatics **3**(1), 205–213 (2013)
19. Mullen, G.L., White, D.: A polynomial representation for logarithms in GF(q). Acta Arith. **47**(3), 255–261 (1986)
20. Niederreiter, H.: A short proof for explicit formulas for discrete logarithms in finite fields. Appl. Algebra Eng. Commun. Comput. **1**(1), 55–57 (1990)
21. Shparlinski, I.E.: Cryptographic Applications of Analytic Number Theory: Complexity Lower Bounds and Pseudorandomness. Progress in Computer Science and Applied Logic, vol. 22. Birkhauser Verlag, Basel (2003)
22. Stadler, M.: Publicly verifiable secret sharing, advances in cryptology. Lect. Notes Comput. Sci. **1070**, 190–199 (1996)
23. Triantafyllou, D.: Numerical linear algebra methods in data encoding and decoding. J. Appl. Math. Bioinformatics **3**(1), 193–203 (2013)
24. Wells, A.L., Jr.: A polynomial form for logarithms modulo a prime. IEEE Trans. Inf. Theory **IT-30**, 845–846 (1984)
25. Wilkinson, J.H.: The Algebraic Eigenvalue Problem. Clarendon Press, Oxford (1965)

Flaws in the Initialisation Process of Stream Ciphers

Ali Alhamdan, Harry Bartlett, Ed Dawson, Leonie Simpson, and Kenneth Koon-Ho Wong

Abstract The initialisation process is a key component in modern stream cipher design. A well-designed initialisation process should not reveal any information about the secret key, or possess properties that may help to facilitate attacks. This paper analyses the initialisation processes of shift register based stream ciphers and identifies four flaws which lead to compression, state convergence, the existence of slid pairs and possible weak Key-IV combinations. These flaws are illustrated using the A5/1 stream cipher as a case study. We also provide some design recommendations for the intialisation process in stream ciphers, to overcome these and other flaws.

Keywords: Stream cipher • Initialisation • Slid pairs • Slide attack • Synchronisation attack • State convergence • A5/1

1 Introduction

Symmetric stream ciphers are used to provide confidentiality in a wide range of real-time applications such as the internet, pay TV and mobile phone transmissions. In these applications, the information being transmitted should not be accessible to unauthorised parties. The most common type of stream cipher is the binary additive stream cipher, in which the plaintext (message) is regarded as a stream of bits and encryption is performed by XORing the plaintext with a sequence of keystream bits to obtain the ciphertext. The keystream is a pseudorandom binary sequence produced by a deterministic finite state machine, known as a keystream generator. An identical keystream must also be generated and used for decryption;

A. Alhamdan
National Information Center, Riyadh, Saudi Arabia
e-mail: alhamdan@nic.gov.sa

H. Bartlett • E. Dawson (✉) • L. Simpson • K.K.-H. Wong
Institute for Future Environments, Science and Engineering Faculty,
Queensland University of Technology, Brisbane, QLD, Australia
e-mail: h.bartlett@qut.edu.au; e.dawson@qut.edu.au; lr.simpson@qut.edu.au;
kk.wong@qut.edu.au

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_2

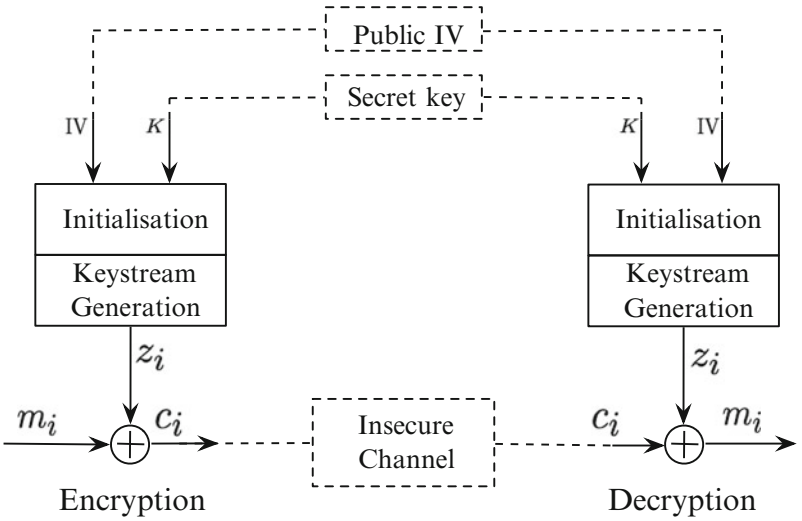


Fig. 1 Binary additive stream cipher

the keystream is XORed with the ciphertext to recover the plaintext, as shown in Fig. 1. Before the keystream generator can begin to produce an output sequence, it must have an initial value or state. Using the inputs to the keystream generator to form this initial value is known as initialisation.

For many applications, the communication is divided into sections known as packets or frames, with a different keystream sequence required for each section of the communication. Most modern keystream generators utilise two inputs: a secret key and an initialisation vector (IV) or frame number [34]. The IVs are assumed to be known information. Generally the same secret key is used for the whole communication, but with different IVs for each packet or frame. For each packet or frame, initialisation using the key and IV must be performed before a sequence of keystream bits of the required length is generated and used for encryption or decryption. This repetition of the initialisation process for the keystream generator is referred to as reinitialising or "rekeying". Examples of packet sizes used in common applications are: digital video broadcasting (DVB), 184 bytes; advanced television systems committee (ATSC), 208 bytes; general packet radio service (GPRS): 160, 240, 288 or 400 bits; and GSM mobile phone: 228 bits. The A5/1 cipher used to encrypt the frames of a GSM conversation is rekeyed every 4.6 ms [17]. The short lengths of these keystream sequences illustrate the importance of an efficient initialisation process for real-time applications such as mobile and wireless communications [20]. Additionally, the requirement for efficient initialisation should not compromise the security of the cipher.

The security provided by a stream cipher depends on the pseudorandom keystream sequences appearing to be random [14, 17]. Most cryptanalysts focus their security analysis on the keystream generation phase and do not consider the

initialisation phase. However, the initialisation process is a necessary operation before keystream generation and also affects the security of the cipher. A good initialisation process should ensure that each key-IV pair generates a distinct and unpredictable keystream and that multiple keystreams produced using the same secret key with different IVs appear unrelated. Also, the initialisation process should ensure that, even if the state of the keystream generator is revealed sometime during keystream generation, relationships between the key-IV pair and the keystreams are hard to establish so state recovery does not reveal any information about the secret key.

This paper focuses on the initialisation process of shift register based keystream generators for stream ciphers. Section 2 describes the phases of the initialisation process for the keystream generators of stream ciphers. In Sect. 3 the security of the initialisation process is investigated, and features of the cipher initialisation process which reduce resistance to common forms of attack are identified. In Sect. 4, these identified flaws are illustrated using the well-known A5/1 stream cipher as a case study. This section is based on results reported by the authors in [2, 4, 5, 45]. Section 5 discusses the existence of these flaws in the initialisation processes of certain other shift register based stream ciphers. Section 6 summarises our findings and gives some design recommendations for the initialisation processes of shift register based stream ciphers.

2 The Initialisation Process

In the initialisation process a secret key (necessary) and an IV (optional but very common) are used to form an initial state for the keystream generator, before keystream generation begins. In this paper we assume the use of an IV. The initialisation process generally consists of two phases: a *loading phase* and a *diffusion* phase. These are discussed in greater detail below.

2.1 Loading Phase

In the *loading phase*, the secret key and IV are loaded into the internal state of the keystream generator. The key and IV loading may be performed sequentially or simultaneously. For example, the A5/1 stream cipher [17] loads the secret key first followed by the IV, whereas the Grain [31] and Trivium [23] ciphers load both secret key and IV simultaneously into the internal state. In some cases, such as the common scrambling algorithm stream cipher (CSA-SC) [47], the IV is loaded during the diffusion phase but that approach is not common.

The size of the internal state relative to the lengths of the key and IV is a factor in the loading options available. For many early stream ciphers, the keystream generator state size is the same as the key length. For example, the A5/1 cipher

has a 64-bit state and uses a 64-bit key. For these ciphers, if an IV is used along with the key, both values cannot be simultaneously placed into the state space; the loading must be sequential. In a sequential process feedback functions are used to introduce the key and IV bits into the state. These functions can be either linear or nonlinear.

More recent stream ciphers generally have a state space that is at least the size of the sum of the lengths of the key and IV; this permits both key and IV values to be placed directly into the state simultaneously. Where the state size is larger than the combined size of the key and IV, if the key and IV values are simultaneously placed into stages in the internal state, predetermined values must be specified for the "unused" stages, a practice known as padding. If the state consists of binary shift registers, the loading phase must specify which stages will hold key bits, which will hold IV bits, and which of the remaining stages will be set to 0 and 1, respectively. The Trivium [23] and Sfinks [18] ciphers are examples of ciphers where the key and IV are loaded simultaneously, and the remainder of the state padded (different padding formats for each cipher). For ciphers like these, the padding specification should be considered in the security analysis.

We refer to the state contents at the end of the loading phase as the cipher's *loaded state* for that particular key and IV pair. Note that in cases where the state size is not greater than the sum of the key and IV lengths, the value of the internal state at any time (during either initialisation or keystream generation) corresponds to a loaded state for some key and IV pair. Where the state space is larger than the sum of the lengths of the key and IV an internal state at any time will only correspond to a legitimate loaded state if it meets the prescribed padding format. This is an important factor in considering the application of slide attacks, discussed in greater detail in Sects. 3.3 and 4.3 below.

2.2 Diffusion Phase

In the *diffusion phase* the internal state of the keystream generator is updated using a specified initialisation state update function but no keystream is produced. The state update function during the diffusion phase is usually a nonlinear function. This may be implemented as Boolean functions or in some cases as S-boxes. We refer to the state contents at the end of the diffusion phase as the cipher's *initial state* for a particular key and IV. Where one secret key is used for a communication, and initial states for the various packets or frames are generated from the same key but different IVs, the initial state may be referred to as a *session key*.

The objective of this phase is to diffuse the secret key and IV across the internal state, so that a state recovery attack which identifies the initial state does not compromise the secret key. That is, if an attacker recovers the initial state (session key) of a stream cipher, then the initialisation process should be sufficiently complicated to prevent them recovering the secret key by any means which is faster than exhaustive key search. Then a state recovery attack must be repeated every time the cipher is rekeyed.

The number of iterations of the state update function performed during the diffusion phase may affect both the security and efficiency of the cipher. If very few iterations are performed, the relationships between key and IV bits and keystream output may be simple and readily exploited in attacks, such as algebraic, differential and correlation attacks. A common belief in symmetric key cryptography is that increasing the number of iterations during a nonlinear initialisation process increases the security provided by the cipher, as performing more mixing of the key and IV should provide resistance to these attacks. However, this does not provide security against all attacks for all keystream generators. If state convergence occurs during the initialisation process, then increasing the number of iterations actually decreases the number of obtainable initial states. This may actually leave the cipher more vulnerable to other attacks such as time memory tradeoff (TMTO) attacks aimed at recovering a session key. This is the situation for the A5/1 stream cipher, discussed in Sect. 4.2. The probability of success of another form of attack, the slide attack, is independent of the number of iterations of the state update function. Finally, performing a greater number of iterations increases the time taken for rekeying; that is, it decreases the efficiency of the initialisation process. This may be critical in some real-time applications.

2.3 Keystream Generation

When the initialisation process is complete, the cipher is in its *initial state* and keystream generation can begin. During keystream generation, the internal state is updated using a prescribed state update function and the keystream is generated from the internal state using an output function. The state update function used during keystream generation may be the same as the state update function used in the diffusion phase of the initialisation process. If it is different, there may be a degree of similarity to the state update function used in the diffusion phase. This is an important factor in considering the application of slide attacks, discussed in Sect. 3.3.

3 Flaws in the Initialisation Process

We identified four common flaws in the initialisation processes of some shift register based stream ciphers. These are compression, state convergence, the existence of slid pairs and the existence of weak Key-IV combinations. These flaws are due to either structural features of the keystream generator or properties of the initialisation processes of these ciphers. In some cases, these flaws may be used to disclose information about the secret key or the encrypted message.

For frame based communications, information may be obtained related to multiple key and IV inputs. Possible cases to consider include input pairs which

have the same secret key but different IVs, (K, IV) and (K, IV'); or different secret keys but the same IV, (K, IV) and (K', IV); or different secret keys and different IVs, (K, IV) and (K', IV'). Compromise in the first case is potentially the most serious, as this is widely applicable in communications. For example, this would apply to a phone call encrypted using A5/1.

3.1 Compression

We noted in Sect. 2.1 that some early stream ciphers had keystream generators with a state space that was smaller than the sum of the key and IV lengths. In such cases, it is clear that multiple key-IV pairs must correspond to the same loaded state and therefore also produce the same initial state and consequently the same keystream sequence. We refer to this situation as *compression* of the key-IV space. The degree of compression can be computed as a ratio of the total number of key-IV pairs to the state size. In these cases, the key and IV are loaded sequentially into the internal state of the keystream generator. The feedback function used for the loading process will determine the actual number of Key-IV pairs per loaded state.

If the feedback functions used to load the key and IV into the internal state are simple (perhaps linear), then recovery of the loaded state may easily be extended to key recovery. Additionally, where identical keystreams are produced for different key IV pairs, the known differences in the IVs may reveal information about corresponding differences in the keys.

If compression occurs, then the effective key-IV space is reduced, and the security provided by the cipher is affected. The cipher may be vulnerable to TMTO attacks aiming to recover the loaded state. Guidelines for appropriate internal state sizes have increased over time. In 1997, Golić [29] advised an internal size larger than the key size be used to prevent TMTO and in 2000, Biryukov and Shamir [15] recommended a state size that was twice the key size. Hong and Sarkar [33, 34] revised TMTO attacks and suggested that the IV size should be at least equal to the key size. Dunkelman and Keller [25] state an IV size of at least 1.5 times the key size is needed to prevent TMTO attacks. To satisfy this condition while avoiding compression, a state size of at least 2.5 times the key size is needed.

3.2 State Convergence

State convergence occurs when a state transition function is not one-to-one. That is, two or more distinct states at one time point are mapped to the same state at the next time point. Note that state convergence is different to compression, discussed above. In fact, it is possible for a cipher to exhibit both compression and state convergence.

For keystream generators state convergence may occur during the initialisation process, during keystream generation, or both, depending on the state update

functions used in these phases. Consider state convergence occurring during the initialisation process. If the initialisation state update function is not one-to-one, then state convergence can occur in each iteration. As the number of iterations of the state update function increases, the number of obtainable initial states may decrease. That is, different key and IV inputs result in distinct loaded states that, through initialisation are mapped to the same initial state and therefore produce the same keystream. Thus, similar to the case for compression outlined above, state convergence reduces the effective size of the key-IV space. This is the case for the A5/1 stream cipher. State convergence for A5/1 is discussed in Sect. 4.2. This may leave the cipher vulnerable to attacks such as distinguishing attacks [41], time-memory-data trade-off attacks [15] or other ciphertext-only attacks [22].

Clearly the efficiency of the initialisation process decreases as the number of iterations of the state update function increases. Note that for ciphers where state convergence occurs during initialisation, as the number of iterations of the initialisation process increases, the entropy of the secret key is effectively decreased. That is, increasing the number of iterations may actually be decreasing the effective security. However, having few iterations during the diffusion phase may make the cipher vulnerable to attacks such as correlation or algebraic attacks. For a given stream cipher, the optimal number of iterations during the initialisation process should be chosen carefully after extensive security analysis.

3.3 Slid Pairs and Shifted Keystream

The state update function of the initialisation process defines a path of transitions of internal state values. The loaded state resulting from a key-IV pair (K, IV) represents one point on such a path. If a later state in this path is the same as the loaded state resulting from another key-IV pair (K', IV'), then the two loaded states associated with the distinct input pairs (K, IV) and (K', IV'), respectively, are said to form a *slid pair*.

If the state update functions for the diffusion phase and for keystream generation are the same, then the keystream sequence obtained from the second key-IV pair will simply be a phase-shifted version of the keystream sequence obtained from the first key-IV pair [16, 24, 37, 40, 49]. Figure 2a illustrates the initialisation and keystream generation processes for two distinct key-IV pairs, (K, IV) and (K', IV'), where the corresponding loaded states are separated by α iterations of the diffusion state update function. The corresponding keystream sequences are shifted by $\epsilon \cdot \alpha$ bit(s) relative to one another, where ϵ is a positive constant that depends on the output function of the stream cipher. (For a bit based stream cipher, $\epsilon = 1$.)

If the update functions for diffusion and keystream generation are similar, but not identical, then the keystream sequence obtained from the second key-IV pair may be a phase shifted version of the keystream sequence obtained from the first key-IV

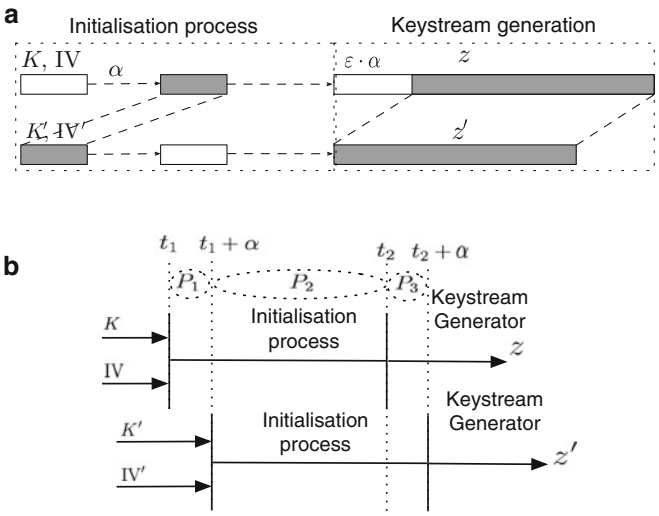


Fig. 2 Slid pairs in stream ciphers. (**a**) Two slid pairs and shifted keystream. (**b**) Analysis of slid pairs and shifted keystream

pair [16, 24, 37, 40, 49], with some probability. A slid pair is guaranteed to generate shifted keystream when the following properties hold:

(a) The state update functions used for each iteration of the diffusion phase of initialisation are the same as each other.
(b) The state update functions used for each iteration of the keystream generation process are the same as each other.
(c) The state update functions used for the initialisation and keystream generation processes are the same as one another.

Conditions (a) and (b) above hold for most stream ciphers. Condition (c) may apply with probability less than one if there is some similarity between the two state update functions. That is, the outputs of two similar functions may be the same for a subset of input values. Therefore, the probability of obtaining a slid pair that produces a correspondingly phase shifted keystream depends on the three probabilities P_1, P_2 and P_3, as shown in Fig. 2b and defined as follows:

– P_1 is the probability that a legitimate loaded state occurs after α iterations of the initialisation process.
– P_2 is the probability that the state updates for the final $t_2 - (t_1 + \alpha)$ iterations of the diffusion phase for the loaded state corresponding to (K, IV) have the same effect as the first $t_2 - (t_1 + \alpha)$ iterations of the diffusion phase for the loaded state corresponding to (K', IV').
– P_3 is the probability that the state updates for the first α iterations of keystream generation for the loaded state corresponding to (K, IV) have the same effect as the state updates during the last α iterations of the diffusion phase for the loaded state corresponding to (K', IV').

The total probability that a randomly chosen key-IV pair has a corresponding slid pair which produces a phase-shifted keystream for a slide distance of α can be calculated as the product of these three probabilities. Note that if condition (a) holds, then $P_2 = 1$, and if conditions (b) and (c) both hold, then $P_3 = 1$.

The relationships between the multiple key-IV pairs that result in the loaded states which are slid pairs, and which produce shifted keystreams may be exploited in known plaintext slide attacks. These are sometimes referred to as slid pair attacks, resynchronisation attacks [37, 48] or related key chosen IV attacks [38]. This form of attack was first developed for block ciphers and has been applied to stream ciphers based on block ciphers such as LEX [48] and WAKE-ROFB [16]. More recently it has been applied to other stream ciphers such as Grain [24, 37, 49] and Trivium [40]. This property means that the applicability of slide attacks to shift register based stream ciphers is independent of the number of iterations of the state update function performed in the diffusion phase. Clearly, increasing the number of iterations of the state update function in the diffusion phase does not increase the security of the cipher with respect to these types of attack, although it does decrease the efficiency of the initialisation process.

3.4 Weak Key-IV Combinations

For some shift register based stream ciphers, certain key-IV pairs result in internal states in which one or more of the component registers have all zero contents. If this occurs in the initial state of a component and that particular component is autonomous during keystream generation, then it will remain in an all-zero state throughout keystream generation. The component therefore contributes a constant value to the output function throughout keystream generation, so that, for that key-IV pair, the keystream generator is equivalent to another generator with fewer components and a smaller internal state size. We refer to such key-IV pairs as *weak key-IV pairs*.

The key and IV bits in each weak key-IV pair must satisfy certain relationships in order for this result to occur. For some ciphers it is possible to distinguish keystreams produced from keystream generators loaded with weak key-IV pairs. If the keystream can be detected, an attacker may use their knowledge of the relationships between key and IV bits which result in weak keys to recover information about the secret key, given the known IV. This has previously been observed in Grain v0, v1 and 128 [49]. In Sect. 4 we will show that it also occurs in the A5/1 cipher.

4 Case Study: A5/1

In this section we demonstrate that the flaws in the initialisation processes of stream ciphers, as discussed in Sect. 3, all exist in the A5/1 stream cipher. The A5/1 stream cipher [17, 19, 29] is used to protect the privacy of GSM mobile

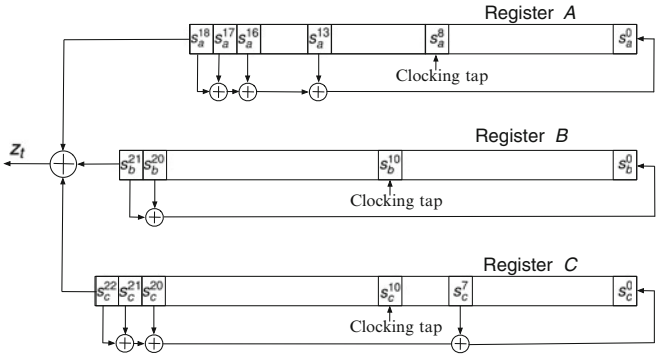


Fig. 3 A5/1 stream cipher

telephone communications. Each telephone conversation uses one secret key for all frames of that conversation, and the frame number is used to form an IV. For each frame, the initialisation process is performed and then a 228-bit keystream is generated and used to encrypt the frame (approximately 4.6 ms duration). A5/1 has received much attention from cryptographers [10, 13, 17, 28–30]. However, most of the analyses have looked at the keystream generation process rather than the initialisation process. We primarily consider the initialisation process in this section.

A5/1 is a bit-based cipher that takes a 64-bit secret key and 22-bit IV (frame number) as inputs into a 64-bit internal state. The state consists of the contents of three binary linear feedback shift registers (LFSRs), denoted A, B and C, with lengths of 19, 22 and 23 bits, respectively, as shown in Fig. 3. Each shift register has a primitive feedback polynomial. We use S to denote the internal state of A5/1 and S_A, S_B and S_C to denote the internal states for the registers A, B and C, respectively. Let $s^i_{a,t}$ denote the content of the ith stage of register A at time t, (for $0 \leq i \leq 18$). Similarly, let $s^j_{b,t}$ and $s^k_{c,t}$ denote the jth stage of register B, (for $0 \leq j \leq 21$) and the kth stage of register C, (for $0 \leq k \leq 22$), respectively, at time t.

The **loading phase** of A5/1 begins with the contents of all stages of the three registers being set to zero. Each LFSR is then regularly clocked 64 times as the key bits are XORed successively into the feedback bit of the register. Following this, the 22-bit IV is loaded in the same manner [17]. Note that the state update function during the loading phase is entirely linear, and that the key and then IV have been loaded into each register separately. This produces the loaded state of the A5/1 keystream generator. The contents of each stage in each register of the loaded state are independent linear combinations of key and IV bits.

The **diffusion phase** consists of 100 iterations of a majority clocking mechanism. To implement this, a clocking tap is designated in each register (namely, stages s^8_a, s^{10}_b and s^{10}_c). The contents of these stages at time t determine which registers will be clocked in the next iteration, at time $(t + 1)$. More specifically, those registers for which the clock control bits agree with the majority value are clocked. For example, if $s^8_{a,t} = 0$, $s^{10}_{b,t} = 1$ and $s^{10}_{c,t} = 0$, then the majority value is 0 and registers A and

C are clocked at time $(t + 1)$. Under this mechanism, either two or three registers are clocked in each iteration. There is no output from the shift registers during the diffusion phase.

After initialisation is complete, **keystream generation** begins. A further 228 iterations of the state update function are performed, using the same majority clocking rule as in the diffusion phase. In each iteration, the keystream bit is obtained by XORing the output bit of each of the three registers. That is, $z_t = s_{a,t}^{18} \oplus s_{b,t}^{21} \oplus s_{c,t}^{22}$. Note that the majority clocking mechanism used in both the diffusion phase of initialisation and during keystream generation is the only nonlinear function in the operation of A5/1.

4.1 Compression and A5/1

The loading phase of the A5/1 initialisation process transfers the 64-bit secret key and 22-bit frame number (IV) into the internal state. Since the total size of the secret key and IV ($64 + 22 = 86$ bits) exceeds the 64-bit state size, it is clear that compression occurs. In fact, as the state-update function is linear during the loading phase, and the three LFSR lengths are coprime, it can be shown that there are 2^{22} key-IV pairs corresponding to each possible loaded state.

Given the use of a 64-bit key and the 64-bit state size, it is clear that A5/1 is vulnerable to a TMTO attack. The attack may be performed to recover either the loaded state or the initial state of the cipher. Note that due to the linear loading process, recovery of the loaded state (for a known IV) translates directly to key recovery. Once the key is recovered for one frame, the contents of all frames in the conversation can be revealed.

4.2 State Convergence in A5/1

For A5/1, state convergence occurs during both the diffusion phase of initialisation and the subsequent keystream generation process. This is due to the majority clocking scheme used for the state update function during these two processes. Convergence after the first iteration of the diffusion phase was first reported by Golić [29, 30], who also stated the extent of convergence at this iteration. Since then, others have attempted to extend this analysis across the diffusion phase, using either experimental or theoretical approaches. Biryukov et al. [17] used experimental data from a random sample of A5/1 states to estimate that the set of possible initial states contains only about 15 % of all possible 64-bit states. Alhamdan [1] performed an exhaustive experimental evaluation on a scaled-down version of the A5/1 stream cipher, and found similar proportions. Kiselev and Tokareva [36] used a theoretical approach to extend Golic's results, but obtained results which conflict with those

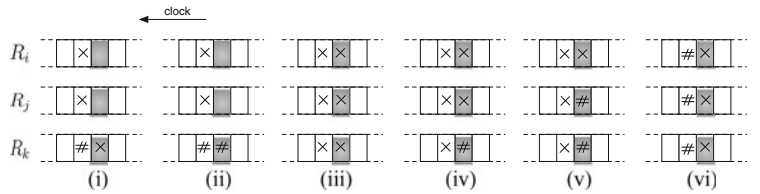


Fig. 4 Graphical representation of Golić's A5/1 pre-image cases

Table 1 Proportions of loaded states for each of Golić's cases

Case	(i)	(ii)	(iii)	(iv)	(v)	(vi)
Proportion of states	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{32}$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{1}{32}$
Number of pre-images	0	1	1	2	3	4

published previously. In this section, we outline these previous analyses, and also provide our extension of Golić's results, based on theory, to a larger number of iterations.

Golić [30] considered the inverse mapping for the A5/1 majority clocking function. He identified the format of states with no pre-image; that is, states which cannot be reached from any loaded state in a single iteration. We refer to these as inaccessible states. Note that these states may occur as loaded states, but cannot occur at any time after that. These inaccessible states are of the format depicted as case (i) in Fig. 4. In this figure, (R_i, R_j, R_k) is any permutation of the set $\{A, B, C\}$ of registers and the shaded stage in each register is its clocking tap. The symbol $\times$ represents either 0 or 1, while # represents the complement of $\times$; a blank stage represents a stage where the contents can take either value. States with this format may occur as loaded states, but cannot be reached from any valid state after the first iteration of the initialisation state update function. Golić demonstrated that states with no preimage comprise $\frac{3}{8}$ of the loaded states of the system. Thus, the usable state space shrinks by a factor of $\frac{5}{8}$ (from 2^{64} to $5 \times 2^{61} \approx 2^{63.32}$) after the first iteration of the diffusion phase. Golić also identified the format of states with unique pre-images and others with up to four pre-image states. Golić's results clearly demonstrate that the majority clocking process is not one-to-one and that state convergence can occur in one iteration. Figure 4 presents a graphical summary of the six cases identified by Golić. The proportion of loaded states for each of the six cases depicted in Fig. 4 is presented in Table 1, along with the corresponding number of pre-images.

In the diffusion phase, once the first iteration of the state update function has occurred, it is not obvious what proportion of the remaining states will become inaccessible in subsequent iterations. Clearly the proportions for the first iteration will not hold for the second iteration, as all of the states of the format depicted as case (i) in Fig. 4 have been removed from the pre-image space. Obtaining precise figures for convergence across the 100 iterations of the diffusion phase using a theoretical approach seems difficult. Biryukov et al. [17] used an experimental approach to try to quantify the level of convergence across the diffusion phase. They

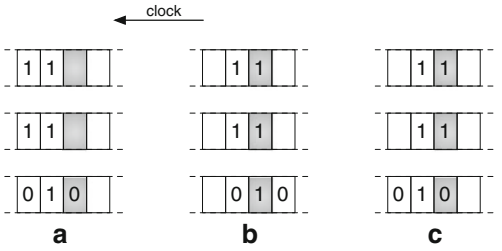


Fig. 5 Counter-example to Kiselev and Tokareva (**a**) state claimed by Kiselev and Tokareva to be inaccessible at second iteration; (**b**) inaccessible state at first iteration which clocks to state (**a**); (**c**) accessible state at first iteration which also clocks to state (**a**)

took a random sample of 100,000,000 A5/1 states and then tried to work through the state transition function in the reverse direction for 100 iterations, to form an estimate of the proportion of all possible 64-bit states that could occur as loaded states. Their results indicate that the set of loaded states contains only about 15 % of all possible 64-bit states.

More recently, Kiselev and Tokareva [36] tried to extend Golić's [30] work to determine theoretically the effective key space reduction in each of the first eight iterations of the diffusion phase. Their results for the number of inaccessible states after the first iteration agree with previously reported results, but the results for further iterations are inconsistent with the experimental results presented in [1, 17]. This is a result of a false assumption on their part: that any state which is accessible from an inaccessible state is also inaccessible. In fact, many of these states can be reached by clocking from other accessible states as well from the inaccessible states. Thus, these authors have included many accessible states in their claimed list of inaccessible states, for each iteration beyond the first. We provide a counter-example to their claims. State (a) in Fig. 5 is one example of a state they claim is inaccessible at the second iteration [36, Figure 4]. Their reasoning is that state (a) can be obtained by clocking state (b), and given that state (b) is inaccessible at the first iteration, they claim that state (a) must therefore be inaccessible at the second iteration. However, state (a) can also be reached by clocking state (c), which is accessible at the first iteration [see Fig. 4(iv)]. Therefore, state (a) is accessible at the second iteration. Thus, Kiselev and Tokareva's analysis is shown to be flawed.

The work summarised below takes a theoretical approach, and extends Golić's logic to identify the states which cannot be reached after each of the first six iterations of the diffusion phase. It shows that state convergence continues with each iteration, though not uniformly at each iteration, contrary to Golić's assumptions in [30].

Consider the first two iterations in the diffusion phase. Applying Golić's logic to identify loaded states of particular formats, a particular state will be inaccessible after two iterations only if it either matches case (i) in Fig. 4 or has a preimage which contains only states which match this case. Since case (i) cannot be reached after the

Table 2 Proportion of available states after α iterations

α (number of iterations)	1	2	3	4	5	6
New proportion inaccessible	$\frac{3}{8}$	$\frac{3}{64}$	$\frac{9}{512}$	$\frac{57}{4096}$	$\frac{423}{32768}$	$\frac{6453}{524288}$
Cumulative proportion inaccessible	0.375	0.422	0.439	0.453	0.466	0.479
Proportion accessible	0.625	0.578	0.561	0.547	0.534	0.521
Number of accessible states	$2^{63.322}$	$2^{63.209}$	$2^{63.165}$	$2^{63.129}$	$2^{63.094}$	$2^{63.061}$

first iteration, a state which can only be reached from this case cannot be reached at the second (or any subsequent) iteration. (Note: Case (i) is a valid loaded state.)

A similar process can be followed to identify patterns for inaccessible states after α iterations. We have done this for $2 \leq \alpha \leq 6$ and found a branching tree of patterns for these inaccessible states: see [45] for some examples and further details. Table 2 presents the cumulative proportion of inaccessible states (out of all possible loaded states) after each of the first six iterations, together with the corresponding proportion and number of accessible states.

The complexity and the number of distinct patterns obtained so far indicates that obtaining a general expression for the number of accessible states after a given number of iterations is not a simple task for large values. Extrapolating from the known values in Table 2 provides an approximation. Using an exponential extrapolation based on the proportion of accessible states as reported above for 2–6 iterations, we estimate that the proportion of accessible states after 100 iterations is around 5 % of the number of loaded states. The extrapolation is based on a linear regression fit to the logarithm of the proportion accessible [6]. The results presented above for small numbers of iterations align closely with those from Alhamdan's exhaustive experimental analysis on a scaled-down version of A5/1 [1]. Table 3 shows a summary of the previous works and the current work. Alhamdan [1] found the proportion of distinct states after 100 iterations was 19.2 % of the original loaded states. This is close to Biryukov, Shamir and Wagner's experimental result for A5/1, which is 15 % [17]. Our extrapolation based on the results in Table 2 for the proportion of accessible states is 5 %.

As noted in Sect. 3.2, state convergence may leave a cipher vulnerable to distinguishing attacks, time-memory-data trade-off attacks and other ciphertext-only attacks. In fact, state convergence in A5/1 was one of the contributing factors in Biryukov, Shamir and Wagner's practical time-memory attack on this cipher [17]. In addition, the presence of state convergence in A5/1 may reduce the search space for attacks such as the ciphertext-only attack we present in Sect. 4.4.

Table 3 Comparison between the analysis of inaccessible states

	Cumulative proportion of state reduction							
No. of clocks	1	2	3	4	5	6	10	100
Golić[a] [30]	0.375	–	–	–	–	–	–	–
BSW[b] [17]	–	–	–	–	–	–	–	0.85
Alhamdan[c] [1]	0.375	0.422	0.439	–	0.466	–	0.524	0.81
KT[a] [36]	0.375	0.578	0.689	0.767	0.826	–	–	–
This work[a]	0.375	0.422	0.439	0.453	0.466	0.479	–	0.95[d]

[a]Theoretical analysis
[b]Based on 10^8 randomly simulated states
[c]Exhaustive search for a scaled-down version
[d]Based on exponential extrapolation

4.3 Slid Pairs and Synchronisation Attacks

Since every state of A5/1 is a valid loaded state, it is clear that the internal state obtained from any loaded state after some number α of iterations is also a legitimate loaded state. That is, for any key-IV pair and any $\alpha > 0$, it is always possible to find a second key-IV pair such that the loaded state from the second pair can be obtained from the loaded state of the first pair by applying α iterations of the (diffusion) state update function. Thus, the loaded states corresponding to these two key-IV pairs form a slid pair separated by α clocks. Further, since the update functions during diffusion and keystream generation of A5/1 are identical, this slid pair will always produce keystream sequences which are out of phase by α bits.

We show below how the state update operations in A5/1 can be represented in terms of matrix equations. An analysis of these equations then enables us to identify the conditions under which a slid pair can occur. We particularly focus on the scenario where the loaded states in the slid pair were both generated from the same secret key (but necessarily with different IVs). As discussed in Sect. 3, this is the most practical scenario for the frame-based communications context in which A5/1 is used.

Since each register is loaded independently, we first develop the equations for a single register. First note that the autonomous operation of any LFSR can be described using a matrix equation [39, 46] of the form

$$S_{t+1} = TS_t$$

where the state transition matrix T shifts the contents of each stage of the register to the subsequent stage and inserts the feedback bit into the relevant stage. This equation can then be adapted to include loading of the key bits by writing:

$$S_{t+1} = TS_t \oplus \sigma k_t$$

where $\sigma = [1\ 0 \ldots 0]^{\mathsf{T}}$ indicates that the relevant key bit is XORed into the feedback of the LFSR. Setting $t = \tau$ to indicate the register state before loading commences, iterating this equation 64 times and collecting terms then gives

$$S_{\tau+64} = T^{64} S_\tau \oplus NK$$

where $N = [T^{63}\sigma \quad T^{62}\sigma \ldots T\sigma \quad \sigma]$ and $K = [k_0 \quad k_1 \ldots k_{62} \quad k_{63}]^{\mathsf{T}}$. Using a similar approach to represent the loading of the 22 IV bits, we obtain

$$S_{\tau+86} = T^{86} S_\tau \oplus T^{22} NK \oplus MV$$

where $M = [T^{21}\sigma \quad T^{20}\sigma \ldots T\sigma \quad \sigma]$ and $V = [v_0 \quad v_1 \ldots v_{20} \quad v_{21}]^{\mathsf{T}}$. Finally, noting that the state S_τ for A5/1 contains all zeros and setting $\tau = -86$, so that $t = 0$ corresponds to the loaded state of the register, we have

$$S_0 = T^{22} NK \oplus MV \tag{1}$$

The above result applies individually to each of the registers A, B and C of A5/1. If we let T_A, T_B and T_C denote the state transition matrices of these registers and denote the corresponding σ and N matrices for these registers as σ_A, σ_B, σ_C and N_A, N_B, N_C, respectively, then Eq. (1) can also be applied to the combined state of A5/1 by defining

$$T = \begin{bmatrix} T_A & 0 & 0 \\ 0 & T_B & 0 \\ 0 & 0 & T_C \end{bmatrix}, \sigma = \begin{bmatrix} \sigma_A \\ \sigma_B \\ \sigma_C \end{bmatrix} \quad \text{and} \quad N = \begin{bmatrix} N_A \\ N_B \\ N_C \end{bmatrix}$$

Now consider the operation of A5/1 during the diffusion phase. In this phase, the registers of A5/1 are clocked using a majority clocking rule, so there are now four different cases to be considered at each iteration. These are:

Case 1. All registers are clocked
Case 2. Registers A and B are clocked
Case 3. Registers A and C are clocked
Case 4. Registers B and C are clocked

The state update process during diffusion can thus be represented as $S_{t+1} = T_D S_t$, where the state transition matrix T_D depends on the case as follows:

Case 1: $T_D = T$ (as above)

Case 2: $T_D = \begin{bmatrix} T_A & 0 & 0 \\ 0 & T_B & 0 \\ 0 & 0 & I \end{bmatrix}$

Case 3: $T_D = \begin{bmatrix} T_A & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & T_C \end{bmatrix}$

Case 4: $T_D = \begin{bmatrix} I & 0 & 0 \\ 0 & T_B & 0 \\ 0 & 0 & T_C \end{bmatrix}$

Table 4 Slid pairs after 1 clock

Case:	1	2	3	4
Free key bits	20	22	21	20
IV bits to specify	4	22	22	22

Now suppose that we are looking for slid pairs with a slide distance of $\alpha = 1$ in which both loaded states arise from the same secret key. We have

$$S_1 = T_D S_0 \quad \text{with} \quad S_0 = T^{22}NK \oplus MV$$

If S_1 is also a loaded state for the same key K and a different IV, V', we have

$$S_1 = T^{22}NK \oplus MV'$$

as well, and hence

$$MV \oplus MV' = S_0 \oplus S_1$$

$$\begin{aligned} \text{or} \quad M\Delta &= (I \oplus T_D)S_0 \\ &= (I \oplus T_D)(T^{22}NK \oplus MV) \end{aligned} \tag{2}$$

where $\Delta = V \oplus V' = [\delta_0\ \delta_1 \ldots \delta_{21}]^\mathsf{T}$.

For each of the cases of T_D, we can use Eq. (2), together with the conditions guaranteeing the relevant type of clocking, to determine a set of conditions on the various bits of Δ, K and V that must be satisfied in order for a slid pair to occur in the manner described above. These equations have been analysed using Gaussian elimination and the results show that the values of certain key bits (the "free" key bits), together with known bits from V, fully determine the remaining key bits when a slid pair of this type occurs. Table 4 presents the number of key and IV bits that must be specified in order to determine the remaining key bits for each of the cases.

For example, in Case 1, there are 20 key bits and 4 IV bits that can be freely chosen in order to form a 64-bit secret key for which two related IVs generate a slid pair. For a particular choice of these 24 bits, the remaining 44 key bits are specified. The rest of the IV bits do not affect the secret key. Considering all possible values for the 20 free key bits and the 22 IV bits, the total number of slid pairs for Case 1 is 2^{42}. The probability that a randomly chosen key satisfies these equations for a given IV is 2^{-44}. The total number of slid pairs in each of Cases 2, 3 and 4 can be calculated similarly, to obtain 2^{44}, 2^{43} and 2^{42} respectively. So the total number of slid pairs with $\alpha = 1$ (for the 4 cases combined) is 2^{45}. Likewise, the probability that a randomly chosen key satisfies the equations for any of these cases (for a given IV) is found to be 2^{-41}.

Table 5 gives two examples generated using the equations discussed above. In each example, a secret key and two different IVs generate a slid pair when loaded,

Table 5 Examples of 1-bit shifted keystreams from the same secret key

Key	0x2D37B6F7292DFFFB
IV1	0x200000
IV2	0xE05A00
Keystream1	{0}0x5E449A6F3414F3CD76F567275D31CFE1A4F4AE4F4D3C954D3CB124D9A
Keystream2	0x5E449A6F3414F3CD76F567275D31CFE1A4F4AE4F4D3C954D3CB124D9A
Key	0xF77832CC89EFFFFB
IV1	0x200000
IV2	0x4001A4
Keystream1	{1}0xF798818F32A6B4772F5B2E55B8808541301E49CA76B11BC46F65C1494
Keystream2	0xF798818F32A6B4772F5B2E55B8808541301E49CA76B11BC46F65C1494

leading in turn to keystream sequences with a one bit phase shift. In each example, the two 228-bit keystream sequences contain a common 227-bit sequence; note that the final byte of keystream only contains three or four bits, which are treated as MSBs in each case.

The above analysis can be readily extended to shifts of $\alpha = 2, 3$ or more. This is done by including additional iterations of T_D in the derivation of Eq. (2) and considering the relevant combinations of cases for T_D that may be involved.

4.3.1 Attack Algorithm

A ciphertext-only key recovery attack on A5/1 can be performed if a pair of frames in a single conversation with one-bit phase shifted keystreams can be identified. (Similar algorithms can also be developed for other phase shift values.) We identify such a pair of frames by noting that the XOR of two frames encrypted with the same keystream is just the XOR combination of the corresponding plaintext frames, and that such a combination can be easily identified due to the redundancy of plaintext [22]. In the following, we assume only that the attacker is able to get enough encrypted speech (ciphertext) and that the IVs (frame numbers) of all frames are known.

The attack algorithm is as follows:

Step 1: Divide the encrypted speech (ciphertext) into separate frames, each with its known frame number (IV).

Step 2: Compare each encrypted frame with a one-bit shifted version of each other encrypted frame, using the redundancy property to identify when a slid pair has occurred.

Step 3: If a phase shifted keystream is identified, use the known IV with each possible value of the free key bits to find the secret key. (Here we use the equations discussed above Table 4 and check all four cases if required.)

Step 4: Use the secret key and known IVs to decrypt all frames and reveal the plaintext of the entire conversation.

Table 6 Complexity of various length of conversation

No. of frames	Conversation time	Comparison complexity	Probability of success
2^{14}	1 min 16 s	2^{28}	2^{-35}
2^{16}	5 min 2 s	2^{32}	2^{-31}
2^{18}	20 min 6 s	2^{36}	2^{-27}
2^{20}	1 h 21 min	2^{40}	2^{-23}
2^{22}	5 h 22 min	2^{44}	2^{-19}

As rekeying is performed in A5/1 every 4.6 ms, the total time needed to use all 2^{22} possible frame numbers is around 5 h and 22 min. Table 6 gives some examples of the number of frames used for various lengths of conversation, together with the corresponding number of comparisons involved in the attack and the probability of success for the attack on a conversation of that length. In a successful attack, up to 2^{23} choices of free key-bits must be checked in Step 3.

4.4 Weak Key-IV Combinations

The registers of A5/1 operate nonautonomously during the loading phase, since the new bit of each register during this phase depends on both the feedback and an external value (the key or IV bit). During the diffusion phase and keystream generation process, however, the feedback for each register is autonomous, as there is no external input at these times. If any of the registers contains all-zero values at the end of the loading phase, it will therefore remain in this state throughout the subsequent diffusion phase and keystream generation process. Furthermore, if two or three registers contain all-zero values at that time, then the keystream generated for that frame will be constant: either all zeros or all ones. This flaw results in sending a frame in clear text if the keystream is zeros or as the complement of cleartext if the keystream is all ones.

Recall that the loading phase of A5/1 can be represented by Eq. (1). The terms $T^{22}NK$ and MV which are XORed together can also be represented by concatenating $T^{22}N$ and M and multiplying by a new vector KV, where $KV = [k_0 \quad k_1 \ldots k_{62} \quad k_{63} \quad v_0 \quad v_1 \ldots v_{20} \quad v_{21}]^{\mathrm{T}}$, as follows

$$S_0 = [T^{22}N||M]KV \tag{3}$$

We now apply Eq. (3) to each register of A5/1 to investigate the conditions under which that register will contain all zeros at the end of the loading phase. Writing this equation separately for each register, we have:

$$\begin{aligned} S_{A,0} &= [T_A^{22} N_A || M_A] KV \\ S_{B,0} &= [T_B^{22} N_B || M_B] KV \\ S_{C,0} &= [T_C^{22} N_C || M_C] KV \end{aligned} \tag{4}$$

We consider two scenarios below: a single register containing all-zero values and two registers containing all-zero values. The case with all three registers containing all-zero values is treated as a special case of the scenario with two registers containing all-zero values.

4.4.1 One Register All Zeros

This section focuses on the situation in which only one register contains all-zero values after the loading phase. We denote the cases in which register A, B or C is the all-zero register as Case 1, Case 2 and Case 3, respectively. Whether or not this register is clocked at any iteration during keystream generation, its contribution to the keystream bit will always be zero. At each iteration of keystream generation, at least one of the other two registers will be clocked and the value of the keystream bit will depend only on the output bits of those two registers. The keystream in these circumstances will therefore contain both zeros and ones.

At this stage, we have been unable to find a way of distinguishing this keystream from the normal keystream of A5/1, so it is not yet possible to mount an attack based on this condition. However, as we show below, the case of a single register containing all zeros at the end of loading is very common, so this would be a widely applicable attack if a distinguisher could be found for the keystream from this situation.

By setting any of the left-hand terms $S_{A,0}$, $S_{B,0}$ or $S_{C,0}$ to zero in Eq. (4), we obtain a corresponding set of conditions on the key and IV bits for that register to contain all-zero values at the end of the loading phase. As in analysing the conditions for slid pairs to occur, we find in each case that the known IV and a subset of the key bits together determine the remaining key bits for that case. Table 7 presents the number of key and IV bits that must be specified in order to determine the remaining key bits for each of these cases.

From Table 7, the number of weak key-IV pairs for each of cases 1, 2 and 3 are 2^{67}, 2^{64} and 2^{63}, respectively. Ignoring a minor degree of overlap, the total number of weak key-IV pairs that result in freezing one register of A5/1 is $2^{67.25}$, so the probability that a randomly chosen key satisfies the equations for any of these cases (for a given IV) is $2^{-18.75}$. For a conversation containing N frames, the probability

Table 7 Weak key-IV pairs (one all-zero register)

Case:	1	2	3
Free key bits	45	42	41
IV bits to specify	22	22	22

that at least one frame has a weak key-IV pair of this type is approximately $2^{-18.75}N$, which is approximately $2^{-2.75}$ for a 5-min conversation (about 2^{16} frames). Thus, we see that this type of weak key-IV pair occurs very frequently in practice.

4.4.2 Two Registers All Zeros

We now focus on the situation in which two registers contain all-zero values after performing the loading phase; we denote the cases in which the pair of all-zero register is (A, B), (A, C) or (B, C) as Cases 4, 5 and 6, respectively. As the clocking stage in each of these two registers will contain a zero, the majority value will be zero and hence these two registers will be clocked every time. The third register will be clocked until the content of its clocking stage has value "1". Since the diffusion phase consists of 100 clocking steps before producing any keystream bits and the largest register has only 23 stages, this process will ensure that the third register will be in its steady state before the keystream generation begins. Thus, the leftmost stage of the non-zero register will be fixed and will be the value of the keystream bit.

By setting the relevant pair of left-hand terms $S_{A,0}$, $S_{B,0}$ or $S_{C,0}$ to zero in Eq. (4), we can again obtain a set of conditions on the key and IV bits for that pair of registers to contain all-zero values at the end of the loading phase and find in each case that the known IV and a subset of the key bits together determine the remaining key bits for that case. Table 8 presents the number of key and IV bits that must be specified in order to determine the remaining key bits for each of these cases, while Table 9 shows two examples of weak key-IV pairs (Case 4) that produce fixed keystream (either all zeros or all ones). (In Table 9, underlining indicates the output stage of each register, while bold face type indicates the stages used to control the register clocking. The keystream is presented in hexadecimal notation.)

From Table 8, the number of weak key-IV pairs for each of cases 4, 5 and 6 are 2^{45}, 2^{44} and 2^{41} respectively. Therefore, the total number of weak key-IV pairs for which two registers contain all-zeros is $2^{45.64}$, so the probability that a randomly chosen key satisfies the equations for any of these cases (for a given IV) is $2^{-40.36}$. For a conversation containing N frames, the probability that at least one frame has a weak key-IV pair of this type is approximately $2^{-40.36}N$, which is approximately $2^{-24.36}$ for a 5-min conversation. For a full set of 2^{22} IVs with a fixed key, the probability is $2^{-18.36}$.

4.4.3 Attack Algorithm

The fact that the keystream is constant for weak key-IV pairs of this sort allows us to recognise when such a pair has been used and to mount a ciphertext-only attack that enable the whole conversation to be decrypted. This attack also allows the secret key to be determined, but this is not necessary to the success of the attack. The essence of this attack is to recognise when such a key-IV pair has been used and then use the relations discussed above to determine the loaded state for that frame. Once this has

Table 8 Weak key-IV pairs (two all-zero registers)

Case:	4	5	6
Free key bits	23	22	19
IV bits to specify	22	22	22

Table 9 Two examples of weak key-IV pairs (Case 4)

Key		01101001010000101001101110111010010010110111111111111111111111001
IV		11100000000000000000
Loaded state	A	00000000**0**00000000000
	B	00000000000**0**000000000
	C	11111110110**0**010110111001
Initial state	A	00000000**0**00000000<u>0</u>
	B	00000000000**0**000000000<u>0</u>
	C	11111111011**0**01011011100<u>0</u>
Keystream		0x00000000000000000000000000000000000000000000000000000000000000000000
Key		10000100010111011010101011001101100110111111111111111111111001
IV		11000000000000000000
Loaded state	A	00000000**0**00000000000
	B	00000000000**0**000000000
	C	11101101110**1**111101001111
Initial state	A	00000000**0**00000000<u>0</u>
	B	00000000000**0**000000000<u>0</u>
	C	01110110111**1**011111010011<u>1</u>
Keystream		0xFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

been confirmed, it is straightforward to decrypt all other frames of the conversation and to recover the secret key, if required.

In this attack, we rely on the ability to recognise a frame of conversation that has been sent as cleartext or the complement of cleartext. If such a frame is identified, we then use the relations discussed above to determine the secret key from a reduced candidate set. We present the attack in two phases:

Phase 1: Identify a weak frame in the conversation.
Phase 2: Recover the loaded state for that frame—and then for all other frames.

Phase 1: Identify a weak frame

Given an encrypted conversation:

1. Divide the ciphertext (encrypted speech) into separate frames.
2. For each frame, check whether the frame or its bitwise complement is intelligible.
3. If such a frame is identified, proceed to Phase 2.

Phase 2: Recover weak loaded state

Given a weak frame and known frame number (IV):

Table 10 Success probability for various lengths of conversation

No. of frames	Conversation time	Probability of weak frame
2^{14}	1 min 16 s	$2^{-26.36}$
2^{16}	5 min 2 s	$2^{-24.36}$
2^{18}	20 min 6 s	$2^{-22.36}$
2^{20}	1 h 21 min	$2^{-20.36}$
2^{22}	5 h 22 min	$2^{-18.36}$

1. Repeat until weak loaded state is identified:
   - For Cases 1, 2 and 3 in turn, guess the free bits of the loaded state.
   - Use previous or subsequent frame to check correctness of this guess:
     (a) Complement last loaded bit in each register.
     (b) Carry out diffusion phase and generate corresponding keystream.
     (c) If known IV is odd (resp. even), attempt to decrypt previous (resp. subsequent) frame using this keystream.
       - If decrypted successfully, correct loaded state has been found;
       - If not, repeat process for next guess.
2. Use the loaded state and known IV to determine
   (a) the state contents immediately after **key** loading ($NK = T^{-22}NK \oplus MV$); and (optionally)
   (b) the secret key for this conversation.
3. Use the secret key or state contents (NK) and known IVs to decrypt all other frames in the conversation.

From the previous discussion, the probability that Phase 1 succeeds is approximately $2^{-40.36}N$ for a conversation containing N frames. If Phase 1 succeeds, Phase 2 requires us to guess and check up to $2^{23.64}$ loaded states and will always succeed in decrypting the conversation; the secret key K can also be determined, if required. Table 10 presents the probability of success for Phase 1 for various lengths of conversation.

5 Initialisation Flaws in Other Ciphers

In this section, we discuss the existence of flaws in the initialisation processes of some other shift register based stream ciphers. For each flaw examined for A5/1, compression, state convergence, slid pairs and weak key-IV pairs, we identify their existence in these other ciphers and discuss the specific causes of these flaws.

5.1 Compression

As discussed in Sect. 3.1, compression occurs during the loading phase of initialisation in many of the stream ciphers designed before the introduction of TMTO attacks. This occurs whenever the state space is smaller than the key-IV space. Other examples include the CSA-SC [12] and LILI-II [21].

The common scrambling algorithm (CSA) used for DVB is a current industrial standard that exhibits key-IV space compression. The CSA-SC uses a 64-bit key and a 64-bit IV to produce an 89-bit internal state (ignoring redundant storage). Thus there are 2^{128} possible key-IV combinations, but only at most 2^{89} possible loaded states. On average, 2^{39} key-IV pairs correspond to each loaded state, generating the same keystream [11].

The LILI-II stream cipher, based on the NESSIE [35] stream cipher candidate LILI-128, also exhibits compression [14]. If two key-IV pairs (K, IV) and (K', IV') are such that $K \oplus K' = IV \oplus IV' = \{1\}^{128}$, then these two key-IV pairs produce the same loaded state.

By the time of the eSTREAM project [26], larger state spaces were widely accepted as a design requirement. Therefore, compression is not readily seen in more recent proposals. For example, the eSTREAM candidates Trivium [23] and Dragon [20] have much larger state spaces than the key-IV spaces, and hence compression does not occur.

5.2 State Convergence

State convergence is the result of a many-to-one state update function, which can have different causes. It can be due to a non-autonomous clocking mechanism as in the A5/1 stream cipher. However, such clocking mechanism does not necessary lead to state convergence. For example, state convergence does not exist in LILI-II [44].

The interaction among components of the state update function can also result in many-to-one state updates even when the individual components are one-to-one. For example, the SFINKS stream cipher [18] consists of a feedback shift register and an S-Box that injects nonlinearly into several stages of the register. While the register and the S-Box are one-to-one individually, the way in which these two components interact results in many-to-one state updates. This leads to state convergence during the initialisation process, which has been estimated to occur with probability $2^{-6.9}$ at any given clock [3]. In the case of CSA-SC, where compression occurs during the loading phase, the complex state update mechanism leads to further state convergence from the compressed key-IV space during the diffusion phase [11].

If the key-IV loading process involves state updates, state convergence can occur during the loading phase. This is the case for the eSTREAM [26] candidate cipher MICKEY v1 [7]. Although in MICKEY v1 the state space is potentially large

enough to avoid compression, the cipher uses a nonlinear state update function to progressively load the key and IV into the state. This results in a many-to-one state updating and reduces the number of possible loaded states. (Hong and Kim [32] identified this property in the keystream generation process of MICKEY, but it also applies during both phases of initialisation.) The revised version, MICKEY 2.0 [8], also exhibits state convergence, although the designers have deliberately increased the state size in order to reduce the negative effects of this property.

To avoid the problem of state convergence, the state update function should be one-to-one. However, a disadvantage of having a one-to-one state update function is that state recovery leads to efficient key recovery, since the cipher can be clocked backwards from an initial state to uniquely determine the key-IV pair used for the session. Therefore, it is sometimes worthwhile to have a degree of state convergence that is not detrimental to the overall security of a cipher.

5.3 Slid Pairs

For ciphers in which an l-bit key and j-bit IV are transferred directly into an internal state of size s, where $l + j < s$, some stages of S remain unfilled. Padding is required to fill these remaining stages to obtain the loaded state. The choice of padding pattern may affect the probability of occurrence of slid pairs. For some ciphers, it would also result in shifted keystreams for small phase shifts.

Consider the eSTREAM candidates Grain v1 [31], Trivium [23] and SFINKS [7]. Grain v1 has a 80-bit key and 64-bit IV, which is directly loaded into a 160-bit state. This means that 144 bits of padding is necessary. It has been shown that related key-IV pairs can be found such that such pairs produce shifted keystream 12 clocks apart [9]. Similar results were found for variant Grain-128, but not for Grain-128a, where the padding pattern has been modified to avoid these slid pairs. Similarly, Trivium has an 80-bit key and 80-bit IV, which are directly loaded into three registers totalling 288 bits. One of the registers with length 111 consists entirely of padding, which gives rise to slid pairs occurring 111 clocks or more apart [40].

Slid pairs can also be found in the SFINKS stream cipher [18]. Related key-IV pairs exists for 17 or more clocks, although with negligible probability [3]. Furthermore, it has been shown that even a slight modification from the current design can result in slid pairs occurring after one clock with high probability.

5.4 Weak Key-IV Combinations

The various components in ciphers can operate either interdependently or autonomously. Many of the desirable properties of sequences produced by these components hold only under the condition that the components are operating autonomously. For example, cycle lengths of linear feedback shift registers (LFSRs)

can be predetermined and all-zero states would not occur given any non-zero initial state. However, when the register is not autonomous, the possibility of an all-zero state cannot be discounted.

The non-autonomous feedback mechanism may result in an undesirable initial state (session key) at the end of the diffusion phase. It may also result in all-zero contents for the register that uses a non-autonomous feedback function. The key-IV pairs that give rise to non-zero states are called weak key-IV combinations.

The existence of weak key-IV combinations has been discovered in the Grain family of stream ciphers [9]. For example, there are 2^{64} such combinations among a total of 2^{144} possible key-IV pairs for Grain v1. Grain consists of a LFSR and a nonlinear feedback shift register (NFSR). Using any of the weak key-IV combinations results in an all-zero state in the LFSR, which renders the cipher vulnerable to distinguishing attacks on the NFSR. MICKEY v1 [7] also has a small class of weak key-IV pairs [32]. Weak key-IV combinations would also exist theoretically in the case of CSA-SC [12], due to the non-autonomous feedback during its initialisation process [2].

Weak key-IV combinations may also result when non-standard feedback mechanisms are used. The RC4 stream cipher is a famous example of a stream cipher that is not based on feedback shift registers [42]. Due to its special construction, RC4 has been found to be susceptible to suboptimal initialisation with certain key-IV pairs. It has been shown that some key-IV combinations allow recovery of state bits [27]. This weakness was used in a practical cryptanalysis of the wired equivalent privacy (WEP) standard [43].

5.5 Summary

Table 11 summarises the existence of security flaws in the stream ciphers discussed in this section. In this table, ✓ indicates that the flaw does exist in the cipher and x indicates that it does not. A blank entry denotes that, to the authors' knowledge,

Table 11 Security flaws in initialisation of certain stream ciphers

	RC4	*Trivium*	*Grain v1*	*Dragon*	*MICKEY v1*	*LILI-II*	*A5/1*	*SFINKS*	*CSA-SC*
Compression	x	x	x	x	x	✓	✓	x	✓
Convergence		x	x		✓	✓	✓	✓	✓
Slid pairs		✓	✓		✓	[a]	✓	✓	✓
Weak key-IV	✓		✓		✓		✓		[b]

Notes: [a]Theoretically, there are slid pairs that lead to shifted keystreams

[b]Theoretically, weak key-IV pairs exist due to non-autonomous feedback during initialisation

whether or not the flaw exists in the cipher is unknown. While all flaws presented in this paper can be found in A5/1, most other ciphers only possess a subset of these flaws.

For those entries not mentioned in this section, the reader is referred to [2, Table 6.1] for further details.

6 Conclusion

Stream cipher proposals usually include both design specifications, including an initialisation or rekeying process, and an analysis section outlining resistance against generic attacks. The focus of the security analysis is generally on keystream generation. Less attention is paid to the analysis of the initialisation process, though it is considered in some proposals. This paper recommends that stream cipher designers consider carefully the initialisation process, and perform sufficient analysis to ensure that both the loading and diffusion phases of this process are secure against known attacks and avoid the known flaws discussed in this paper.

A well-designed initialisation process (comprising both loading and diffusion phases) should not reveal any information about the secret key, or possess properties that may help to facilitate attacks. The initialisation process should ensure that performing a key recovery attack is hard even if state recovery has occurred, because the mathematical relationships between the key-IV pair and the keystream are hard to establish.

Considering both the A5/1 case study in Sect. 4 and the other examples given in Sect. 5, we provide the following recommendations for initialisation process for stream ciphers:

1. **State, key and IV sizes.** The state size should be larger than the sum of the key and IV lengths, and the IV should be at least the same length as the key. This is necessary to provide resistance to TMTO attacks.
2. **Padding pattern.** The padding pattern should not be a repetitive pattern. Given a total state size which is larger than sum of the key and IV lengths, specifying a means for loading contents into all stages in the state is likely to involve padding. The padding pattern should not be a series of identical values (either all-zeros or all-ones), or consist of repeated copies of a specific pattern (such as 010101 or 001001), as slid pairs for small phase shifts are readily found in such cases. Avoiding repetitive patterns will defer the occurrence of slid pairs to the maximum possible shift, increase the complexity of finding slid pairs and reduce the probability of obtaining phase shifted keystreams for small shift values.
3. **One-to-one functions.** The state update function should be one-to-one. Both the individual components and the combination of these components should result in an update function that is one-to-one. This is required to prevent state convergence. Conversely, the use of a one-to-one function leads to another

problem: in the case of state recovery, it may be possible to clock back to the loaded state and effect key recovery.

4. **Dissimilar state update functions.** The similarity of the state update functions within and between the initialisation and keystream generation processes should be reduced. This should reduce the occurrence of slid pairs and the corresponding shifted keystreams for small shift values. However, implementing this recommendation may have a negative effect on efficiency.
5. **Non-autonomous feedback functions.** The use of non-autonomous feedback functions during either the loading or diffusion phases requires careful consideration to prevent key-IV combinations resulting in weak session keys. Weak session keys result in ineffective components; for that key-IV pair the keystream generator is equivalent to another design with a smaller state. This may leave the cipher vulnerable to attacks.
6. **Nonlinear diffusion process.** The diffusion process should ensure that both the key and IV bits are distributed across the entire state, and combined in a non-linear way. An appropriate diffusion process provides resistance against differential and fault attacks.
7. **Optimal number of iterations.** The number of iterations of the state update function to be performed during the diffusion phase requires careful consideration. Increasing the number of iterations may increase the resistance to some attacks (algebraic, differential, correlation) but this may leave the cipher vulnerable to other potential attacks (such as TMTO attacks, if state convergence is present). Increasing the number of iterations also decreases the efficiency of the cipher. This process is a trade-off between the security aspects and efficiency.

References

1. Alhamdan, A.: A study of the initialisation process of the A5/1 stream cipher. Master's Thesis, Queensland University of Technology, October 2008
2. Alhamdan, A.A.: Secure stream cipher initialisation processes. Ph.D. Thesis, Queensland University of Technology (2014)
3. Alhamdan, A., Bartlett, H., Simpson, L., Dawson, E., Wong, K.K.-H.: State convergence in the initialisation of the sfinks stream cipher. In: Pieprzyk, J., Thomborson, C. (eds.) 10th Australasian Information Security Conference (AISC 2012), Volume 125 of Conference in Research and Practice in Information Technology (CRPIT), pp. 27–32. Australian Computer Society, Melbourne (2012)
4. Alhamdan, A., Bartlett, H., Dawson, E., Simpson, L., Wong, K.K.: Slid pairs in the initialisation of the a5/1 stream cipher. In: Thomborson, C., Parampalli, U. (eds.) Information Security 2013 (AISC 2013), Volume 138 of CRPIT, pp. 3–12. ACS, Adelaide (2013)
5. Alhamdan, A., Bartlett, H., Dawson, E., Simpson, L., Wong, K.K.: Weak key-iv pairs in the a5/1 stream cipher. In: Parampalli, U., Welch, I. (eds.) The 12th Australasian Information Security Conference (AISC 2014), Volume 149 of CRPIT, pp. 23–36. ACS, Auckland (2014)
6. Arsham, H.: Performance extrapolation in discrete-event systems simulation. Int. J. Syst. Sci. **27**(9), 863–869 (1996)

7. Babbage, S., Dodd, M.: The stream cipher MICKEY (version 1). eSTREAM, ECRYPT Stream Cipher Project, Report 2005/015. http://www.ecrypt.eu.org/stream (2005)
8. Babbage, S., Dodd, M.: The stream cipher MICKEY 2.0. eSTREAM, ECRYPT Stream Cipher Project. http://www.ecrypt.eu.org/stream/p3ciphers/mickey/mickey_p3.pdf (2006)
9. Banik, S., Maitra, S., Sarkar, S.: Some results on related key-IV pairs of grain. In: Bogdanov, A., Sanadhya, S. (eds.) Security, Privacy, and Applied Cryptography Engineering. Lecture Notes in Computer Science, pp. 94–110. Springer, Berlin/Heidelberg (2012)
10. Barkan, E., Biham, E., Keller, N.: Instant ciphertext-only cryptanalysis of GSM encrypted communication. In: Boneh, D. (ed.) Advances in Cryptology - CRYPTO 2003. Lecture Notes in Computer Science, vol. 2729, pp. 600–616. Springer, Berlin/Heidelberg (2003)
11. Bartlett, H., Al-Hamdan, A., Simpson, L., Dawson, E., Wong, K.K.-H.: Weaknesses in the initialisation process of the common scrambling algorithm stream cipher. In: Schmidt, K.-U., Winterhof, A. (eds.) Sequences and Their Applications (SETA 2014). Lecture Notes in Computer Science, vol. 8865, Melbourne. Springer, New York (2014)
12. Bewick., S.: Descrambling DVB data according to ETSI common scrambling standard. UK Patent GB2322995A (1998)
13. Biham, E., Dunkelman, O.: Cryptanalysis of the A5/1 GSM stream cipher. In: Roy, B., Okamoto, E. (eds.) Progress in Cryptology - INDOCRYPT 2000. Lecture Notes in Computer Science, vol. 1977, pp. 43–51. Springer, Berlin/Heidelberg (2000)
14. Biham, E., Dunkelman, O.: Differential cryptanalysis in stream ciphers. Cryptology ePrint Archive, Report 2007/218. http://www.eprint.iacr.org/ (2007)
15. Biryukov, A., Shamir, A.: Cryptanalytic time/memory/data tradeoffs for stream ciphers. In: Okamoto, T. (ed.) Advances in Cryptology - ASIACRYPT 2000. Lecture Notes in Computer Science, vol. 1976, pp. 1–13. Springer, Berlin/Heidelberg (2000)
16. Biryukov, A., Wagner, D.: Slide attacks. In: Knudsen, L. (ed.) Fast Software Encryption. Lecture Notes in Computer Science, vol. 1636, pp. 245–259. Springer, Berlin/Heidelberg (1999)
17. Biryukov, A., Shamir, A., Wagner, D.: Real time cryptanalysis of A5/1 on a PC. In: Goos, G., Hartmanis, J., Leeuwen, J., Schneier, B. (eds.) Fast Software Encryption. Lecture Notes in Computer Science, vol. 1978, pp. 1–18. Springer, Berlin/Heidelberg (2001)
18. Braeken, A., Lano, J., Mentens, N., Preneel, B., Verbauwhede, I.: SFINKS: a synchronous stream cipher for restricted hardware environments. eSTREAM, ECRYPT Stream Cipher Project, Report 2005/026. www.ecrypt.eu.org/stream/ciphers/sfinks/sfinks.ps (2005)
19. Briceno, M., Goldberg, I., Wagner, D.: A pedagogical implementation of A5/1. http://www.cryptome.org/jya/a51-pi.htm (1999)
20. Chen, K., Henricksen, M., Millan, W., Fuller, J., Simpson, L., Dawson, E., Lee, H.J., Moon, S.J.: Dragon: a fast word based stream cipher. In: Park, C., Chee, S. (eds.) Information Security and Cryptology - ICISC 2004. Lecture Notes in Computer Science, vol. 3506, pp. 33–50. Springer, Berlin/Heidelberg (2005)
21. Clark, A., Dawson, E., Fuller, J., Golić, J., Lee, H.-J., Millan, W., Moon, S.-J., Simpson, L.: The LILI-II keystream generator. In: Batten, L., Seberry, J. (eds.) Information Security and Privacy. Lecture Notes in Computer Science, vol. 2384, pp. 25–39. Springer, Berlin/Heidelberg (2002)
22. Dawson, E., Nielsen, L.: Automated cryptanalysis of XOR plaintext strings. Cryptologia **20**(2), 165–181 (1996)
23. De Cannière, C., Preneel, B.: Trivium - a stream cipher construction inspired by block cipher design principles. eSTREAM, ECRYPT Stream Cipher Project, Report 2005/030 and 2006/021. http://www.ecrypt.eu.org/stream (2005)
24. De Cannière, C., Kücük, Ö., Preneel, B.: Analysis of Grain's initialization algorithm. In: Vaudenay, S. (ed.) Progress in Cryptology - AFRICACRYPT 2008. Lecture Notes in Computer Science, vol. 5023, pp. 276–289. Springer, Berlin/Heidelberg (2008)
25. Dunkelman, O., Keller, N.: Treatment of the initial value in time-memory-data tradeoff attacks on stream ciphers . Inf. Process. Lett. **107**(5), 133–137 (2008)
26. European Network of Excellence for Cryptology: The eSTREAM project. (2004) http://www.ecrypt.eu.org/stream/

27. Fluhrer, S., Mantin, I., Shamir, A.: Weaknesses in the key scheduling algorithm of RC4. In: Vaudenay, S., Youssef, A. (eds.) Selected Areas in Cryptography. Lecture Notes in Computer Science, vol. 2259, pp. 1–24. Springer, Berlin/Heidelberg (2001)
28. Gendrullis, T., Novotnỳ, M., Rupp, A.: A real-world attack breaking A5/1 within hours. In: Oswald, E., Rohatgi, P. (eds.) Cryptographic Hardware and Embedded Systems - CHES 2008. Lecture Notes in Computer Science, vol. 5154 , pp. 266–282. Springer, Berlin/Heidelberg (2008)
29. Golić, J.: Cryptanalysis of alleged A5 stream cipher. In: Fumy, W. (ed.) Advances in Cryptology - EUROCRYPT '97. Lecture Notes in Computer Science, vol. 1233, pp. 239–255. Springer, Berlin/Heidelberg (1997)
30. Golić, J.: Cryptanalysis of three mutually clock-controlled stop/go shift registers. IEEE Trans. Inf. Theory **46**(3), 1081–1090 (2000)
31. Hell, M., Johansson, T., Meier, W.: Grain: a stream cipher for constrained environments. Int. J. Wirel. Mob. Comput. **2**(1), 86–93 (2007). http://www.ecrypt.eu.org/stream
32. Hong, J., Kim, W.-H.: Tmd-tradeoff and state entropy loss considerations of streamcipher mickey. In: Proceedings of the 6th International Conference on Cryptology in India, INDOCRYPT'05, pp. 169–182. Springer, Berlin/Heidelberg (2005)
33. Hong, J., Sarkar, P.: New applications of time memory data tradeoffs. In: Roy, B. (ed.) Advances in Cryptology - ASIACRYPT 2005. Lecture Notes in Computer Science, vol. 3788, pp. 353–372. Springer, Berlin/Heidelberg (2005)
34. Hong, J., Sarkar, P.: Rediscovery of time memory tradeoffs. Cryptology ePrint Archive, Report 2005/090. http://www.eprint.iacr.org/ (2005)
35. Information Society Technologies (IST) Programme: NESSIE. http://www.cosic.esat.kuleuven.be/nessie/. Accessed 20 May 2010
36. Kiselev, S.A., Tokareva, N.N.: Reduction of the key space of the cipher A5/1 and invertibility of the next-state function for a stream generator. J. Appl. Ind. Math. **6**(2), 194–202 (2012)
37. Küçük, Ö.: Slide resynchronization attack on the initialization of Grain 1.0. eSTREAM, ECRYPT Stream Cipher Project, Report 2006/044. http://www.ecrypt.eu.org/stream (2006)
38. Lee, Y., Jeong, K., Sung, J., Hong, S.: Related-key chosen IV attacks on Grain-v1 and Grain-128. In: Mu, Y., Susilo, W., Seberry, J. (eds.) Information Security and Privacy. Lecture Notes in Computer Science, vol. 5107, pp. 321–335. Springer, Berlin/Heidelberg (2008)
39. Lidl, R., Niederreiter, H.: Finite Fields, vol. 20. Cambridge University Press, Cambridge (1997)
40. Priemuth-Schmid, D., Biryukov, A.: Slid Pairs in Salsa20 and trivium. In: Chowdhury, D., Rijmen, V., Das, A. (eds.) Progress in Cryptology - INDOCRYPT 2008. Lecture Notes in Computer Science, vol. 5365, pp. 1–14. Springer, Berlin/Heidelberg (2008)
41. Rose, G., Hawkes, P.: On the applicability of distinguishing attacks against stream ciphers. Cryptology ePrint Archive, Report 2002/142. http://www.eprint.iacr.org/ (2002)
42. Stallings, W.: Cryptography and Network Security, 4th edn. Pearson Prentice Hall, Upper Saddle River (2006)
43. Stubblefield, A., Ioannidis, J., Rubin, A.D.: Using the Fluhrer, Mantin, and Shamir attack to break WEP. In: Network and Distributed Systems Security Symposium (NDSS), vol. 1722. Citeseer (2002)
44. Teo, S.: Analysis of nonlinear sequences and stream ciphers. Ph.D. Thesis, Queensland University of Technology (2013)
45. Teo, S.-G., Al-Hamdan, A., Bartlett, H., Simpson, L., Wong, K.K.-H., Dawson, E.: State convergence in the initialisation of stream ciphers. In: Parampalli, U., Hawkes, P. (eds.) Information Security and Privacy. Lecture Notes in Computer Science, vol. 6812, pp. 75–88. Springer, Berlin/Heidelberg (2011)
46. Wardlaw, W.P.: A matrix model for the linear feedback shift register. Dtic Document, Naval Research Lab, Washington, DC (1989)
47. Weinmann, R., Wirt, K.: Analysis of the DVB common scrambling algorithm. In: Chadwick, D., Preneel, B. (eds.) Communications and Multimedia Security. IFIP - The International Federation for Information Processing, vol. 175, pp. 195–207. Springer, New York (2005)

48. Wu, H., Preneel, B.: Resynchronization attacks on WG and LEX. In: Robshaw, M. (ed.) Fast Software Encryption. Lecture Notes in Computer Science, vol. 4047, pp. 422–432. Springer, Berlin/Heidelberg (2006)
49. Zhang, H., Wang, X.: Cryptanalysis of stream cipher grain family. Cryptology ePrint Archive, Report 2009/109. http://www.eprint.iacr.org/ (2009)

Producing Fuzzy Inclusion and Entropy Measures

Athanasios C. Bogiatzis and Basil K. Papadopoulos

Abstract Inclusion and entropy measurements are significant for a variety of applications in fuzzy logic area. Several authors and researchers have tried to axiomatize fuzzy inclusion and entropy indicators. Others have introduced such measures based on specific desired properties. Significant results have been obtained; results that have led to a number of alternative solutions concerning several different applications. Apart from these interesting and innovative ideas, open matters of further discussion and research have occurred in these studies as well. Following the work of these authors, we propose an alternative axiomatization of fuzzy inclusion based on an already existing one. This allows us to introduce a category of subsethood and entropy measures which contains well-known indicators as well as new ones.

Keywords: Inclusion grade • Fuzzy entropy • Fuzzy implications • Fuzzy subsethood

1 Introduction

There are two well-known axiomatizations concerning fuzzy inclusion. The first one belongs to Sihna and Dougherty [13]. It has been proven that it can be reconstructed with less axioms [2, 3] and that it can be covered by Willmott's axioms, which were earlier introduced in [15, 16]. The second axiomatization of fuzzy inclusion was proposed by Young [17]. Young disagrees with some of Sihna–Dougherty's properties and chooses her axioms from a different point of view (in order to relate them with fuzzy entropy as a continuity of Kosko's work [9, 10]).

Fuzzy subsethood measures are mainly obtained by fuzzy implications, something consistent with classic logic. Thus, they are mainly the infimum (satisfying S-D's properties) or the mean value (i.e., Goguen's inclusion measure [6]) of an

A.C. Bogiatzis (✉) • B.K. Papadopoulos
Department of Civil Engineering, Section of Mathematics and Informatics, Democritus University of Thrace, Vas. Sofias 12, Xanthi 67100, Greece
e-mail: abogiatz@hotmail.com; papadob@civil.duth.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_3

implication operator ([7], [8]). Authors like Bandler and Kohout [1] or Willmott [15, 16] deal with several such measures. There is also Kosko's subsethood measure which doesn't seem to belong to either of these categories. Later, we will see how it is included in our category of inclusion measures.

The main goals of this paper are the proposal of an alternative axiomatization of fuzzy inclusion (based on Young's original one), the introduction of a formula for producing subsethood and their corresponding entropy measures, and the production of such indicators (new or already known). There is also a first comparison between them through some specific examples and graphs.

In Sect. 2, we briefly remind the axioms of fuzzy intersections and implications as presented in [18] and [14]. Next, we remind the axioms of fuzzy entropy measures, as presented by De Luca and Termini [4], and some main parts of Young's work. We also give some basic concepts of Kosko's work [9–11] and its relation with Young's research. In Sect. 3, we present our alternative axiomatization and a formula for producing fuzzy inclusion and entropy measures. We produce some new possible measures as well. Section 4 consists of specific examples on these inclusion operators and a brief comparison of their corresponding results. In Sect. 5, we do the same for the entropy measures. These examples are accompanied by some figures as well. We conclude in Sect. 6 with some matters of additional discussion and future research.

2 Preliminaries and Notation

2.1 Basic Notation

We mainly keep Young's notation which is quite common in the literature. So, let X denote the universal set (which is a finite set) and $F(X)$ its power set. Members of $F(X)$ are represented by capital letters A, B, C, etc., whereas their membership functions are denoted as m_A, m_B, m_C, respectively. Set A will be a refinement of set B if $m_A(x) \leqslant m_B(x)$ when $m_B(x) \leqslant \frac{1}{2}$ and $m_A(x) \geqslant m_B(x)$ when $m_B(x) \geqslant \frac{1}{2}$. P will be the fuzzy set whose membership function is equal to $\frac{1}{2}$, for all $x \in X$. The cardinality of a fuzzy set A will be $|A| = \sum_{x \in X} m_A(x)$.

Furthermore, we call A a subset of B (we write $A \subseteq B$) if $m_A(x) \leqslant m_B(x)$ for all $x \in X$. Superscript c stands for the standard fuzzy complement, meaning that the membership function of A^c is $1 - m_A(x)$. Finally A_{near} will denote the set whose membership value of $x \in X$ is 0 if $m_A(x) < \frac{1}{2}$ and 1 otherwise. A_{far} will be the complement of A_{near}.

2.2 Fuzzy Intersections and Implications

Next, we recall the definition and basic properties of fuzzy intersections and implications:

Definition 1. A function $T : [0, 1] \times [0, 1] \to [0, 1]$ is a fuzzy intersection (t-norm) if and only if it satisfies the following conditions for all $a, b, d \in [0, 1]$:

$T(a, 1) = a$ (boundary condition) **(*t*1)**
$b \leq d \Rightarrow T(a, b) \leq T(a, d)$ (monotonicity) **(*t*2)**
$T(a, b) = T(b, a)$ (commutativity) **(*t*3)**
$T(a, T(b, d)) = T(T(a, b), d)$ (associativity) **(*t*4)**

Three of the most important additional requirements, which restrict the class of fuzzy intersections, are the following:

$T(a, a) < a$ (subidempotency) **(*t*5)**
$a_1 < a_2$ and $b_1 < b_2 \Rightarrow T(a_1, b_1) < T(a_2, b_2)$ (strict monotonicity) **(*t*6)**
T is a continuous function (continuity) **(*t*7)**

A subidempotent, continuous t-norm is called Archimedean; if it is also strictly monotonous, it is called strict Archimedean. An additional property of all fuzzy intersections that will be needed later in the paper is that: $T(a, 0) = 0$

Definition 2. A function $I : [0, 1] \times [0, 1] \to [0, 1]$ is a fuzzy implication if and only if it satisfies the following conditions for all $a, b, d \in [0, 1]$:

$a \leq b \Rightarrow I(a, d) \geq I(b, d)$ (monotonicity in first argument) **(*i*1)**
$a \leq b \Rightarrow I(d, a) \leq I(d, b)$ (monotonicity in second argument) **(*i*2)**
$I(0, a) = 1$ (dominance of falsity) **(*i*3)**
$I(1, b) = b$ (neutrality of truth) **(*i*4)**
$I(a, a) = 1$ (identity) **(*i*5)**
$I(a, I(b, d)) = I(b, I(a, d))$ (exchange property) **(*i*6)**
$I(a, b) = 1 \Leftrightarrow a \leq b$ (boundary condition) **(*i*7)**
$I(a, b) = I(c(b), c(a))$ for a fuzzy complement c **(*i*8)**
I is a continuous function (continuity) **(*i*9)**

but not necessarily all of them—different applications demand different additional attributes of I.

(Not all these axioms are independent. For instance, axioms (i3) and (i5) can be derived from axiom (i7) but not vice versa. Anyway we mainly deal with axioms (i1), (i2), and (i7) in this paper.)

2.3 Fuzzy Entropy

According to De Luca and Termini:

Definition 3. A fuzzy entropy measure is a function $E : F(X) \rightarrow [0, 1]$ which satisfies the following:

$E(A) = 0$ if and only if $m_A(x) \in \{0, 1\}$ for all $x \in X$ (***E*1**)
$E(A) = 1$ if and only if $A = P$ (***E*2**)
$E(A) \leqslant E(B)$ if A is a refinement of B (***E*3**)
$E(A) = E(A^c)$ (***E*4**)

As Young states, these axioms have been further studied and supplemented by other authors [5, 12]. In this paper, we will refer only to these four properties.

2.4 Young's Axioms and Theorem

In [9], Kosko shows that the entropy of a fuzzy set A is the degree to which $A \cup A^c$ is a subset of its complement $A \cap A^c$. So, after examining the axioms of Sihna and Dougherty and partially disagreeing with them, Young presents three axioms for fuzzy inclusion which are sufficient to lead to entropy measures according to Kosko's proposition. Considering S is a function defined as $S : F(X) \times F(X) \rightarrow I$, these three axioms are:

- $S(A, B) = 1$ if and only if $A \subseteq B$ in Zadeh's sense (***S*1**)
- If $P \subseteq A$ in Zadeh's sense, then $S(A, A^c) = 0 \Leftrightarrow A = X$ if and only if $A = X$ (***S*2**)
- If $B \subseteq A_1 \subseteq A_2$, then $S(A_1, B) \geqslant S(A_2, B)$ (***S*3**)

  and if $B_1 \subseteq B_2$ then $S(A, B_1) \leqslant S(A, B_2)$
  (From now on, let ***S*3*a*** denote the first half of ***S*3** and ***S*3*b*** the second one).
  The following theorem is then presented:

Theorem 1. *If S is a fuzzy subsethood measure (meaning a function satisfying* ***S*1**–***S*3***), then E defined as:*

$$E(A) = S(A \cup A^c, A \cap A^c),\ A \in F(X)$$

is a fuzzy entropy measure of fuzzy set A, where:

$$m_{A \cup A^c}(x) = \max(m_A(x), 1 - m_A(x)) \text{ and } m_{A \cap A^c}(x) = \min(m_A(x), 1 - m_A(x))$$

This way, we have an alternative axiomatization of fuzzy subsethood and a theorem which "allows" every inclusion measure to produce a corresponding entropy measure. In this category belong subsethood measures such as Kosko's:

$$S_K(A,B) = \begin{cases} \frac{\sum_{x\in X} \min(m_A(x), m_B(x))}{\sum_{x\in X} m_A(x)} = \frac{|A\cap B)|}{|A|} & \text{if } A \neq \varnothing \\ 1 & \text{if } A = \varnothing \end{cases}$$

or Goguen's inclusion grade:

$$S_I(A,B) = \frac{1}{n}\sum_{x\in X} \min(1, 1 - m_A(x) + m_B(x)), \quad n = |X|$$

which is actually the mean value of Lukasiewicz's implication. Their corresponding entropy measures are:

$$E_K(A) = \frac{\sum_{x\in X} \min(1 - m_A(x), m_A(x))}{\sum_{x\in X} \max(1 - m_A(x), m_A(x)}$$

and

$$E_I(A) = \frac{2}{n}\sum_{x\in X} \min(m_A(x), 1 - m_A(x)),$$

respectively.

3 Our Proposition

3.1 Basic Idea

Our basic idea was to introduce a category of subsethood measures based upon the following:

Proposition 1. *Let $A, B \in F(X)$, and $S : F(X)\times F(X) \to I$ be a function defined as:*

$$S(A,B) = \begin{cases} \frac{\sum_{x\in X} T(I(m_A(x), m_B(x)), m_A(x))}{\sum_{x\in X} m_A(x)} & \text{if } A \neq \varnothing \\ 1 & \text{if } A = \varnothing \end{cases}$$

where T and I denote a fuzzy intersection and a fuzzy implication, respectively. If:

1. *Implication I satisfies axioms (i1), (i2), (i7) and property:*

   For all $x \in X$, If $m_A(x) \geq \frac{1}{2}$, then
   $I(m_A(x), 1 - m_A(x)) = 0 \Leftrightarrow m_A(x) = 1$ **(i∗)**

2. *T and I (when combined) satisfy the following properties, for every* $x \in X$ *:*

$$T(I(m_A(x), m_B(x)), m_A(x)) = m_A(x) \Leftrightarrow m_A(x) \leqslant m_B(x) \quad \textbf{(ti1)}$$

$$m_B(x) \leq m_{A_1}(x) \leq m_{A_2}(x) \Rightarrow \\ T(I(m_{A_1}(x), m_B(x)), m_{A_1}(x)) \geq T(I(m_{A_2}(x), m_B(x)), m_{A_2}(x)) \quad \textbf{(ti2)}$$

then S is a subsethood measure of set A into set B.

(*These conditions are sufficient but not always necessary*)

Proof. **(S1):** It's obvious for $A = \varnothing$. Let $A \neq \varnothing$ and $S(A, B) = 1$. Then:

$$S(A, B) = 1 \Leftrightarrow$$

$$\sum T(I(m_A(x), m_B(x)), m_A(x)) = \sum_{x \in X} m_A(x) \Leftrightarrow$$

(since $T(I(m_A(x), m_B(x)), m_A(x)) \leqslant m_A(x)$)

$$T(I(m_A(x), m_B(x)), m_A(x)) = m_A(x) \text{ for every } x \in X \overset{\text{property } \textbf{ti1}}{\Leftrightarrow}$$

$$m_A(x) \leqslant m_B(x) \text{ for every } x \in X \Leftrightarrow$$

$$A \subseteq B$$

(S2): If $P \subseteq A$ or $m_A(x) \geqslant \frac{1}{2}$ for all $x \in X$, then:

$$S(A, A^c) = 0 \Leftrightarrow$$

$$T(I(m_A(x), 1 - m_A(x)), m_A(x)) = 0 \text{ for every } x \in X \overset{m_A(x) \neq 0}{\Leftrightarrow}$$

$$I(m_A(x), 1 - m_A(x)) = 0 \text{ for every } x \in X \overset{\text{property } i*}{\Leftrightarrow}$$

$$m_A(x) = 1 \text{ for every } x \in X \Leftrightarrow$$

$$A = X$$

(axiom ***i*4** is sufficient but not necessary for I to have property ***i****)

(S3): As far as property ***S*3*a*** is concerned:

$$B \subseteq A_1 \subseteq A_2 \Rightarrow m_B(x) \leqslant m_{A_1}(x) \leqslant m_{A_2}(x)) \text{ for all } x \in X \overset{\text{property } \textbf{ti2}}{\Rightarrow}$$

$$T(I(m_{A_1}(x), m_B(x)), m_{A_1}(x)) \geq T(I(m_{A_2}(x), m_B(x)), m_{A_2}(x)) \text{ for all } x \in X \Rightarrow$$

$$\sum_{x\in X} T(I(m_{A_1}(x), m_B(x)), m_{A_1}(x)) \geq \sum_{x\in X} T(I(m_{A_2}(x), m_B(x)), m_{A_2}(x))$$

and since $\sum_{x\in X} m_{A_1}(x) \leq \sum_{x\in X} m_{A2}(x)$ we'll have:

$$\frac{\sum_{x\in X} T(I(m_{A_1}(x), m_B(x)), m_{A_1}(x))}{\sum_{x\in X} m_{A_1}(x)} \geqslant \frac{\sum_{x\in X} T(I(m_{A_2}(x), m_B(x)), m_{A_2}(x))}{\sum_{x\in X} m_{A2}(x)} \Rightarrow$$

$$S(A_1, B) \geqslant S(A_2, B)$$

For ***S3b*** we have:

$$B_1 \subseteq B_2 \Rightarrow m_{B_1}(x) \leqslant m_{B_2}(x) \overset{\text{axiom } i2}{\Rightarrow}$$

$$I(m_A(x), m_{B_1}(x)) \leqslant I(m_A(x), m_{B_2}(x)) \overset{\text{axiom } i2}{\Rightarrow}$$

$$T(I(m_A(x), m_{B_1}(x)), m_A(x)) \leqslant T(I(m_A(x), m_{B_2}(x)), m_A(x)) \Rightarrow$$

$$\frac{\sum_{x\in X} T(I(m_A(x), m_{B_1}(x)), m_{B_1}(x))}{\sum_{x\in X} m_A(x)} \leqslant \frac{\sum_{x\in X} T(I(m_A(x), m_{B_2}(x)), m_{B_2}(x))}{\sum_{x\in X} m_A(x)} \Rightarrow$$

$$S(A, B_1) \leqslant S(A, B_2)$$

For the rest of the paper, let $\tilde{I}$ denote the set of all implications that satisfy ***i1***, ***i2***, ***i7***, and ***i**. Most common implications satisfy ***i1***, ***i2***, and ***i**. Some of these, like:

$$I_{\text{God}}(m_A(x), m_B(x)) = \begin{cases} 1 & \text{if } m_A(x) \leq m_B(x) \\ m_B(x) & \text{if } m_A(x) > m_B(x) \end{cases} \quad \text{(Gödel 's)}$$

$$I_{\text{Gog}}(m_A(x), m_B(x)) = \min(1, \tfrac{m_B(x)}{m_A(x)}) \quad \text{(Goguen's)}$$

$$I_{\text{Luc}}(m_A(x), m_B(x)) = \min(1, m_{A^c}(x) + m_B(x)) \quad \text{(Lucasiewichz's)}$$

$$I_{\text{Wu}}(m_A(x), m_B(x)) = \begin{cases} 1 & \text{if } m_A(x) \leq m_B(x) \\ \min(m_{A^c}(x), m_B(x)) & \text{if } m_A(x) > m_B(x) \end{cases} \quad \text{(Wu's)}$$

satisfy axiom ***i7*** as well, whereas others like:

$$I_{\text{Z}}(m_A(x), m_B(x)) = \max(m_{A^c}(x), \min(m_A(x), m_B(x))) \quad \text{(Zadeh's)}$$

$$I_{\text{KD}}(m_A(x), m_B(x)) = \max(m_{A^c}(x), m_B(x)) \quad \text{(Kleene-Dienes')}$$

$$I_{\text{R}}(m_A(x), m_B(x)) = 1 - m_A(x) + m_A(x)m_B(x) \quad \text{(Reichenbach's)}$$

$$I_{\text{Y}}(m_A(x), m_B(x)) = [m_B(x)]^{m_A(x)} \quad \text{(Yager's)}$$

don't. However, these can be easily adjusted so that they satisfy axiom ***i7*** in the following sense:

$$I_{\mathrm{Z}}^{*} = I_{\mathrm{KD}}^{*}(m_A(x), m_B(x)) = \begin{cases} 1 & \text{if } m_A(x) \leq m_B(x) \\ \max(m_{A^c}(x), m_B(x)) & \text{if } m_A(x) > m_B(x) \end{cases}$$

$$I_{\mathrm{R}}^{*}(m_A(x), m_B(x)) = \begin{cases} 1 & \text{if } m_A(x) \leq m_B(x) \\ 1 - m_A(x) + m_A(x)m_B(x) & \text{if } m_A(x) > m_B(x) \end{cases}$$

$$I_{\mathrm{Y}}^{*}(m_A(x), m_B(x)) = \begin{cases} 1 & \text{if } m_A(x) \leq m_B(x) \\ [m_B(x)]^{m_A(x)} & \text{if } m_A(x) > m_B(x) \end{cases}$$

3.2 R-Implications and Kosko's Measure

R-implications (along with Mamdani's and Sugeno's implications) are commonly used in fuzzy logic controllers. *R*-implications are produced by the following formula:

$$I(m_A(x), m_B(x)) = \sup\{s \in [0, 1] / T(m_A(x), s) \leq m_B(x)\}$$

By using different *t*-norms T in this formula, we obtain different implications. Goguen's implication is obtained by using *t*-norm T_P:

$$T_P(m_A(x), m_B(x)) = m_A(x) \cdot m_B(x)$$

Lukasiewicz's implication is obtained when we use bounded difference T_{BD}:

$$T_{\mathrm{BD}}(m_A(x), m_B(x)) = \max(0, m_A(x) + m_B(x) - 1)$$

Gödel 's implication is obtained when we use *t*-norm T_m:

$$T_m(m_A(x), m_B(x)) = \min(m_A(x), m_B(x))$$

T_P, T_{BD}, and T_m satisfy axioms ***t*1**, ***t*2** and I_{Gog}, I_{Luc} and I_{God} belong into $\tilde{I}$. Moreover, combinations (T_P, I_{Gog}), $(T_{\mathrm{BD}}, I_{\mathrm{Luc}})$, and (T_m, I_{God}) satisfy properties ***ti*1** and ***ti*2** and can be used in formula of Proposition 1. However, all three combinations return nothing else but Kosko's subsethood measure $S_K(A, B) = \frac{|A \cap B|}{|A|}$.

Also, other fuzzy operators like Mamdani's or Sugeno's—which cannot be considered as implications from a mathematical point of view and of course do not satisfy ***i*1** or ***i*7**—can be used in the formula. Specifically, if we combine Mamdani's operator with T_m, we obtain S_K. Furthermore, Mamdani's and Sugeno's implications could be applied along with T_P in the formula of Proposition 1 and give us possible inclusion measures. That's why we outlined that conditions of Proposition 1 are sufficient yet not always necessary. Function I could be any "proper" fuzzy operator and not necessarily an implication. But this is something we'll discuss in a future

research. In this paper, we'll only use fuzzy implications (or operators satisfying the basic axioms of fuzzy implications, something consistent with classic logic) and we'll present some basic results. Similar discussion could be made concerning operators T as well. Nevertheless, we will only use t-norms.

From what we've seen so far, the main question that rises is: "Can we produce any other than Kosko's inclusion and entropy measure using Proposition 1 and Theorem 1?" It seems as if Kosko's subsethood measure is the only one (apart from measures based on the mean value of an implication) consistent with Young's work in the sense we described above. But that's because Young's axiom S3 is in fact stricter than needed.

3.3 An Alternative Axiomatization

Axiom ***S3*** is connected with axiom ***E3*** in the following way:

If A is a refinement of B, then $A \cap A^c \subseteq B \cap B^c \subseteq B \cup B^c \subseteq A \cup A^c$ and:

$$E(A) = S(A \cup A^c, A \cap A^c) \overset{(S3a)}{\leqslant} S(B \cup B^c, A \cap A^c)$$

$$\overset{(S3b)}{\leqslant} S(B \cup B^c, B \cap B^c) = E(B)$$

What we see for property ***S3a*** is that it is stronger than needed in order for the above to be valid. In fact ***S3*** can be replaced by the following:

$$\begin{array}{l}\text{If } A_2^c \subseteq A_1^c \subseteq A_1 \subseteq A_2, \text{ then } S(A_1, A_2^c) \geqslant S(A_2, A_2^c) \\ \text{and if } B_1 \subseteq B_2 \text{ then } S(A, B_1) \leqslant S(A, B_2)\end{array} \qquad (\boldsymbol{S3*})$$

without invalidating the rest of Young's work (it goes without saying that ***S3*** $\Rightarrow$ ***S3****).

So using ***S1***, ***S2***, ***S3**** as an axiomatization of fuzzy inclusion, Young's theorem is still valid and we can obtain new subsethood and their corresponding entropy measures using the following:

Proposition 2. *Let $A, B \in F(X)$, and $S : F(X) \times F(X) \to I$ a function defined as:*

$$S(A,B) = \begin{cases} \dfrac{\sum_{x \in X} T(I(m_A(x), m_B(x)), m_A(x))}{\sum_{x \in X} m_A(x)} & \text{if } A \neq \varnothing \\ 1 & \text{if } A = \varnothing \end{cases}$$

where T and I denote a fuzzy intersection and a fuzzy implication, respectively. If:

1. *Implication I satisfies axioms* ***i1***, ***i2***, ***i7*** *and property:*

$$\begin{array}{l}\text{For all } x \in X, \text{ If } m_A(x) \geqslant \frac{1}{2}, \text{ then} \\ I(m_A(x), 1 - m_A(x)) = 0 \Leftrightarrow m_A(x) = 1\end{array} \qquad (\boldsymbol{i*})$$

2. *T and I (when combined) satisfy the following properties for every* $x \in X$*:*

$$T(I(m_A(x), m_B(x)), m_A(x)) = m_A(x) \Leftrightarrow m_A(x) \leqslant m_B(x) \quad \textbf{(ti1)}$$

$$1 - m_{A_2}(x) \leq 1 - m_{A_1}(x) \leq m_{A_1}(x) \leq m_{A_2}(x) \Rightarrow \\ T(I(m_{A_1}(x), m_B(x)), m_{A_1}(x)) \geq T(I(m_{A_2}(x), m_B(x)), m_{A_2}(x)) \quad \textbf{(ti2*)}$$

then S is a subsethood measure of set A into set B and function $E : F(X) \to I$ *defined as:*

$$E(A) = S(A \cup A^c, A \cap A^c)$$

is a fuzzy entropy measure of set A.

Proof. What have changed since Proposition 1 are axiom ***S3*** to ***S3**** and property ***ti2*** to ***ti2****. So, we'll just prove the first part of ***S3****:

$$A_2^c \subseteq A_1^c \subseteq A_1 \subseteq A_2 \Rightarrow$$

$$1 - m_{A_2}(x) \leqslant 1 - m_{A_1}(x) \leqslant m_{A_1}(x) \leqslant m_{A_2}(x) \text{ for all } x \in X \overset{\text{property } \textbf{ti2*}}{\Rightarrow}$$

$$\sum_{x \in X} T(I(m_{A_1}(x), m_{A_2^c}(x)), m_{A_1}(x)) \geq \sum_{x \in X} T(I(m_{A_2}(x), m_{A_2^c}(x)), m_{A_2}(x)) \Rightarrow$$

$$\frac{\sum_{x \in X} T(I(m_{A_1}(x), m_{A_2^c}(x)), m_{A_1}(x))}{\sum_{x \in X} m_{A_1}(x)} \leqslant \frac{\sum_{x \in X} T(I(m_{A_2}(x), m_{A_2^c}(x)), m_{A_2}(x))}{\sum_{x \in X} m_{A2}(x)} \Rightarrow$$

$$S(A_1, A_2^c) \geqslant S(A_2, A_2^c)$$

3.4 New Measures

We've already seen that "couples" (T_P, I_{Gog}), (T_{BD}, I_{Luc}), (T_m, I_{God}) apply to Proposition 2 and that they all return Kosko's inclusion and entropy measure. Two other couples that can be used are:

– (T_P, I_Z^*) returning:

$$S_{\text{pro}-Z}(A, B) = \begin{cases} \frac{\sum_{x \in X} I_Z^*(m_A(x), m_B(x)) \cdot m_A(x))}{\sum_{x \in X} m_A(x)} & \text{if } A \neq \varnothing \\ 1 & \text{if } A = \varnothing \end{cases}$$

and

$$E_{\text{pro}-\text{Z}}(A,B) = \begin{cases} \frac{\sum_{x\in X} \min(m_{A^c}(x), m_A(x)) \cdot \max(m_{A^c}(x), m_A(x))}{\sum_{x\in X} \max(m_{A^c}(x), m_A(x))} & \text{if } A \neq P \\ 1 & \text{if } A = P \end{cases}$$

– (T_m, I_{Wu}) returning:

$$S_{\min-\text{Wu}}(A,B) = \begin{cases} \frac{\sum_{x\in X} \min(I_{\text{Wu}}(m_A(x), m_B(x)), m_A(x))}{\sum_{x\in X} m_A(x)} & \text{if} A \neq \varnothing \\ 1 & \text{if } A = \varnothing \end{cases}$$

and

$$E_{\min-\text{Wu}}(A) = E_K(A) = \frac{\sum_{x\in X} \min(m_{A^c}(x), m_A(x))}{\sum_{x\in X} \max(m_{A^c}(x), m_A(x))}$$

So there you go. We have two new inclusion functions and one new entropy indicator. In Sects. 4 and 5, we'll have a first look at their values and their behavior. We close this subsection by pointing out that axiom ***S3**** can be loosened even more. We know this; and this will probably allow us to have even more suitable (T, I) combinations and their respective inclusion and entropy measures. It's something we're currently working on. However, we would like to publish these results along with a classification and a more profound study on all these inclusion and entropy measures in a future paper.

3.5 A Noticeable Observation

As Kosko and Young mention in their respective presentations, S_K and generally subsethood measures under Young's axiomatization are connected with conditional probabilities. Since the previously introduced measures satisfy Young's axioms (although S3 is slightly altered) and are being produced according to Young's work, they are also connected with probability theory (not in all cases though). We will not further examine this connection in this presentation but we wish to mention that in the case of the "product-based" measures:

Proposition 3. *All "product-based" subsethood measures satisfy the following—respective to the additive law—property:*

$$S(C, A \cup B) = S(C, A) + S(C, B) - S(C, A \cap B)$$

Proof. Let $A = \{m_A(x_1), \ldots, m_A(x_k), \ldots, m_A(x_n)\}$
$B = \{m_B(x_1), \ldots, m_B(x_k), \ldots, m_B(x_n)\}$
$C = \{m_C(x_1), \ldots, m_C(x_n)\}$ be our fuzzy sets, where—without any loss of generality—$m_A(x_i) \geq m_B(x_i)$ for $i = 1(1)k$ and $m_A(x_i) \leq m_B(x_i)$ for $i = k(1)n$. Then:

$$\sum_{x\in X} I\left(m_C(x), [\max(m_A(x), m_B(x)) + \min(m_A(x), m_B(x))]\right) \cdot m_C(x)$$

$$= \sum_{i=1}^{k} I(m_C(x_i), m_A(x_i)) \cdot m_C(x_i) + \sum_{i=k}^{n} I(m_C(x_i), m_B(x_i)) \cdot m_C(x_i)$$

$$+ \sum_{i=1}^{k} I(m_C(x_i), m_B(x_i)) \cdot m_C(x_i) + \sum_{i=k}^{n} I(m_C(x_i), m_A(x_i)) \cdot m_C(x_i)$$

$$= \sum_{x\in X} I(m_C(x), m_A(x)) \cdot m_C(x) + \sum_{x\in X} I(m_C(x), m_B(x)) \cdot m_C(x)$$

$$\Rightarrow \frac{\sum_{x\in X} I\left(m_C(x), [\max(m_A(x), m_B(x)) + \min(m_A(x), m_B(x))]\right) \cdot m_C(x)}{\sum_{x\in X} m_C(x)}$$

$$= \frac{\sum_{x\in X} I(m_C(x), m_A(x)) \cdot m_C(x)}{\sum_{x\in X} m_C(x)} + \frac{\sum_{x\in X} I(m_C(x), m_B(x)) \cdot m_C(x)}{\sum_{x\in X} m_C(x)}$$

$$\Rightarrow S(C, A \cup B) + S(C, A \cap B) = S(C, A) + S(C, B)$$

Now we can show that:

$$\begin{aligned} S(B, A_1 \cup A_2 \cup A_3) = {} & S(B, A_1) + S(B, A_2) + S(B, A_3) \\ & - S(B, A_1 \cap A_2) - S(B, A_1 \cap A_3) \\ & - S(B, A_2 \cap A_3) + S(B, A_1 \cap A_2 \cap A_3) \end{aligned}$$

Proof.

$$\begin{aligned} & S(B, A_1 \cup A_2 \cup A_3) \\ & = S(B, (A_1 \cup A_2) \cup A_3) \\ & = S(B, A_1 \cup A_2) + S(B, A_3) - S(B, (A_1 \cup A_2) \cap A_3) \\ & = S(B, A_1) + S(B, A_2) - S(B, A_1 \cap A_2) + S(B, A_3) - S(B, (A_1 \cap A_3) \cup (A_2 \cap A_3)) \\ & = S(B, A_1) + S(B, A_2) + S(B, A_3) - S(B, A_1 \cap A_2) - S(B, A_1 \cap A_3) - S(B, A_2 \cap A_3) \\ & \quad + S(B, A_1 \cap A_2 \cap A_3) \end{aligned}$$

Then, we can generalize this for the inclusion grade of a set B into the union of n sets A_i, $i = 1(1)n$ and get the respective property of possibility theory.

4 A First Comparison of The Inclusion Measures

Let's see some examples:

Example 1. Let's take the following fuzzy sets:

$$A_1 = \{0.9, 0.8, 0.5, 0.5, 0.2, 0.1\}, A_2 = \{0.7, 0.7, 0.8, 0.5, 0.2, 0.1\},$$
$$A_3 = \{0.6, 0.7, 0.7, 0.8, 0.8, 1\}, B = \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$$

The corresponding inclusion grades are:

	$S(A_1, B)$	$S(A_2, B)$	$S(A_3, B)$
S_K	0.8	0.866667	0.956522
$S_{\text{pro}-\text{Z}}$	0.743333	0.71	0.873913
$S_{\min-\text{Wu}}$	0.533333	0.533333	0.869565

In the first case, S_K and $S_{\text{pro}-\text{Z}}$ are quite close, whereas $S_{\min-\text{Wu}}$ is close to $\frac{1}{2}$. Inequality $m_A(x) \leq m_B(x)$ is violated two out of six times and it's:

$$a = \frac{\sum_{x \in X \text{ and } m_A(x) > m_B(x)} (m_A(x) - m_B(x))}{\sum_{x \in X \text{ and } m_A(x) < m_B(x)} (m_B(x) - m_A(x))} = \frac{6}{21} = 28\,\%$$

According to a, the value of $S_{\text{pro}-\text{Z}}(A_1, B)$ being close to 74 % seems quite reasonable. On the other hand, $S_{\min-\text{Wu}}$ sees set A as a subset of set B, almost as much as it doesn't. It seems that $S_{\min-\text{Wu}}$ is affected by fraction a in a "harsher" way. In the second case, while a is the same, it is distributed among more elements of A. Kosko's measure gives a larger inclusion grade whereas $S_{\text{pro}-\text{Z}}$ a smaller one—a fact quite reasonable in our opinion. $S_{\min-\text{Wu}}$ returns the same value, something that could also seem justifiable to some. Finally, regarding $S(A_3, B)$, $S_{\text{pro}-\text{Z}}$ and $S_{\min-\text{Wu}}$ are almost equal. On the other hand, the fact that $S_K(A_3, B)$ is so close to 1 may seem a little excessive since $m_{A_3}(x_1) > m_B(x_1)$ and $m_{A_3}(x_2) > m_B(x_2)$.

Example 2. Let's examine now some smaller inclusion grades. Let

$$A_1 = \{0.9, 0.8, 0.5, 0.5, 0.2, 0.1\}, A_2 = \{0.7, 0.7, 0.8, 0.5, 0.2, 0.1\}$$

(same as above) and

$$A_3 = \{0.4, 0.5, 0.4, 0.4, 0.4, 0.5\} \text{ (near set } P), B = \{0.1, 0.2, 0.2, 0.3, 0.4, 0.5\}.$$

The corresponding inclusion grades are:

	$S(A_1, B)$	$S(A_2, B)$	$S(A_3, B)$
S_K	0.366667	0.366667	0.653846
$S_{\text{pro}-\text{Z}}$	0.35	0.376667	0.719231
$S_{\min-\text{Wu}}$	0.366667	0.366667	0.653846

S_K and $S_{\min-\text{Wu}}$ return the exact same values. In the first case $S_{\text{pro}-\text{Z}}$ is very close to them. As the fuzziness of set A_1 is distributed among more of its elements (set A_2), $S_K(A_2, B)$ and $S_{\min-\text{Wu}}(A_2, B)$ remain the same while $S_{\text{pro}-\text{Z}}(A_2, B)$ gets a bit larger. This again shows a sensitivity of indicator $S_{\text{pro}-\text{Z}}$ regarding the distribution of fuzziness. The difference between $S_{\text{pro}-\text{Z}}$ and the other two measures gets larger in the third case. Someone could say that S_K returns a relatively low value, taking into concern $S_K(A_2, B)$ of Example 1.

Now let's see some cases where the differences between the indicators are more noticeable:

Example 3. Sets are:

$$A_1 = \{0.8, 0.8, 0.8, 0.8, 0.8, 0.8\}, A_2 = \{0.7, 0.9, 0.7, 0.9, 0.7, 0.9\},$$
$$A_3 = \{0.7, 0.7, 0.7, 0.7, 1, 1\}, B = \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$$

and inclusion grades are:

	$S(A_1, B)$	$S(A_2, B)$	$S(A_3, B)$
S_K	0.6875	0.6875	0.6875
$S_{\text{pro}-\text{Z}}$	0.583333	0.6	0.575
$S_{\min-\text{Wu}}$	0.375	0.333333	0.25

The first thing that one should notice is that S_K returns the same value in all three cases (maybe a rather large value, taking into concern fraction a and the number of violations of $m_A(x) \leq m_B(x)$). As before, S_{pro-Z} shows a greater sensitivity and seems to have "logical" variations. Once again, $S_{\min-\text{Wu}}$ is "stricter" and closer to the concept of crisp inclusion. To some readers, values of $S_{\min-\text{Wu}}$ may seem more logical than those of the others inclusion measures. $S_{\min-\text{Wu}}$ could be more suitable concerning specific applications (having strict control rules). Finally let's see some examples where set A remains the same and set B doesn't:

Example 4. Let's take sets

$$A = \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}, B_1 = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.9\},$$
$$B_2 = \{0.6, 0.7, 0.4, 0.5, 0.3, 0.8\}, B_3 = \{0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$$

We have:

	$S(A, B_1)$	$S(A, B_2)$	$S(A, B_3)$
S_K	0.848485	0.818182	0.727273
$S_{\text{pro}-\text{Z}}$	0.672727	0.684848	0.612121
$S_{\min-\text{Wu}}$	0.727273	0.787879	0.636364

Here, we can see considerable differences between S_K and $S_{\text{pro}-\text{Z}}$, while $S_{\min-\text{Wu}}$ is somewhere between them.

Example 5. Finally, let's alter set $A = \{0.1, 0.2, 0.2, 0.8, 0.8, 0.9\}$ (more crisp) and keep sets B_1, B_2, and B_3 the same as above. We have:

	$S(A, B_1)$	$S(A, B_2)$	$S(A, B_3)$
S_K	0.833333	0.7	0.566667
$S_{\text{pro}-\text{Z}}$	0.76	0.62	0.496667
$S_{\min-\text{Wu}}$	0.6	0.333333	0.333333

Here measures behave in the same way though their respective values have noticeable differences. $S_{\min-\text{Wu}}$ is once again the stricter one.

We could see the behavior of the indicators through some graphs as well. Below, we can better see their similarities as well as their differences concerning several cases of sets A and B. S_K is represented by the dotted line, $S_{\text{pro}-\text{Z}}$ by the thick line, and $S_{\min-\text{Wu}}$ by the dashed line:

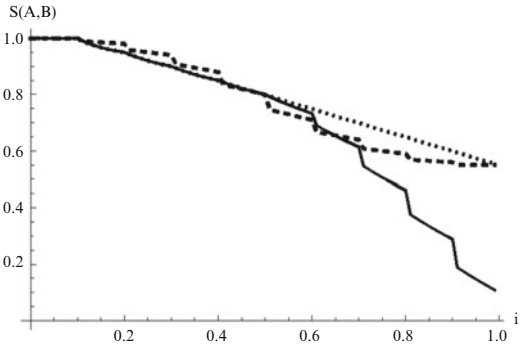


Fig. 1 $A = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}$, $i \in [0, 1]$ and $B = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$

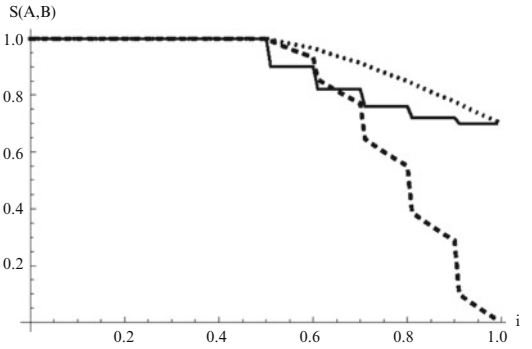


Fig. 2 $A = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}, i \in [0, 1]$ and $B = \{0.9, 0.9, 0.8, 0.8, 0.7, 0.7, 0.6, 0.6, 0.5, 0.5\}$

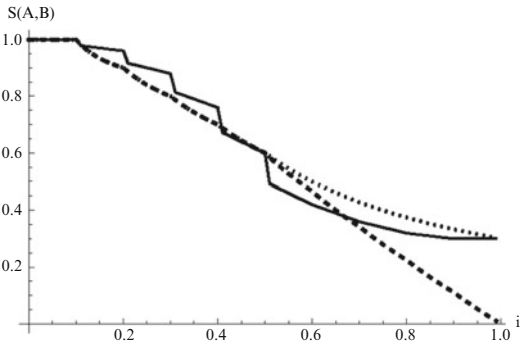


Fig. 3 $A = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}, i \in [0, 1]$ and $B = \{0.1, 0.1, 0.2, 0.2, 0.3, 0.3, 0.4, 0.4, 0.5, 0.5\}$

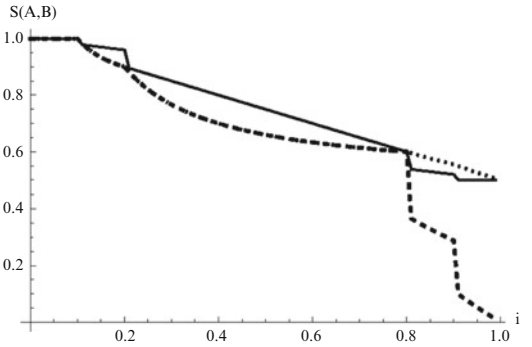


Fig. 4 $A = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}, i \in [0, 1]$ and $B = \{0.1, 0.1, 0.2, 0.2, 0.2, 0.8, 0.8, 0.8, 0.9, 0.9\}$

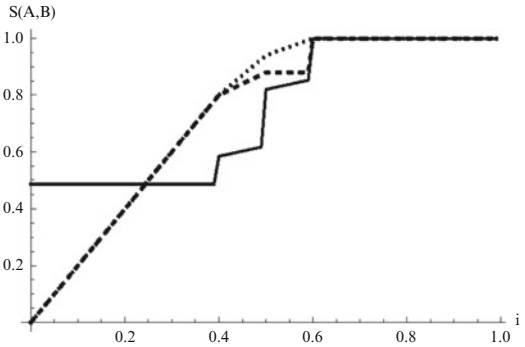


Fig. 5 $A = \{0.1, 0.1, 0.2, 0.2, 0.2, 0.8, 0.8, 0.8, 0.9, 0.9\}$ and $B = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}, i \in [0, 1]$

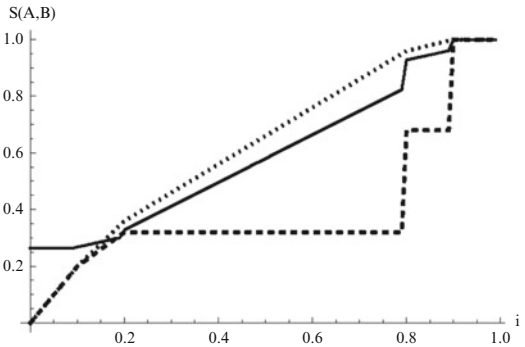


Fig. 6 $A = \{0.4, 0.4, 0.4, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.6\}$ and $B = \{\underbrace{i, \ldots, i}_{10 \text{ times}}\}, i \in [0, 1]$

To have an even better view of the three measures, here are their 3D plots accompanied by their respective Contour plots when applied to sets $A = \{i, \ldots, i\}$ and $B = \{j, \ldots, j\}$, $i, j \in [0, 1]$:

These examples and graphs are only indicative and not of course sufficient to extensively study the behavior of the measures. We only wish to show some differences and some similarities between the indicators concerning specific cases. The graphics show these even more comprehensively. The main purpose of this paper is to present a new formula of producing inclusion and entropy measures accompanied by some specific applications and examples of this procedure. Moreover, no supremacy of one measure over the other is implied. It depends on what application fuzzy inclusion is being used in, the rules and their strictness someone is setting, each researcher's different perspective, and many other factors. We personally believe that different piece of information could be obtained by each measure, whereas their combination (mean value, confidence intervals, etc.) could give more suitable answers when it comes to applications. In any case, a further

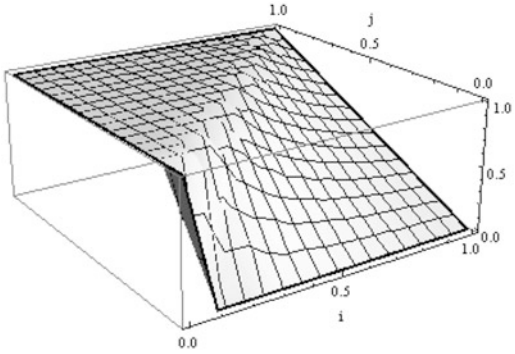


Fig. 7 3D plot of S_K

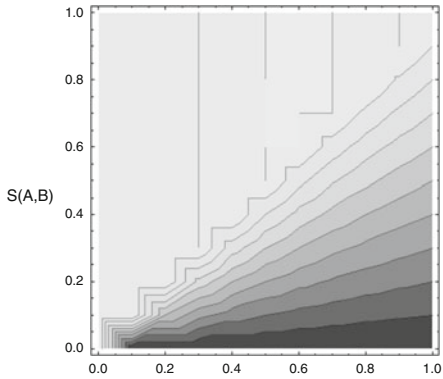


Fig. 8 Contour plot of S_K

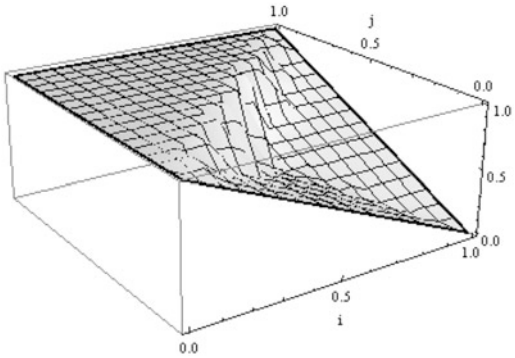


Fig. 9 3D plot of $S_{\text{pro}-\text{Z}}$

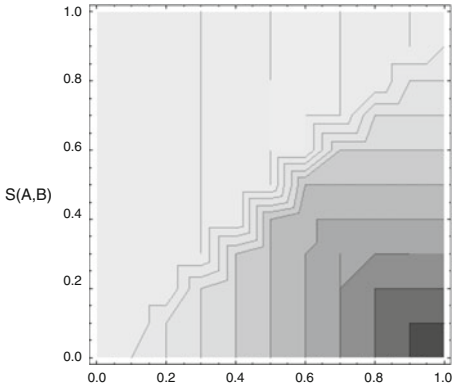


Fig. 10 Contour plot of $S_{\text{pro}-\text{Z}}$

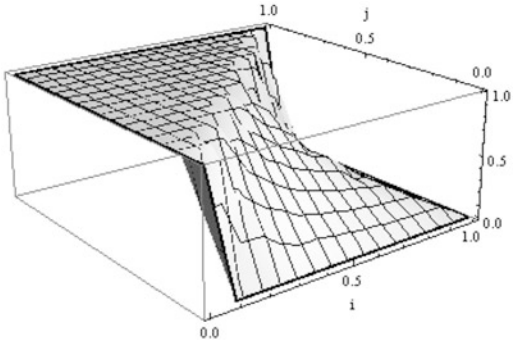


Fig. 11 3D plot of $S_{\min-\text{Wu}}$

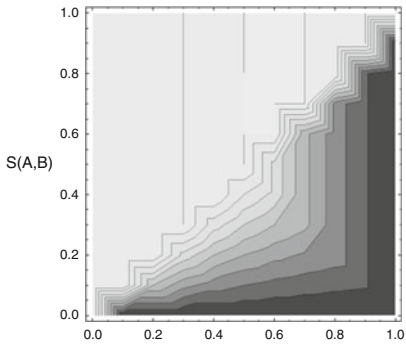


Fig. 12 Contour plot of $S_{\min-\text{Wu}}$

theoretical study and comparison between them—as well as others produced using different implications and intersections—is necessary. Then we need to see each measure's behavior as well as what their combination could offer us in specific applications. All these are things we're currently working on and their results will be presented in future papers. Same things apply for their corresponding entropy measures which we will see next.

5 Entropy Measures

Previously, we saw that our formula combined with Young's theorem allowed us to produce two entropy measures: E_K (Kosko's) and $E_{\text{pro}-\text{Z}}$. However, when it comes to applications, $E_{\text{pro}-\text{Z}}$ is practically useless since it can't be larger than $\frac{1}{2}$ (apart from the case when $A = P$). Nevertheless, we can easily turn it into a sufficient entropy measure by doubling its value:

$$E_1(A) = 2E_{\text{pro}-\text{Z}}(A) = \frac{2\sum_{x\in X} \min(m_{A^c}(x), m_A(x)) \cdot \max(m_{A^c}(x), m_A(x))}{\sum_{x\in X} \max(m_{A^c}(x), m_A(x))}$$

E_1 is clearly an entropy measure (according to De Luca and Termini) and it's

$$E_1(A) \geq E_K(A) \text{ for every } A \in F(X)$$

Furthermore, in contradiction with E_K, E_1 has a linear behavior. This can be clearly seen in Fig. 13 where E_1 is represented by the dashed line:

It also appears to be more "sensitive" than E_K. That's logical if we think that algebraic product $m_A(x) \cdot m_B(x)$ is evaluated by both m_A and m_B, whereas $\min(m_A(x), m_B(x))$ simply returns one out of the two values. This means that E_1 takes into consideration not just the distance of set A from A_{near} but the intersection

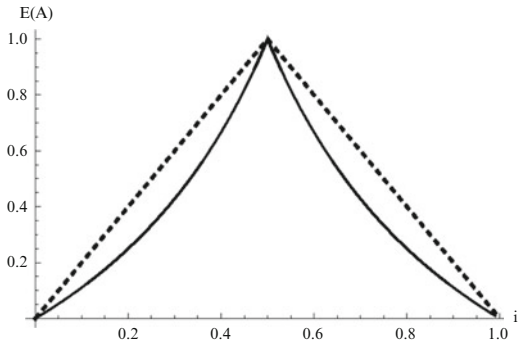


Fig. 13 $E_K(A)$ and $E_1(A)$ when $A = \{i, \ldots, i\}, i = [0, 1]$

of the distances of set A from both A_{near} and A_{far}. Thus, the quantity of elements of A which have certain distances from A_{near} and A_{far} is playing a crucial role to its entropy.

Let's see what we mean through some brief examples:

Example 6. Let's take fuzzy sets:

$$A = \{0.25, 0.25, 0.25, 0.25, 0.25, 0.25\}, B = \{0.5, 0.5, 0.5, 1, 1, 1\},$$
$$C = \{0.4, 0.3, 0.1, 0.3, 0.3, 0.1\}, D = \{0.5, 0.2, 0.1, 0.3, 0.3, 0.1\}$$

Then:

	A	B	C	D
E_K	0.333333	0.333333	0.333333	0.333333
E_1	0.5	0.333333	0.466667	0.448889

These sets have six elements and the largest total amount of vagueness (meaning $\sum_{x \in X} \min(m_A(x), 1 - m_A(x))$) they can have is 3. In this case, both functions are of course equal to 1.

All elements of A have half the fuzziest truth value. At a certain point, someone would expect for an entropy indicator to return $\frac{1}{2}$. That's true for E_1, whereas Kosko's measure returns $\frac{1}{3}$. Kosko's indicator is equal to $\frac{1}{2}$ when truth values of A are all $\frac{1}{3}$ or $\frac{2}{3}$. The same applies to the case when truth values of A are all equal to 0.75. So, what we see for E_1 is that if all elements of a fuzzy set lose a certain amount of their fuzziness (all elements having the same value), the same percentage is also lost by its entropy value (Fig. 13).

Sets B and C have also a total amount of fuzziness equal to 1.5. Nevertheless, half elements of set B have maximum fuzziness and half have zero. The total vagueness of set C, though equal with that of set B, is distributed among more of its elements (but not equally as in set A). We see that since the total fuzziness of sets A, B, and C is the same, E_K always returns the same value (although set B definitely contains three elements). E_1 doesn't have this property. However, we doubt if this a negative fact. We believe that there should be some difference in the entropy of sets A, B, and C. We see that since set B definitely contains half of its elements, E_1 considers it less fuzzy than A or C. As far as set C is concerned, although its total vagueness is equal with that of set A, 4 of its elements are more fuzzy and only two otherwise. E_1 seems to recognize this fact by giving set C a larger entropy value.

Sets D and C differ in only two truth values. First element of D is 0.1 more fuzzy whereas its second is 0.1 less. In this case, we must give best to Kosko's measure since we believe we should have the same entropy for both sets. $E_1(D)$ is slightly smaller than $E_1(C)$, at a percentage of 3.8 %.

Now, let's see an example where sets have different total fuzziness:

Example 7. These are:

$$A = \{0.2, 0.3, 0.2, 0.4, 0.5, 0.3\}, B = \{0.2, 0.2, 0.8, 0.3, 0.8, 0.3\},$$
$$C = \{1, 1, 1, 0.5, 0.45, 0.45\}, D = \{0.1, 0.2, 0.9, 0.8, 0.9, 0.3\}$$

and we have:

	A	B	C	D
E_K	0.463415	0.304348	0.304348	0.2
E_1	0.6	0.46087	0.323913	0.32

Their sums of fuzziness are 1.8, 1.4, 1.4, and 1.0, respectively. The corresponding fractions of the vagueness of the sets to the largest possible are:

$$f_1 = \frac{3}{5}, f_2 = \frac{7}{15}, f_3 = \frac{7}{15}, f_4 = \frac{1}{3}$$

We can easily see that E_1 returns values closer to f_1, f_2, and f_4. That is not the case for set C which has three elements with truth value equal to 1. Furthermore, the percentage reductions in vagueness between sets A and B is 22.2 %, B and D is 28.6 %, A and D is 44.4 %. The corresponding reductions in the entropy of the sets are:

	A and B	B and D	A and D
E_K	34.3 %	34.3 %	56.84 %
E_1	23.2 %	30.5 %	46.6 %

Once again, E_1 seems to have more "normal" changes in its values concerning the changes in the fuzziness of a set. However, this doesn't happen when the change is less "uniform," like between sets A and C. Then the respective percentages are 34.3 % for E_K and 46 % for E_1; something expected taking into consideration Example 6. In other words E_1 is affected by more parameters.

Finally, let's have a set that gradually becomes more fuzzy:

Example 8. We have:

$$A = \{0.1, 0.2, 0.1, 0.2, 0.1, 0.3\}, B = \{0.2, 0.2, 0.2, 0.3, 0.3, 0.3\},$$
$$C = \{0.1, 0.2, 0.2, 0.5, 0.5, 0.5\}, D = \{0.4, 0.4, 0.3, 0.3, 0.1, 0.5\}$$

and

	A	B	C	D
E_K	0.2	0.333333	0.5	0.5
E_1	0.32	0.493333	0.58	0.62

Their sums of fuzziness are 1, 1.5, 2, and 2. Inclusion measures return:

	A	B	C	D
E_K	0.2	0.333333	0.5	0.5
E_1	0.32	0.493333	0.58	0.62

Once again, these results lead us to conclusions similar with those derived from the previous examples. The increase in the entropy of a fuzzy set seems to be more "normally" depicted by E_1. There is only a difference between $E_1(C)$ and $E_1(D)$ which, however, is about 4 %.

Similar observations can be made through a variety of examples. We don't intend to argue whether E_K or E_1 is best when it comes to applications (especially when we've seen just a few specific short examples). But we see that Kosko's indicator takes into concern only the total sum of fuzziness of a set, whereas E_1 seems to be more "sensitive" to the way that this vagueness is distributed among its elements (something wishful to some cases). Sporadic changes seem to affect E_1 which is more stable when fluctuations in the vagueness of a set are more normally and symmetrically distributed among its elements. Anyway, as with inclusion measures, different entropy measures are not meant to be adversary. They can be cooperative since the usage of different measures can give a more spherical view of the behavior of fuzzy sets. Having a variety of indicators is something welcomed when it comes to applications. We have more choices and we can have more information. Something we are currently working on is making rules of choosing the proper entropy measure for specific applications (image thresholds, decision making, etc.)

6 Conclusion

In this paper, we gave an alternative axiomatization of fuzzy inclusion (based on Young's axiomatization) and proposed a formula for producing inclusion and—their corresponding—entropy measures. Then, we produced already known inclusion and entropy indicators as well as possible new ones. We compared their results through some examples and graphs.

As far as our future work is concerned, we are currently working on several matters. Our next step will be the production of even more inclusion and entropy measures by further loosening our axiomatization (specifically ***S3a***). Other fuzzy operators—apart from classic fuzzy implications and intersections—can be used

in our formula and we would like to further examine this process and its results. Apart from these, something truly significant would be to include to our propositions necessary conditions as well. Finally, we intend to use all these possible measures in specific applications (fuzzy controllers, image thresholds, etc.). This way, we could compare and classify them more effectively, have a better look on their behavior, and see if their combination could give us further information and better results. Then, we could set some rules of choosing the "right" inclusion and entropy measure depending on the application it is meant to be used in.

References

1. Bandler, W., Kohout, L.: Fuzzy power sets and fuzzy implication operators. Fuzzy Set. Syst. **4**, 183–190 (1980)
2. Burillo, P., Frago, N., Fuentes, R.: Inclusion grade and fuzzy implication operators. Fuzzy Set. Syst. **114**, 417–429 (2000)
3. Cornelis, C., Van der Donck, C., Kerre, E.: Sinha-Dougherty approach to the fuzzification of set inclusion revisited. Fuzzy Set. Syst. **134**, 283–295 (2003)
4. DeLuca, A., Termini, S.: A definition of a non probabilistic entropy in the setting of fuzzy sets theory. Inf. Control **20**, 301–312 (1972)
5. Ebanks, B.: On measures of fuzziness and their representations. J. Math. Anal. Appl. **94**, 24–37 (1983)
6. Goguen, J.A.: The logic of inexact concepts. Systhese **19**, 325–373 (1969)
7. Kitainik, L.: Fuzzy inclusions and fuzzy dichotomous decision procedures. In: Kacprzyk, J., Orlovski, S. (eds.) Optimization Models Using Fuzzy Sets and Possibility Theory, pp. 154–170. Reidel, Dordrecht (1987)
8. Kitainik, L.: Fuzzy Decision Procedures with Binary Relations. Kluwer, Dordrecht (1993)
9. Kosko, B.: Fuzzy entropy and conditioning. Inf. Sci. **40**, 165–174 (1986)
10. Kosko, B.: Fuzziness vs. probability. Int. J. Gen. Syst. **17**, 211–240 (1990)
11. Kosko, B.: Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence. Prentice-Hall, Englewood Cliffs (1992)
12. Sander, W.: On measures of fuzziness. Fuzzy Set. Syst. **29**, 49–55 (1989)
13. Sinha, D., Dougherty, E.: Fuzzification of set inclusion theory and applications. Fuzzy Set. Syst. **55**, 15–42 (1993)
14. Wang, Z., Klir, G.: Fuzzy Measure Theory. Plenum Press, New York (1992)
15. Willmott, R.: Two fuzzier implication operators in the theory of fuzzy power sets. Fuzzy Set. Syst. **4**, 31–36 (1980)
16. Willmott, R.: Mean measures of containment and equality between fuzzy sets. In: Proceedings of the 11th International Symposium on MultipleValued Logic, Silver Spring, Md.: IEEE Computer Society Press, c1981, pp. 183–190. Oklahoma (1981)
17. Young, V.R.: Fuzzy subsethood. Fuzzy Set. Syst. **77**, 371–384 (1996)
18. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)

On Some Recent Results on Asymptotic Behavior of Orthogonal Polynomials on the Unit Circle and Inserting Point Masses

Kenier Castillo and Francisco Marcellán

Abstract In the present paper, we formulate and reprove, in a brief and self-contained presentation, some recent results concerning the asymptotic behavior of orthogonal polynomials on the unit circle by inserting point masses recently obtained by the authors and co-workers. In a first part, we deal with a spectral transformation of a Hermitian linear functional by the addition of the first derivative of a complex Dirac linear functional supported either in a point on the unit circle or in two symmetric points with respect to the unit circle. In this case, outer relative asymptotics for the new sequences of orthogonal polynomials in terms of the original ones are obtained. Necessary and sufficient conditions for the quasi-definiteness of the new linear functionals are given. The relation between the corresponding sequence of orthogonal polynomials in terms of the original one is presented. The second part is devoted to the study of a relevant family of orthogonal polynomials associated with perturbations of the original orthogonality measure by means of mass points: discrete Sobolev orthogonal polynomials. We compare the discrete Sobolev orthogonal polynomials with the initially ones. Finally, we analyze the behavior of their zeros.

Keywords: Orthogonal polynomials on the unit circle • Asymptotic behavior • Inserting point masses

1 Orthogonal Polynomials on the Unit Circle

In this section, a brief introduction to *orthogonal polynomials on the unit circle* (OPUC, in short) is given, in order to have a self-contained and accessible presentation of the results for a reader who is not familiar with the OPUC theory

K. Castillo (✉)
CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal
e-mail: kcastill@math.uc3m.es; kenier@mat.uc.pt

F. Marcellán
Instituto de Ciencias Matemáticas (ICMAT) and Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Spain
e-mail: pacomarc@ing.uc3m.es

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_4

[18, 22, 38, 39, 43]. We denote by $\Lambda := \mathbb{C}[z, z^{-1}]$ the vector space of Laurent polynomials in the variable z with complex coefficients. Associated with every pair of integers $(p, q), p \leq q$, we define the vector subspace $\Lambda_{p,q}$ of Laurent polynomials of the form

$$\sum_{n=p}^{q} a_n z^n, \quad a_n \in \mathbb{C}.$$

The vector subspace of complex polynomials will be denoted by $\mathbb{P} := \mathbb{C}[z]$ and we write $\mathbb{P}_q \equiv \Lambda_{0,q}$ for the vector subspace of polynomials of degree (at most) q, while $\mathbb{P}_{-1} \equiv \{0\}$ is the trivial subspace.

Let $\mathcal{L}$ be a linear functional in Λ satisfying

$$c_n = \langle \mathcal{L}, z^n \rangle = \overline{\langle \mathcal{L}, z^{-n} \rangle} = \overline{c}_{-n}, \qquad n \in \mathbb{Z}. \tag{1}$$

$\mathcal{L}$ is said to be a Hermitian linear functional. A bilinear functional associated with $\mathcal{L}$ can be introduced in $\mathbb{P}$ as follows:

$$\langle f, g \rangle_{\mathcal{L}} = \left\langle \mathcal{L}, f(z)\overline{g}(z^{-1}) \right\rangle, \qquad f, g \in \mathbb{P}.$$

The complex numbers $\{c_n\}_{n\in\mathbb{Z}}$ are said to be the canonical moments of $\mathcal{L}$ and the infinite matrix

$$\mathbf{T} = \left[\left\langle z^i, z^j \right\rangle_{\mathcal{L}} \right]_{i,j\geq 0} = \begin{bmatrix} c_0 & c_1 & \cdots & c_n & \cdots \\ c_{-1} & c_0 & \cdots & c_{n-1} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \\ c_{-n} & c_{-n+1} & \cdots & c_0 & \cdots \\ \vdots & \vdots & & \vdots & \ddots \end{bmatrix},$$

is the Gram matrix of the above bilinear functional in terms of the canonical basis $\{z^n\}_{n\geqslant 0}$ of $\mathbb{P}$. It is known in the literature as a Toeplitz matrix, a matrix in which each descending diagonal from left to right is constant [21]. Subdiagonal perturbations of Toeplitz matrices and their relation with OPUC are considered in [10]

If $\mathbf{T}_n$, the $(n + 1) \times (n + 1)$ principal leading submatrix of $\mathbf{T}$, is non-singular for every $n \geqslant 0$, $\mathcal{L}$ is said to be quasi-definite, and there exists a sequence of monic polynomials $\{\Phi_n\}_{n\geqslant 0}$, orthogonal with respect to $\mathcal{L}$,

$$\langle \Phi_n, \Phi_m \rangle_{\mathcal{L}} = \mathbf{k}_n \delta_{n,m}, \quad \mathbf{k}_n \neq 0, \quad m \geqslant 0.$$

1.1 Recurrence Relations

The OPUC satisfy the following forward and backward recurrence relations

$$\Phi_{n+1}(z) = z\Phi_n(z) + \Phi_{n+1}(0)\Phi_n^*(z), \quad n \geqslant 0, \tag{2}$$

$$\Phi_{n+1}(z) = \left(1 - |\Phi_{n+1}(0)|^2\right) z\Phi_n(z) + \Phi_{n+1}(0)\Phi_{n+1}^*(z), \quad n \geqslant 0, \tag{3}$$

where $\Phi_n^*(z) = z^n\overline{\Phi}_n(z^{-1}) = z^n(\Phi_n)_*(z)$ is the so-called reversed polynomial, and the complex numbers $\{\Phi_n(0)\}_{n\geqslant 1}$, with

$$|\Phi_n(0)| \neq 1, \quad n \geqslant 1,$$

are known as Verblunsky, Schur, or reflection coefficients. The OPUC are therefore completely determined by the sequence $\{\Phi_n(0)\}_{n\geq 1}$. To obtain the recurrence formula, we take into account the fact that the reversed polynomial $\Phi_n^*(z)$ is the unique polynomial of degree at most n orthogonal to z^k, $1 \leqslant k \leqslant n$. Equations (2) and (3) are called either the Szegő recurrence or Szegő difference relations. Moreover, we have

$$\langle \Phi_n, \Phi_n\rangle_{\mathcal{L}} = \mathbf{k}_n = \frac{\det \mathbf{T}_n}{\det \mathbf{T}_{n-1}}, \qquad n \geq 1, \quad \mathbf{k}_0 = c_0.$$

In a recent paper [7], new sequences of OPUC associated with finite perturbation of (2) and (3) are considered.

We can derive a recurrence formula which does not involve the reversed polynomials,

$$\Phi_n(0)\Phi_{n+1}(z) = (z\Phi_n(0)+\Phi_{n+1}(0))\Phi_n(z)-z(1-|\Phi_n(0)|^2)\Phi_{n+1}(0)\Phi_{n-1}(z), \quad n \geqslant 0,$$

if we assume $\Phi_{-1} = 0$. The polynomials Φ_{n+1} can be found from Φ_{n-1} and Φ_n, if $\Phi_n(0) \neq 0$. This is an analogue of the three-term recurrence relation for *orthogonal polynomials on the real line* (OPRL, in short), except for the factor z in the last term. In [5, 8], the authors show that similar recurrence relations are associated with the *para-orthogonal polynomials on the unit circle* (POPUC, in short).

1.2 Integral Representation and Kernel Polynomials

If $c_0 = 1$ and $\det \mathbf{T}_n > 0$, for every $n \geqslant 0$, then $\mathcal{L}$ is said to be positive definite and it has the following integral representation

$$\langle \mathcal{L}, f\rangle = \int_{\mathbb{T}} f(z)d\sigma(z), \quad f \in \mathbb{P},$$

where σ is a non-trivial probability measure supported on the unit circle $\mathbb{T}$. In such a case, there exists a unique sequence of polynomials $\{\phi_n\}_{n\geqslant 0}$, with positive leading coefficients, such that

$$\int_{\mathbb{T}} \phi_n(z)\overline{\phi_m(z)}d\sigma(z) = \delta_{m,n}, \quad m \geq 0.$$

$\{\phi_n\}_{n\geqslant 0}$ is said to be the sequence of orthonormal polynomials with respect to $d\sigma$. Denoting by κ_n the leading coefficient of ϕ_n, $\Phi_n = \kappa_n^{-1}\phi_n$ is the corresponding OPUC of degree n. Moreover, $\langle \Phi_n, \Phi_n\rangle_{\mathcal{L}} = \|\Phi_n\|_\sigma^2 = \mathbf{k}_n > 0$.

Using the Pythagoras theorem, (2) yields

$$\frac{\|\Phi_n\|_\sigma^2}{\|\Phi_{n-1}\|_\sigma^2} = 1 - |\Phi_n(0)|^2 > 0, \quad n \geqslant 1.$$

This shows that in the positive definite case the Verblunsky coefficients always satisfy

$$|\Phi_n(0)| < 1, \quad n \geqslant 1. \tag{4}$$

In this situation, we have an analogous of the Favard theorem [36, 37, 46], formulated as follows. Any sequence of complex numbers obeying (4) arises as the Verblunsky coefficients of a unique non-trivial probability measure supported on the unit circle. In the POPUC theory, the Favard Theorem was recently proved in [8].

We use the notation $\rho_n = \sqrt{1-|\Phi_n(0)|^2} = \|\Phi_n\|_\sigma / \|\Phi_{n-1}\|_\sigma = \kappa_{n-1}/\kappa_n$. Hence, for the orthonormal polynomials ϕ_n, the recurrence relations (2)–(3) become

$$\begin{aligned} \rho_{n+1}\phi_{n+1}(z) &= z\phi_n(z) - \Phi_{n+1}(0)\phi_n^*(z), \quad n \geqslant 0, \\ \phi_{n+1}(z) &= \rho_{n+1} z\phi_n(z) + \Phi_{n+1}(0)\phi_{n+1}^*(z), \quad n \geqslant 0, \end{aligned} \tag{5}$$

In the case of OPUC we have a simple expression for the reproducing kernel [1, 17, 38], similar to the Christoffel–Darboux formula for OPRL. The nth polynomial kernel $K_n(z, y)$ associated with $\{\Phi_n\}_{n\geqslant 0}$ is defined by

$$\begin{aligned} K_n(z,y) = \sum_{j=0}^{n} \frac{\overline{\Phi_j(y)}\Phi_j(z)}{\mathbf{k}_j} &= \frac{\overline{\Phi_{n+1}^*(y)}\Phi_{n+1}^*(z) - \overline{\Phi_{n+1}(y)}\Phi_{n+1}(z)}{\mathbf{k}_{n+1}(1-\bar{y}z)} \\ &= \frac{\overline{\phi_{n+1}^*(y)}\phi_{n+1}^*(z) - \overline{\phi_{n+1}(y)}\phi_{n+1}(z)}{1-\bar{y}z}, \end{aligned} \tag{6}$$

and it satisfies the reproducing property,

$$\int_{\mathbb{T}} K_n(z,y)\overline{f(z)}d\sigma(z) = \overline{f(y)}, \tag{7}$$

for every polynomial f of degree at most n. Taking into account $\phi_{n+1}^*(0) = \kappa_{n+1}\Phi_{n+1}^*(0) = \kappa_{n+1}$, we get

$$\Phi_n^*(z) = \frac{1}{\kappa_n^2}K_n(z,0) = \mathbf{k}_n K_n(z,0), \quad n \geq 0,$$

which is an expression for the reversed polynomials as a linear combination of the OPUC up to degree n.

1.3 GGT Matrices

Using the forward recurrence formula (2), we are able to express $z\phi_n(z)$ as a linear combination of $\{\phi_k\}_{k=0}^{n+1}$,

$$z\phi_n(z) = \frac{\kappa_n}{\kappa_{n+1}}\phi_{n+1}(z) - \frac{1}{\kappa_n}\Phi_{n+1}(0)\sum_{k=0}^{n}\kappa_k\overline{\Phi_k(0)}\phi_k(z),$$

or, in the matrix form,

$$z\boldsymbol{\phi}(z) = \mathbf{H}_\sigma\boldsymbol{\phi}(z),$$

where $\boldsymbol{\phi}(z) = [\phi_0(z), \phi_1(z), \dots]^T$, and the matrix $\mathbf{H}_\sigma$ is defined by

$$[\mathbf{H}_\sigma]_{i,j} = \langle z\phi_i, \phi_j\rangle_{\mathcal{L}} = \begin{cases} -\frac{\kappa_j}{\kappa_i}\Phi_{i+1}(0)\overline{\Phi_j(0)}, & j \le i, \\ \frac{\kappa_i}{\kappa_{i+1}}, & j = i+1, \\ 0, & j > i+1. \end{cases}$$

This lower Hessenberg matrix [21], where the jth row has at most its first $j+1$ components non-zero, is called GGT representation of the multiplication by z, after [18, 20, 45].

In an analog way to the real line case, the zeros of the OPUC $\Phi_n(z)$ are the eigenvalues of $(\mathbf{H}_\sigma)_n$, the $n \times n$ principal leading sub-matrix of the GGT matrix $\mathbf{H}_\sigma$. Hence, $\Phi_n(z)$ is the characteristic polynomial of $(\mathbf{H}_\sigma)_n$,

$$\Phi_n(z) = \det\left(z\mathbf{I}_n - (\mathbf{H}_\sigma)_n\right).$$

1.4 Szegő Extremum Problem and $\mathcal{S}$ Class

The measure of orthogonality $d\sigma$ can be decomposed as the sum of a purely absolutely continuous measure with respect to the Lebesgue measure and a singular part. Thus, if we denote by σ', the Radon–Nikodym derivative [35] of the measure σ supported in $[-\pi, \pi]$, then

$$d\sigma(\theta) = \sigma'(\theta)\frac{d\theta}{2\pi} + d\sigma_s,$$

where σ_s is the singular part of σ.

The Szegő extremum problem on the unit circle consists of finding

$$\lambda(z) = \lim_{n\to\infty}\lambda_n(z),$$

with

$$\lambda_n(z) = \inf_{f(z)=1} \left\{ \int_{-\pi}^{\pi} \left| f(e^{i\theta}) \right|^2 d\sigma(\theta);\ f \in \mathbb{P}_n \right\}.$$

$\lambda(z)$ is said to be the Christoffel function. The solution of this problem for $|z| < 1$ was obtained by Szegő in [41, 42].

In the literature, an important class of measures is the Szegő class $\mathcal{S}$. We summarize some relevant characterizations to the $\mathcal{S}$ class. Indeed, the following conditions are equivalent:

(i) $\sigma \in \mathcal{S}$. (ii) $\displaystyle\int_{-\pi}^{\pi} \log \sigma'(\theta) \frac{d\theta}{2\pi} > -\infty$.

(iii) $\displaystyle\sum_{n=0}^{\infty} |\Phi_n(0)|^2 < \infty$. (iv) $\displaystyle\lambda(0) = \prod_{n=0}^{\infty} (1 - |\Phi_{n+1}(0)|^2) < +\infty$.

From this we deduce that if the measure σ does not belong to the $\mathcal{S}$ class, the GGT matrix $\mathbf{H}_\sigma$ is unitary. In general, $\mathbf{H}_\sigma$ satisfies

(i) $\mathbf{H}_\sigma \mathbf{H}_\sigma^H = \mathbf{I}$; (ii) $\mathbf{H}_\sigma^H \mathbf{H}_\sigma = \mathbf{I} - \lambda(0) \boldsymbol{\phi}(0) \boldsymbol{\phi}(0)^H$.

As a part of the analysis when $\sigma \in \mathcal{S}$, one can construct the Szegő function D, defined in $\mathbb{D}$ as

$$D(z) = \exp\left(\frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} \log \sigma'(\theta) d\theta \right), \quad z \in \mathbb{D}.$$

Thus, $|D|^2 = \sigma'$ almost everywhere on $\mathbb{T}$, and the solution of the Szegő extremum problem is given by

$$\lambda(z) = (1 - |z|^2) |D(z)|^2, \quad z \in \mathbb{D}.$$

1.5 $\mathcal{N}$ Class

We say that σ belongs to the Nevai class $\mathcal{N}$ if

$$\lim_{n\to\infty} \Phi_n(0) = \lim_{n\to\infty} \frac{\phi_n(0)}{\kappa_n} = 0.$$

The relation between the classes $\mathcal{S}$ and $\mathcal{N}$ can be viewed using the results in [30]. If $\sigma \in \mathcal{S}$, then it has a normal L^2-derivative behavior, i.e.,

$$\lim_{n\to\infty} \left(\int_{-\pi}^{\pi} \frac{\left| \phi_n'(e^{i\theta}) \right|^2}{n^2} \sigma'(\theta) d\theta \right)^{\frac{1}{2}} = 1,$$

and thus $\sigma \in \mathcal{N}$. Furthermore, if $\sigma \in \mathcal{N}$,

$$\left| \frac{\Phi_n(z)}{\Phi_{n-1}(z)} - z \right| \leqslant |\Phi_n(0)|, \qquad z \in \mathbb{C} \setminus \mathbb{D}.$$

Thus,

$$\lim_{n\to\infty} \frac{\Phi_n(z)}{\Phi_{n-1}(z)} = z,$$

uniformly in compact subsets of $\mathbb{C} \setminus \overline{\mathbb{D}}$.

This result can be obtained under weaker conditions. A well-known result of Rakhmanov (see [34]) states that any probability measure σ with $\sigma' > 0$, almost everywhere on $\mathbb{T}$, belongs to the class $\mathcal{N}$. The converse is not true at all.

2 Adding the Derivative of a Dirac's Delta

Let $\mathcal{L}$ be a Hermitian linear functional given by (1). Its derivative $D\mathcal{L}$ (see [44]) is defined by

$$\langle D\mathcal{L}, f\rangle = -i\left\langle \mathcal{L}, zf'(z)\right\rangle, \quad f \in \Lambda.$$

In this section we first deal with r a perturbation of a linear functional $\mathcal{L}$ by the addition of a derivative of a Dirac's delta, i.e.,

$$\langle \mathcal{L}_1, f\rangle = \langle \mathcal{L}, f\rangle + m\,\langle D\delta_\alpha, f\rangle, \quad m \in \mathbb{R}, \quad |\alpha| = 1. \tag{8}$$

Let $\mathcal{L}_U$ be a linear functional such that

$$\langle \mathcal{L}_U, f\rangle = \langle \mathcal{L}, f\rangle + mf(\alpha), \quad m \in \mathbb{R}, \quad |\alpha| = 1.$$

We say that $\mathcal{L}_U$ is the Uvarov spectral transformation of the linear functional $\mathcal{L}$ [15]. The connection between the corresponding sequences of monic OPUC as well as the associated GGT matrices using LU and QR factorization has been studied in [15]. Asymptotic properties for the corresponding sequences of OPUC have been obtained in [47]. Notice that the addition of a Dirac's delta derivative (on a point of the unit circle) to a linear functional can be considered as the limit case of two Uvarov spectral transformations with equal masses and opposite sign, located on two nearby points on the unit circle $\alpha_1 = e^{i\theta_1}$ and $\alpha_2 = e^{i\theta_2}$, $0 \leqslant \theta_1, \theta_2 \leqslant 2\pi$, when $\theta_1 \to \theta_2$, but the difficulties to deal with them yield a different approach.

2.1 Mass Point on the Unit Circle

In terms of the associated bilinear functional (8) becomes

$$\langle f, g\rangle_{\mathcal{L}_1} = \langle f, g\rangle_{\mathcal{L}} - im\left(\alpha f'(\alpha)\overline{g(\alpha)} - \overline{\alpha} f(\alpha)\overline{g'(\alpha)}\right). \tag{9}$$

In the next theorem we obtain necessary and sufficient conditions for $\mathcal{L}_1$ to be a quasi-definite linear functional, as well as an expression for its corresponding family of OPUC.

Theorem 1 ([9]). *Let us assume $\mathcal{L}$ is a quasi-definite linear functional and denote by $\{\Phi_n\}_{n\geqslant 0}$ its corresponding sequence of monic OPUC. Let us consider $\mathcal{L}_1$ as in (9). Then, the following statements are equivalent:*

(*i*) *$\mathcal{L}_1$ is quasi-definite.*
(*ii*) *The matrix $\mathbf{D}(\alpha) + m\mathbb{K}_{n-1}(\alpha,\alpha)$, with*

$$\mathbb{K}_{n-1}(\alpha,\alpha) = \begin{bmatrix} K_{n-1}(\alpha,\alpha) & K_{n-1}^{(0,1)}(\alpha,\alpha) \\ K_{n-1}^{(1,0)}(\alpha,\alpha) & K_{n-1}^{(1,1)}(\alpha,\alpha) \end{bmatrix}, \quad \mathbf{D}(\alpha) = \begin{bmatrix} 0 & -i\alpha \\ i\alpha^{-1} & 0 \end{bmatrix},$$

is non-singular, and

$$\mathbf{k}_n + m\mathbf{\Phi}_n(\alpha)^H \left(\mathbf{D}(\alpha) + m\mathbb{K}_{n-1}(\alpha,\alpha)\right)^{-1} \mathbf{\Phi}_n(\alpha) \neq 0, \quad n \geqslant 1.$$

Furthermore, the sequence $\{\Psi_n\}_{n\geqslant 0}$ of monic OPUC associated with $\mathcal{L}_1$ is given by

$$\Psi_n(z) = \Phi_n(z) - m\begin{bmatrix} K_{n-1}(z,\alpha) \\ K_{n-1}^{(0,1)}(z,\alpha) \end{bmatrix}^T \left(\mathbf{D}(\alpha) + m\mathbb{K}_{n-1}(\alpha,\alpha)\right)^{-1} \mathbf{\Phi}_n(\alpha), \tag{10}$$

where $\mathbf{\Phi}_n(\alpha) = \left[\Phi_n(\alpha), \Phi_n'(\alpha)\right]^T$.

Proof. Assume $\mathcal{L}_1$ is quasi-definite and denote by $\{\Psi_n\}_{n\geqslant 0}$ its corresponding family of monic OPUC. Let us consider the Fourier expansion

$$\Psi_n(z) = \Phi_n(z) + \sum_{k=0}^{n-1} \lambda_{n,k}\Phi_k(z),$$

where, for $n \geq 1$,

$$\lambda_{n,k} = \frac{\langle \Psi_n(z), \Phi_k(z)\rangle_{\mathcal{L}}}{\mathbf{k}_k} = \frac{im\left(\alpha\Psi_n'(\alpha)\overline{\Phi_k(\alpha)} - \overline{\alpha}\Psi_n(\alpha)\overline{\Phi_k'(\alpha)}\right)}{\mathbf{k}_k}, \quad 0 \leq k \leq n-1.$$

Thus,

$$\Psi_n(z) = \Phi_n(z) + \sum_{k=0}^{n-1} \frac{im\left(\alpha\Psi_n'(\alpha)\overline{\Phi_k(\alpha)} - \overline{\alpha}\Psi_n(\alpha)\overline{\Phi_k'(\alpha)}\right)}{\mathbf{k}_k}\Phi_k(z),$$

$$= \Phi_n(z) + im\left(\alpha\Psi_n'(\alpha)K_{n-1}(z,\alpha) - \overline{\alpha}\Psi_n(\alpha)K_{n-1}^{(0,1)}(z,\alpha)\right). \tag{11}$$

Taking the derivative with respect to z in the previous expression and evaluating at $z = \alpha$, we obtain the system of linear equations

$$\Psi_n^{(i)}(\alpha) = \Phi_n^{(i)}(\alpha) + im\left(\alpha\Psi_n'(\alpha)K_{n-1}^{(i,0)}(\alpha,\alpha) - \overline{\alpha}\Psi_n(\alpha)K_{n-1}^{(i,1)}(\alpha,\alpha)\right), \quad i = 0, 1,$$

which yields

$$\begin{bmatrix}\Phi_n(\alpha)\\ \Phi_n'(\alpha)\end{bmatrix} = \begin{bmatrix}1 + im\overline{\alpha}K_{n-1}^{(0,1)}(\alpha,\alpha) & -im\alpha K_{n-1}(\alpha,\alpha)\\ im\overline{\alpha}K_{n-1}^{(1,1)}(\alpha,\alpha) & 1 - im\alpha K_{n-1}^{(1,0)}(\alpha,\alpha)\end{bmatrix}\begin{bmatrix}\Psi_n(\alpha)\\ \Psi_n'(\alpha)\end{bmatrix},$$

and denoting $\mathbf{Q} = [Q, Q']^T$, we get

$$\mathbf{\Phi}_n(\alpha) = \left(\mathbf{I}_2 + m\mathbb{K}_{n-1}(\alpha,\alpha)\mathbf{D}(\alpha)\right)\mathbf{\Psi}_n(\alpha).$$

Thus, the necessary condition for regularity is that $\mathbf{I}_2 + m\mathbb{K}_{n-1}(\alpha,\alpha)\mathbf{D}(\alpha)$ must be non-singular. Taking into account $\mathbf{D}^{-1}(\alpha) = \mathbf{D}(\alpha)$ we have the first part of our statement. Furthermore, from (11),

$$\Psi_n(z) = \Phi_n(z) + m\left(K_{n-1}(z,\alpha), K_{n-1}^{(0,1)}(z,\alpha)\right)\begin{bmatrix}0 & i\alpha\\ -i\overline{\alpha} & 0\end{bmatrix}\begin{bmatrix}\Psi_n(\alpha)\\ \Psi_n'(\alpha)\end{bmatrix}$$

$$= \Phi_n(z) - m\begin{bmatrix}K_{n-1}(z,\alpha)\\ K_{n-1}^{(0,1)}(z,\alpha)\end{bmatrix}^T\left(\mathbf{D}(\alpha) + m\mathbb{K}_{n-1}(\alpha,\alpha)\right)^{-1}\mathbf{\Phi}_n(\alpha).$$

This yields (10). Conversely, if $\{\Psi_n\}_{n\geqslant 0}$ is given by (11), then, for $0 \leqslant k \leqslant n-1$,

$$\langle\Psi_n, \Psi_k\rangle_{\mathcal{L}_1} = \left\langle\Phi_n(z) + im\left(\alpha\Psi_n'(\alpha)K_{n-1}(z,\alpha) - \overline{\alpha}\Psi_n(\alpha)K_{n-1}^{(0,1)}(z,\alpha)\right), \Psi_k(z)\right\rangle_{\mathcal{L}_1}$$

$$= \left\langle\Phi_n(z) + im\left(\alpha\Psi_n'(\alpha)K_{n-1}(z,\alpha) - \overline{\alpha}\Psi_n(\alpha)K_{n-1}^{(0,1)}(z,\alpha)\right), \Psi_k(z)\right\rangle_{\mathcal{L}}$$

$$- im\left(\alpha\Psi_n'(\alpha)\overline{\Psi_k(\alpha)} - \overline{\alpha}\Psi_n(\alpha)\overline{\Psi_k'(\alpha)}\right) = 0.$$

On the other hand, for $n \geqslant 1$,

$$\tilde{\mathbf{k}}_n = \langle\Psi_n(z), \Psi_n(z)\rangle_{\mathcal{L}_1} = \langle\Psi_n(z), \Phi_n(z)\rangle_{\mathcal{L}_1}$$

$$= \mathbf{k}_n + m\mathbf{\Phi}_n(\alpha)^H\left(\mathbf{D}(\alpha) + m\mathbb{K}_{n-1}(\alpha,\alpha)\right)^{-1}\mathbf{\Phi}_n(\alpha) \neq 0,$$

where we are using the reproducing property (7). As a conclusion, $\{\Psi_n\}_{n\geqslant 0}$ is the sequence of monic OPUC with respect to $\mathcal{L}_1$. □

From the Christoffel–Darboux formula (6), another way to express (10) is the following.

Corollary 1. *Let $\{\Psi_n\}_{n\geqslant 0}$ be the sequence of monic OPUC associated with $\mathcal{L}_1$ defined as in (9). Then,*

$$(z-\alpha)^2\Psi_n(z) = A(z,n,\alpha)\Phi_n(z) + B(z,n,\alpha)\Phi_n^*(z), \tag{12}$$

where $A(z,n,\alpha)$ and $B(z,n,\alpha)$ are polynomials of degree 2 and 1, respectively, in the variable z, given by

$$\begin{aligned}
A(z,n,\alpha) &= (z-\alpha)^2 - \frac{m\alpha}{\mathbf{k}_n\Delta_{n-1}}\Big((Y_{1,1}\Phi_n(\alpha) + Y_{1,2}\Phi_n'(\alpha))\overline{\Phi_n(\alpha)}(z-\alpha)\\
&\quad + \big(Y_{2,1}\Phi_n(\alpha) + Y_{2,2}\Phi_n'(\alpha)\big)\big(\Phi_n(\alpha)(z-\alpha) + \alpha\Phi_n(\alpha)z\big)\Big),\\
B(z,n,\alpha) &= \frac{m\alpha}{\mathbf{k}_n\Delta_{n-1}}\Big((Y_{1,1}\Phi_n(\alpha) + Y_{1,2}\Phi_n'(\alpha))\overline{\Phi_n^*(\alpha)}\\
&\quad + \Big(Y_{2,1}\Phi_n(\alpha) + Y_{2,2}\Phi_n'(\alpha)\big)\big(\overline{\Phi_n^*(\alpha)}\big)'(z-\alpha) + \alpha\overline{\Phi_n^*(\alpha)}z\big)\Big),
\end{aligned}$$

where $Y_{1,1} = mK_{n-1}^{(1,1)}(\alpha,\alpha)$, $Y_{1,2} = im\alpha K_{n-1}^{(0,1)}(\alpha,\alpha)$, $Y_{2,1} = -im\overline{\alpha}K_{n-1}^{(1,0)}(\alpha,\alpha)$, $Y_{2,2} = m\alpha K_{n-1}(\alpha,\alpha)$, and Δ_{n-1} is the determinant of the matrix $\mathbf{D}(\alpha) + im\mathbb{K}_{n-1}(\alpha,\alpha)$.

2.2 Outer Relative Asymptotics

In this subsection we assume $\mathcal{L}$ is a positive definite linear functional, with associated positive Borel measure σ. We are interested in the asymptotic behavior of the OPUC associated with the addition of the derivative of a Dirac's delta on the unit circle given in (12). We assume that σ is regular in the sense of Stahl and Totik [40], so that

$$\lim_{n\to\infty}\kappa_n^{1/n} = 1.$$

Regularity is a necessary and sufficient condition for the existence of nth root asymptotics, i.e., $\lim_{n\to\infty}|\phi_n|^{1/n} < \infty$. It is easy to see that the existence of the outer ratio asymptotics $\lim_{n\to\infty}\phi_n/\phi_{n-1}$ implies the existence of the root asymptotics, and, in general, the converse is not true.

In particular, we study its outer relative asymptotics with respect to $\{\Phi_n\}_{n\geqslant 0}$. First, we state some results that are useful in our study.

Theorem 2 ([23]). *Let σ be a regular finite positive Borel measure supported on $(-\pi,\pi]$. Let $J\subset(-\pi,\pi)$ be a compact subset such that σ is absolutely continuous*

in an open set containing J. Assume that σ' is positive and continuous at each point of J. Let i, j be non-negative integers. Then, uniformly for $\theta \in J$, $z = e^{i\theta}$,

$$\lim_{n\to\infty} \frac{z^{i-j}}{n^{i+j}} \frac{K_n^{(i,j)}(z,z)}{K_n(z,z)} = \frac{1}{i+j+1}.$$

Lemma 1 ([19]). *Let f, g be two polynomials in $\mathbb{P}$ with degree at least j. Then*

$$\frac{f^{(j)}(z)}{g^{(j)}(z)} = \frac{g^{(j-1)}(z)}{g^{(j)}(z)} \left(\frac{f^{(j-1)}(z)}{g^{(j-1)}(z)}\right)' + \frac{f^{(j-1)}(z)}{g^{(j-1)}(z)}.$$

Using the previous lemma, the outer ratio asymptotics for the derivatives of orthonormal polynomials are deduced.

Lemma 2 ([12]). *Let us assume that $\mathcal{L}$ is a positive definite linear functional, with associated positive Borel measure σ and denote by $\{\phi_n\}_{n\geqslant 0}$ its corresponding sequence of OPUC. If $\sigma \in \mathcal{N}$, then uniformly in $\mathbb{C}\backslash\overline{\mathbb{D}}$*

$$\lim_{n\to\infty} \frac{\phi_{n+1}^{(j)}(z)}{\phi_n^{(j)}(z)} = z, \quad \lim_{n\to\infty} \frac{\phi_n^{(j)}(z)}{\phi_n^{(j+1)}(z)} = 0, \quad j \geqslant 0.$$

Proof. According to Lemma 1,

$$\frac{\phi_{n+1}^{(j)}(z)}{\phi_n^{(j)}(z)} = \frac{\phi_n^{(j-1)}(z)}{\phi_n^{(j)}(z)} \left(\frac{\phi_{n+1}^{(j-1)}(z)}{\phi_n^{(j-1)}(z)}\right)' + \frac{\phi_{n+1}^{(j-1)}(z)}{\phi_n^{(j-1)}(z)}. \tag{13}$$

Using induction in j, we get uniformly in $\mathbb{C}\backslash\overline{\mathbb{D}}$,

$$\lim_{n\to\infty} \left(\frac{\phi_{n+1}^{(j-1)}(z)}{\phi_n^{(j-1)}(z)}\right)' = 1, \quad \lim_{n\to\infty} \frac{\phi_n^{(j-1)}(z)}{\phi_n^{(j)}(z)} = 0.$$

Therefore, if n tends to infinity in (13), the result follows. □

Corollary 2 ([12]). *If $\sigma \in \mathcal{N}$, then uniformly in $\mathbb{C}\backslash\overline{\mathbb{D}}$*

$$\lim_{n\to\infty} \frac{\phi_n^{*(j)}(z)}{\phi_n^{(j)}(z)} = 0, \quad \lim_{n\to\infty} \frac{K_{n-1}^{(l,r)}(z,y)}{\phi_n^{(i)}(z)\overline{\phi_n^{(j)}(y)}} = 0, \quad 0 \leqslant l < i, 0 \leqslant r < j.$$

From the expression (12),

$$\frac{\Psi_n(z)}{\Phi_n(z)} = \frac{A(z,n,\alpha)}{(z-\alpha)^2} + \frac{B(z,n,\alpha)}{(z-\alpha)^2} \frac{\Phi_n^*(z)}{\Phi_n(z)}.$$

Since, for $z \in \mathbb{C} \setminus \overline{\mathbb{D}}$ by Corollary 2,

$$\lim_{n\to\infty} \frac{\Phi_n^*(z)}{\Phi_n(z)} = 0,$$

it suffices to show that, for $|\alpha| = 1$,

$$\lim_{n\to\infty} \frac{A(z,n,\alpha)}{(z-\alpha)^2} = 1.$$

Notice that $\lim_{n\to\infty} \Phi_n(\alpha) = \mathcal{O}(1)$, $\lim_{n\to\infty} \Phi_n'(\alpha) = \mathcal{O}(n)$, $\lim_{n\to\infty} \Phi_n^*(\alpha) = \mathcal{O}(1)$, $\lim_{n\to\infty} \Phi_n^{*\prime}(\alpha) = \mathcal{O}(n)$, and $\lim_{n\to\infty} K_n(\alpha,\alpha) = \mathcal{O}(n)$.

On the other hand, dividing the numerator and denominator of $\dfrac{A(z,n,\alpha)}{(z-\alpha)^2} - 1$ by $n^2K_{n-1}(\alpha,\alpha)$, and using Theorem 2, we obtain

$$\lim_{n\to\infty} \frac{\Phi_n(\alpha)Y_{2,1}}{n^2K_{n-1}(\alpha,\alpha)} = \mathcal{O}(1/n), \qquad \lim_{n\to\infty} \frac{\Phi_n'(\alpha)Y_{2,2}}{n^2K_{n-1}(\alpha,\alpha)} = \mathcal{O}(1/n),$$

$$\lim_{n\to\infty} \frac{\Phi_n(\alpha)Y_{1,1}}{n^2K_{n-1}(\alpha,\alpha)} = \mathcal{O}(1), \qquad \lim_{n\to\infty} \frac{\Phi_n'(\alpha)Y_{1,2}}{n^2K_{n-1}(\alpha,\alpha)} = \mathcal{O}(1).$$

As a consequence, the numerator of $\dfrac{A(z,n,\alpha)}{(z-\alpha)^2} - 1$ behaves as $\mathcal{O}(1)$. Similarly, one can show that the denominator behaves as $\mathcal{O}(n)$ and, therefore,

$$\lim_{n\to\infty} \frac{A(z,n,\alpha)}{(z-\alpha)^2} = 1.$$

The same arguments can be applied to $B(z,n,\alpha)$. Thus, we get

Theorem 3 ([9]). *Let $\mathcal{L}$ be a positive definite linear functional, whose associated measure σ satisfies the conditions of Theorem 2. Let $\{\Psi_n\}_{n\geq 0}$ be the sequence of monic OPUC associated with $\mathcal{L}_1$ defined as in (9). Then, uniformly in $\mathbb{C}\setminus\overline{\mathbb{D}}$,*

$$\lim_{n\to\infty} \frac{\Psi_n(z)}{\Phi_n(z)} = 1.$$

2.3 Mass Points Outside the Unit Circle

Now, consider a hermitian linear functional $\mathcal{L}_2$ such that its associated bilinear functional satisfies

$$\langle f,g\rangle_{\mathcal{L}_2} = \langle f,g\rangle_{\mathcal{L}} + im\left(\alpha^{-1}f(\alpha)\overline{g'(\overline{\alpha}^{-1})} - \alpha f'(\alpha)\overline{g(\overline{\alpha}^{-1})}\right)$$
$$+ i\overline{m}\left(\overline{\alpha}f(\overline{\alpha}^{-1})\overline{g'(\alpha)} - \overline{\alpha}^{-1}p'(\overline{\alpha}^{-1})\overline{q(\alpha)}\right), \tag{14}$$

with $m, \alpha \in \mathbb{C}$, $|\alpha| \neq 0$, and $|\alpha| \neq 1$. As in the previous section, we are interested in the regularity conditions for this linear functional and the corresponding family of OPUC. Assuming that $\mathcal{L}_2$ is a quasi-definite linear functional and following the method used in the proof of Theorem 1, we get

$$\begin{aligned}\Psi_n(z) = \Phi_n(z) &+ im\left(\alpha\Psi_n'(\alpha)K_{n-1}(z,\overline{\alpha}^{-1}) - \alpha^{-1}\Psi_n(\alpha)K_{n-1}^{(0,1)}(z,\overline{\alpha}^{-1})\right)\\ &+ i\overline{m}\left(\overline{\alpha}^{-1}\Psi_n'(\overline{\alpha}^{-1})K_{n-1}(z,\alpha) - \overline{\alpha}\Psi_n(\overline{\alpha}^{-1})K_{n-1}^{(0,1)}(z,\alpha)\right). \end{aligned} \tag{15}$$

Evaluating the above expression and its first derivative in α and $\overline{\alpha}^{-1}$, we get the following systems of linear equations

$$\begin{aligned}\begin{bmatrix}\Phi_n(\alpha)\\ \Phi_n'(\alpha)\end{bmatrix} &= \begin{bmatrix}1 + im\alpha^{-1}K_{n-1}^{(0,1)}(\alpha,\overline{\alpha}^{-1}) & -im\alpha K_{n-1}(\alpha,\overline{\alpha}^{-1})\\ im\alpha^{-1}K_{n-1}^{(1,1)}(\alpha,\overline{\alpha}^{-1}) & 1 - im\alpha K_{n-1}^{(1,0)}(\alpha,\overline{\alpha}^{-1})\end{bmatrix}\begin{bmatrix}\Psi_n(\alpha)\\ \Psi_n'(\alpha)\end{bmatrix}\\ &+ \begin{bmatrix}i\overline{m}\overline{\alpha}K_{n-1}^{(0,1)}(\alpha,\alpha) & -i\overline{m}\overline{\alpha}^{-1}K_{n-1}(\alpha,\alpha)\\ i\overline{m}\overline{\alpha}K_{n-1}^{(1,1)}(\alpha,\alpha) & -i\overline{m}\overline{\alpha}^{-1}K_{n-1}^{(1,0)}(\alpha,\alpha)\end{bmatrix}\begin{bmatrix}\Psi_n(\overline{\alpha}^{-1})\\ \Psi_n'(\overline{\alpha}^{-1})\end{bmatrix}, \end{aligned} \tag{16}$$

$$\begin{aligned}\begin{bmatrix}\Phi_n(\overline{\alpha}^{-1})\\ \Phi_n'(\overline{\alpha}^{-1})\end{bmatrix} &= \begin{bmatrix}im\alpha^{-1}K_{n-1}^{(0,1)}(\overline{\alpha}^{-1},\overline{\alpha}^{-1}) & -im\alpha K_{n-1}(\overline{\alpha}^{-1},\overline{\alpha}^{-1})\\ im\alpha^{-1}K_{n-1}^{(1,1)}(\overline{\alpha}^{-1},\overline{\alpha}^{-1}) & -im\alpha K_{n-1}^{(1,0)}(\overline{\alpha}^{-1},\overline{\alpha}^{-1})\end{bmatrix}\begin{bmatrix}\Psi_n(\alpha)\\ \Psi_n'(\alpha)\end{bmatrix}\\ &+ \begin{bmatrix}1 + i\overline{m}\overline{\alpha}K_{n-1}^{(0,1)}(\overline{\alpha}^{-1},\alpha) & -i\overline{m}\overline{\alpha}^{-1}K_{n-1}(\overline{\alpha}^{-1},\alpha)\\ i\overline{m}\overline{\alpha}K_{n-1}^{(1,1)}(\overline{\alpha}^{-1},\alpha) & 1 - i\overline{m}\overline{\alpha}^{-1}K_{n-1}^{(1,0)}(\overline{\alpha}^{-1},\alpha)\end{bmatrix}\begin{bmatrix}\Psi_n(\overline{\alpha}^{-1})\\ \Psi_n'(\overline{\alpha}^{-1})\end{bmatrix}, \end{aligned} \tag{17}$$

which reads as a system of four linear equations with four unknowns

$$\begin{bmatrix}\mathbf{\Phi}_n(\alpha)\\ \mathbf{\Phi}_n(\overline{\alpha}^{-1})\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 + m\mathbb{K}_{n-1}(\alpha,\overline{\alpha}^{-1})\mathbf{D}(\alpha) & \overline{m}\mathbb{K}_{n-1}(\alpha,\alpha)\mathbf{D}(\overline{\alpha}^{-1})\\ m\mathbb{K}_{n-1}(\overline{\alpha}^{-1},\overline{\alpha}^{-1})\mathbf{D}(\alpha) & \mathbf{I}_2 + \overline{m}\mathbb{K}_{n-1}(\overline{\alpha}^{-1},\alpha)\mathbf{D}(\overline{\alpha}^{-1})\end{bmatrix}\begin{bmatrix}\mathbf{\Psi}_n(\alpha)\\ \mathbf{\Psi}_n(\overline{\alpha}^{-1})\end{bmatrix},$$

where $(\mathbf{Q},\mathbf{R})^T = (Q,Q',R,R')^T$. Thus, in order $\mathcal{L}_2$ to be a quasi-definite linear functional, we need that the 4×4 matrix defined as above must be non-singular. On the other hand,

$$\begin{bmatrix}\mathbf{\Psi}_n(\alpha)\\ \mathbf{\Psi}_n(\overline{\alpha}^{-1})\end{bmatrix} = \begin{bmatrix}\mathbf{I}_2 + m\mathbb{K}_{n-1}(\alpha,\overline{\alpha}^{-1})\mathbf{D}(\alpha) & \overline{m}\mathbb{K}_{n-1}(\alpha,\alpha)\mathbf{D}(\overline{\alpha}^{-1})\\ m\mathbb{K}_{n-1}(\overline{\alpha}^{-1},\overline{\alpha}^{-1})\mathbf{D}(\alpha) & \mathbf{I}_2 + \overline{m}\mathbb{K}_{n-1}(\overline{\alpha}^{-1},\alpha)\mathbf{D}(\overline{\alpha}^{-1})\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{\Phi}_n(\alpha)\\ \mathbf{\Phi}_n(\overline{\alpha}^{-1})\end{bmatrix}.$$

As a consequence, from (15), we get

$$\Psi_n(z) = \Phi_n(z) - m \begin{bmatrix} K_{n-1}(z, \overline{\alpha}^{-1}) \\ K_{n-1}^{(0,1)}(z, \overline{\alpha}^{-1}) \end{bmatrix}^T \mathbf{D}(\alpha)\mathbf{\Psi}_n(\alpha) - \overline{m} \begin{bmatrix} K_{n-1}(z, \alpha) \\ K_{n-1}^{(0,1)}(z, \alpha) \end{bmatrix}^T \mathbf{D}(\overline{\alpha}^{-1})\mathbf{\Psi}_n(\overline{\alpha}^{-1}), \tag{18}$$

where $\mathbf{\Psi}_n(\alpha)$ and $\mathbf{\Psi}_n(\overline{\alpha}^{-1})$ can be obtained from the above linear system. Assuming that the regularity conditions hold and following the method used in the proof of Theorem 1, it is not difficult to show that $\{\Psi_n\}_{n\geq 0}$, defined as in (18), is the sequence of monic OPUC with respect to $\mathcal{L}_2$.

The following result was proved in [16] using a different method, and it has been generalized for rectifiable Jordan curves or arcs in [3]. We show here another proof of the same result.

Lemma 3 ([12]). *If $\sigma \in \mathcal{N}$, then uniformly in $\mathbb{C}\backslash\overline{\mathbb{D}}$,*

$$\lim_{n\to\infty} \frac{K_{n-1}^{(i,j)}(z, y)}{\phi_n^{(i)}(z)\overline{\phi_n^{(j)}(y)}} = \frac{1}{z\bar{y} - 1}, \qquad i, j \geq 0.$$

Proof. From the Christoffel–Darboux formula (6), we obtain

$$\phi_n^*(z)\overline{\phi_n^{*(j)}(y)} - \phi_n(z)\overline{\phi_n^{(j)}(y)} = (1 - z\bar{y})K_{n-1}^{(0,j)}(z, y) - jzK_{n-1}^{(0,j-1)}(z, y),$$

and, as a consequence,

$$\begin{aligned}\phi_n^{*(i)}(z)\overline{\phi_n^{*(j)}(y)} - \phi_n^{(i)}(z)\overline{\phi_n^{(j)}(y)} = {} & (1 - z\bar{y})K_{n-1}^{(i,j)}(z, y) - k\bar{y}K_{n-1}^{(i-1,j)}(z, y) \\ & - j\left(zK_{n-1}^{(i,j-1)}(z, y) + kK_{n-1}^{(k-1,j-1)}(z, y)\right).\end{aligned}$$

Thus, dividing by $\phi_n^{(i)}(z)\overline{\phi_n^{(j)}(y)}$ and using Corollary 2 when n tends to infinity, the result follows. □

It is possible to obtain a generalization of Theorem 3 for the sequence of monic OPUC associated with (14). As above, we can express (18) as in (12). Using the Christoffel–Darboux formula (6), we obtain

$$\Psi_n(z) = \left(1 + \tilde{A}(z, n, \alpha)\right)\Phi_n(z) + \tilde{B}(z, n, \alpha)\Phi_n^*(z),$$

with

$$\begin{aligned}\tilde{A}(z, n, \alpha) = {} & im\overline{\alpha}^{-1}\frac{\overline{\Phi_n'(\overline{\alpha}^{-1})}(1 - \alpha^{-1}z) + z\overline{\Phi_n(\overline{\alpha}^{-1})}}{\mathbf{k}_n(1 - \alpha^{-1})^2}\Psi_n(\alpha) - im\alpha\frac{\overline{\Phi_n(\overline{\alpha}^{-1})}}{\mathbf{k}_n(1 - \alpha^{-1}z)}\Psi_n'(\alpha) \\ & + i\overline{m}\alpha\frac{\overline{\Phi_n'(\alpha)}(1 - \overline{\alpha}z) + z\overline{\Phi_n(\alpha)}}{\mathbf{k}_n(1 - \overline{\alpha})^2}\Psi_n(\overline{\alpha}^{-1}) - i\overline{m}\overline{\alpha}^{-1}\frac{\overline{\Phi_n(\alpha)}}{\mathbf{k}_n(1 - \overline{\alpha}z)}\Psi_n'(\overline{\alpha}^{-1}),\end{aligned}$$

$$\tilde{B}(z,n,\alpha)=im\alpha\frac{\overline{\Phi_n^*(\overline{\alpha}^{-1})}}{\mathbf{k}_n(1-\alpha^{-1}z)}\Psi_n'(\alpha)-im\overline{\alpha}^{-1}\frac{\overline{\Phi_n'^*(\overline{\alpha}^{-1})}(1-\alpha^{-1}z)+z\overline{\Phi_n^*(\overline{\alpha}^{-1})}}{\mathbf{k}_n(1-\alpha^{-1})^2}\Psi_n(\alpha)$$
$$+\,i\overline{m}\overline{\alpha}^{-1}\frac{\overline{\Phi_n^*(\alpha)}}{\mathbf{k}_n(1-\overline{\alpha}z)}\Psi_n'(\overline{\alpha}^{-1})-i\overline{m}\alpha\frac{\overline{\Phi_n'^*(\alpha)}(1-\overline{\alpha}z)+z\overline{\Phi_n^*(\alpha)}}{\mathbf{k}_n(1-\overline{\alpha})^2}\Psi_n(\overline{\alpha}^{-1}),$$

where the values of $\Psi_n(\alpha)$, $\Psi_n'(\alpha)$, $\Psi_n(\overline{\alpha}^{-1})$, and $\Psi_n'(\overline{\alpha}^{-1})$ can be obtained by solving the 4×4 linear system shown above. Denoting the entries of the 2×2 matrices in (16), (17) by $\{b_{i,j}\}$, $\{c_{i,j}\}$, $\{a_{i,j}\}$ and $\{d_{i,j}\}$, respectively, we get

$$\Psi_n(\alpha)=\left(d_{1,1}\Phi_n(\alpha)+d_{1,2}\Phi_n'(\alpha)+c_{1,1}\Phi_n(\overline{\alpha}^{-1})+c_{1,2}\Phi_n'(\overline{\alpha}^{-1})\right)/\Delta,$$
$$\Psi_n'(\alpha)=\left(d_{2,1}\Phi_n(\alpha)+d_{2,2}\Phi_n'(\alpha)+c_{2,1}\Phi_n(\overline{\alpha}^{-1})+c_{2,2}\Phi_n'(\overline{\alpha}^{-1})\right)/\Delta,$$
$$\Psi_n(\overline{\alpha}^{-1})=\left(a_{1,1}\Phi_n(\alpha)+a_{1,2}\Phi_n'(\alpha)+b_{1,1}\Phi_n(\overline{\alpha}^{-1})+b_{1,2}\Phi_n'(\overline{\alpha}^{-1})\right)/\Delta,$$
$$\Psi_n(\overline{\alpha}^{-1})=\left(a_{2,1}\Phi_n(\alpha)+a_{2,2}\Phi_n'(\alpha)+b_{2,1}\Phi_n(\overline{\alpha}^{-1})+b_{2,2}\Phi_n'(\overline{\alpha}^{-1})\right)/\Delta,$$

where Δ is the determinant of the 4×4 matrix. To get the asymptotic result, it suffices to show that $\tilde{A}(z,n,\alpha)\to 0$ and $\tilde{B}(z,n,\alpha)\to 0$ as $n\to\infty$. First, notice that applying the corresponding derivatives to the Christoffel–Darboux formula (6), we obtain

$$K_{n-1}^{(0,1)}(z,y)=\frac{\overline{\Phi_n^{*'}(y)}\Phi_n^*(z)-\overline{\Phi_n'(y)}\Phi_n(z)}{\mathbf{k}_n(1-\overline{y}z)}+\frac{zK_{n-1}(z,y)}{1-\overline{y}z},$$
$$K_{n-1}^{(1,0)}(z,y)=\frac{\overline{\Phi_n^*(y)}\Phi_n^{*'}(z)-\overline{\Phi_n(y)}\Phi_n'(z)}{\mathbf{k}_n(1-\overline{y}z)}+\frac{\overline{y}K_{n-1}(z,y)}{1-\overline{y}z},$$
$$K_{n-1}^{(1,1)}(z,y)=\frac{\overline{\Phi_n^{*'}(y)}\Phi_n^{*'}(z)-\overline{\Phi_n'(y)}\Phi_n'(z)}{\mathbf{k}_n(1-\overline{y}z)}$$
$$+\frac{zK_{n-1}^{(1,0)}(z,y)+\overline{y}K_{n-1}^{(0,1)}(z,y)+K_{n-1}(z,y)}{1-\overline{y}z}.$$

On the other hand, if $\mathcal{L}$ is positive definite and its corresponding measure $\sigma\in\mathcal{N}$, then by Corollary 2 (see also [31]) we have $\Phi_n(\alpha)=\mathcal{O}(\alpha^n)$, $\Phi_n'(\alpha)=\mathcal{O}(n\alpha^n)$, and

$$\lim_{n\to\infty}\frac{\Phi_n(\alpha)}{\Phi_n^*(\alpha)}=0,\quad |\alpha|<1,\qquad \lim_{n\to\infty}\frac{\Phi_n^*(\alpha)}{\Phi_n(\alpha)}=0,\quad |\alpha|>1.$$

Assume, without loss of generality, that $|\alpha|<1$. If $|\alpha|<1$ and $\sigma\in\mathcal{S}$, notice that $\Phi_n(\alpha)$ and $\Phi_n^*(\alpha)$ are $\mathcal{O}(\alpha^n)$, then $\lim_{n\to\infty}K_n(\alpha,\alpha)<\infty$ and $K_n(\overline{\alpha}^{-1},\overline{\alpha}^{-1})=\mathcal{O}(|\alpha|^{-2n})$, as well as $\overline{K_n(\alpha,\overline{\alpha}^{-1})}=K_n(\overline{\alpha}^{-1},\alpha)=\mathcal{O}(n)$. Observe that, except for the entries containing $K_{n-1}(\alpha,\alpha)$ and their derivatives, all other entries of the 4×4 matrix diverge, and thus its determinant diverges faster than any other term in the expressions for $\Psi_n(\alpha)$, $\Psi_n'(\alpha)$, $\Psi_n(\overline{\alpha}^{-1})$ and $\Psi_n'(\overline{\alpha}^{-1})$, so that $\tilde{A}(z,n,\alpha)\to 0$ and $\tilde{B}(z,n,\alpha)\to 0$ as n tends to ∞. As a consequence,

Theorem 4 ([9]). *Let $\mathcal{L}$ be a positive definite linear functional, whose associated measure $\sigma \in \mathcal{S}$. Let $\{\Psi_n\}_{n\geqslant 0}$ be the sequence of monic OPUC associated with $\mathcal{L}_2$ defined as in (14). Then, uniformly in $\mathbb{C} \setminus \mathbb{T}$,*

$$\lim_{n\to\infty} \frac{\Psi_n(z)}{\Phi_n(z)} = 1.$$

3 Sobolev Inner Products

In the last few years, some attention has been paid to the asymptotic properties of OPUC with respect to non-standard inner products. In particular, the algebraic and analytic properties of orthogonal polynomials associated with a Sobolev inner product have attracted the interest of many researchers, see [28] for an updated overview with more than 300 references.

A discrete Sobolev inner product in $\mathbb{C} \setminus \overline{\mathbb{D}}$ is given by

$$\langle f, g\rangle_S = \int_{\mathbb{T}} f(z)\overline{g(z)}d\sigma(z) + \mathbf{f}(Z)\ \mathbf{A}\ \mathbf{g}(Z)^H, \tag{19}$$

where

$$\mathbf{f}(Z) = \left(f(\alpha_1), \ldots, f^{(l_1)}(\alpha_1), \ldots, f(\alpha_m), \ldots, f^{(l_m)}(\alpha_m)\right),$$

$\mathbf{A}$ is an $M \times M$ positive semi-definite hermitian matrix, with $M = l_1 + \cdots + l_m + m$, and $|\alpha_i| > 1$, $i = 1, \ldots, m$. Since $\mathbf{A}$ is a positive semi-definite matrix, the inner product (19) is positive definite. Therefore, there exists a sequence of polynomials $\{\psi_n\}_{n\geqslant 0}$,

$$\psi_n(z) = \gamma_n z^n + (\text{lower degree terms}), \quad \gamma_n > 0,$$

which is orthonormal with respect to (19). We are interested in the outer relative asymptotic behavior of $\{\psi_n\}_{n\geqslant 0}$ with respect to the sequence $\{\phi_n\}_{n\geqslant 0}$ of OPUC with respect to σ. We show that if $\sigma \in \mathcal{N}$ and $\mathbf{A}$ is positive definite, then this outer relative asymptotics follows. Similar results have been obtained when the measure is supported on a bounded interval of the real line [25, 29].

3.1 Outer Relative Asymptotics

In [13, 16, 24, 26], the relative asymptotic behavior of orthogonal polynomials with respect to a discrete Sobolev inner product on the unit circle was studied. In this section, we propose a slightly modified outline.

The nondiagonal structure of the matrix $\mathbf{A}$ makes the analysis of the situation much more difficult. First of all, let us prove an important result which gives a precise information about the matrix $\mathbf{A}$.

Lemma 4 ([6]). *The outer relative asymptotic behavior of orthogonal polynomials with respect to the inner product* (19) *does not depend on the matrix* $\mathbf{A}$.

Proof. Let $\{\tilde{\psi}_n\}_{n\geq 0}$ be the sequence of orthonormal polynomials with respect to the inner product

$$\langle f, g\rangle_{\tilde{S}} = \int_{\mathbb{T}} f(z)\overline{g(z)}d\sigma(z) + \mathbf{f}(Z)\ \mathbf{B}\ \mathbf{g}(Z)^H,$$

where $\mathbf{B}$ is an arbitrary positive definite Hermitian matrix of order M. Expanding ψ_n in terms of $\{\phi_n\}_{n\geq 0}$, we have

$$\psi_n(z) = \frac{\gamma_n}{\kappa_n}\phi_n(z) + \sum_{k=0}^{n-1}\lambda_{n,k}\phi_k(z) \tag{20}$$

where

$$\lambda_{n,k} = \int_{\mathbb{T}} \psi_n(z)\overline{\phi(z)}d\sigma(z) = -\boldsymbol{\psi}_n(Z)\mathbf{A}\boldsymbol{\phi}_n(Z).$$

Substituting this expression in (20), we obtain

$$\psi_n(z) = \frac{\gamma_n}{\kappa_n}\phi_n(z) - \boldsymbol{\psi}_n(Z)\ \mathbf{A}\ \mathbf{K}_n(z, Z)^T, \tag{21}$$

where $\mathbf{K}_n(z, Z) = (K_n(z, \alpha_1), \ldots, K_n^{(0,l_1)}(z, \alpha_1)), \ldots, K_n(z, \alpha_m), \ldots K_n^{(0,l_m)}(z, \alpha_m))$ and $K_n^{(i,j)}(z, y)$ denotes the ith (resp. jth) partial derivative of $K_n(z, y)$ with respect to the variable z (resp. y). In an analogous way, we get

$$\tilde{\psi}_n(z) = \frac{\tilde{\gamma}_n}{\kappa_n}\phi_n(z) - \tilde{\boldsymbol{\psi}}_n(Z)\ \mathbf{B}\ \mathbf{K}_n(z, Z)^T, \tag{22}$$

where $\tilde{\gamma}_n$ is the leading coefficient of ψ_n. From (21) and (22) and following the method used in the proof of Theorem 1, we get [16, 24]

$$\frac{\tilde{\gamma}_n}{\gamma_n}\frac{\tilde{\psi}_n(z)}{\psi_n(z)} = \frac{\det(\mathbf{I} + \mathbf{A}\mathbb{T}_n)}{\det(\mathbf{I} + \mathbf{B}\mathbb{T}_n)}\frac{\det(\mathbf{I} + \mathbf{B}\mathbb{K}_n)}{\det(\mathbf{I} + \mathbf{A}\mathbb{K}_n)},$$

$$\left(\frac{\tilde{\gamma}_n}{\gamma_n}\right)^2 = \frac{\det(\mathbf{I} + \mathbf{B}\mathbb{K}_n)}{\det(\mathbf{I} + \mathbf{A}\mathbb{K}_n)}\frac{\det(\mathbf{I} + \mathbf{A}\mathbb{K}_{n+1})}{\det(\mathbf{I} + \mathbf{B}\mathbb{K}_{n+1})},$$

where $\mathbb{K}_n$ is a positive definite matrix of order M, $n \geq M$, which can be described by blocks. The r, s block of $\mathbb{K}_n$ is the $(l_r + 1) \times (l_s + 1)$ matrix

$$\left(K_n^{(i,j)}(z_r, \overline{z_s})\right)_{i=0,\ldots,l_r}^{j=0,\ldots,l_s}, \qquad r, s = 0, \ldots, m.$$

$\mathbb{T}_n$ is obtained through the following equation $\mathbb{T}_n = \mathbb{K}_n + \mathbb{V}_n$, where $\mathbb{V}_n = -\dfrac{1}{\phi_n(z)}$ $\mathbf{K}_n(z, Z)^T \boldsymbol{\phi}_n(Z)$. Since [16, 26]

$$\lim_{n\to\infty} \frac{\det(\mathbf{I} + \mathbf{A}\mathbb{K}_n)}{\det(\mathbf{I} + \mathbf{B}\mathbb{K}_n)} = \lim_{n\to\infty} \frac{\det(\mathbf{I} + \mathbf{A}\mathbb{T}_n)}{\det(\mathbf{I} + \mathbf{B}\mathbb{T}_n)} = \frac{\det \mathbf{A}}{\det \mathbf{B}},$$

we can deduce that

$$\lim_{n\to\infty} \frac{\tilde{\gamma}_n}{\gamma_n} \frac{\tilde{\psi}_n(z)}{\psi_n(z)} = 1, \quad \lim_{n\to\infty} \left(\frac{\tilde{\gamma}_n}{\gamma_n}\right)^2 = 1,$$

and the lemma is proved. □

For the discrete Sobolev inner product with a single mass point associated with (19),

$$\langle f, g\rangle_{S_1} = \int_{\mathbb{T}} f(z)\overline{g(z)} d\sigma(z) + \lambda f^{(j)}(\alpha)\overline{g^{(j)}(\alpha)}, \quad |\alpha| > 1, \tag{23}$$

we have

Lemma 5 ([6]). *Let $\{\psi_{n;1}\}_{n\geq 0}$, $\psi_{n;1} = \gamma_{n;1} z^n + (\text{lower degree terms})$ be the sequence of orthonormal polynomials with respect to (23). If $\sigma \in \mathcal{N}$, then*

$$\lim_{n\to\infty} \frac{\gamma_{n,1}}{\kappa_n} = \frac{1}{|\alpha|}.$$

Proof. From (21) we have

$$\psi_{n;1}(z) = \frac{\gamma_{n;1}}{\kappa_n}\phi_n(z) - \lambda\psi_{n;1}^{(j)}(\alpha)K_{n-1}^{(0,j)}(z, \alpha). \tag{24}$$

Taking derivatives in (24) and evaluating at $z = \alpha$, we get

$$\psi_{n;1}^{(j)}(\alpha) = \frac{\gamma_{n;1}/\kappa_n\, \phi_n^{(j)}(\alpha)}{1 + \lambda K_{n-1}^{(j,j)}(\alpha, \alpha)}. \tag{25}$$

Thus, (25) yields

$$\left(\frac{\beta_n}{\alpha_n}\right)^2 = \frac{1 + \lambda K_{n-1}^{(j,j)}(\alpha, \alpha)}{1 + \lambda K_n^{(j,j)}(\alpha, \alpha)}.$$

Using the previous identity and Lemma 2,

$$\lim_{n\to\infty}\frac{\gamma_{n,1}^2}{\kappa_n^2}=\lim_{n\to\infty}\frac{1+\lambda K_{n-1}^{(j,j)}(\alpha,\alpha)}{1+\lambda K_n^{(j,j)}(\alpha,\alpha)}=\lim_{n\to\infty}\frac{|\phi_{n-1}^{(j)}(\alpha)|^2}{|\phi_n^{(j)}(\alpha)|^2}=\frac{1}{|\alpha|^2},$$

and the lemma is proved. □

Using the previous lemma, we prove the relative asymptotics in $\mathbb{C}\backslash\overline{\mathbb{D}}$.

Theorem 5 ([6]). *If $\sigma\in\mathcal{N}$, then uniformly in $\mathbb{C}\backslash\overline{\mathbb{D}}$*

$$\lim_{n\to\infty}\frac{\psi_{n;1}(z)}{\phi_n(z)}=B(z;\alpha),\quad B(z;\alpha)=\frac{\overline{\alpha}(z-\alpha)}{|\alpha|(\overline{\alpha}z-1)}.\tag{26}$$

Proof. From (24), we have

$$\frac{\psi_{n;1}(z)}{\phi_n(z)}=\frac{\gamma_{n;1}}{\kappa_n}-\lambda\psi_{n;1}^{(j)}(\alpha)\overline{\phi_n^{(j)}(\alpha)}\frac{K_{n-1}^{(0,j)}(z,\alpha)}{\phi_n(z)\overline{\phi_n^{(j)}(\alpha)}}.\tag{27}$$

Using (25), we obtain

$$\lim_{n\to\infty}\lambda\psi_{n;1}^{(j)}(\alpha)\overline{\phi_n^{(j)}(\alpha)}=\left(|\alpha|-\frac{1}{|\alpha|}\right).\tag{28}$$

The outer relative asymptotics (26) follows letting n tends to infinity in (27), using Lemmas 5, 3, and (28). □

From Theorem 5 we can see that the outer relative asymptotic behavior of orthogonal polynomials associated with (23) does not depend on the specific choice of j and λ.

Lemma 6 ([6]). *$\sigma\in\mathcal{N}$, then $S_1\in\mathcal{N}$.*

Proof. Assume, without loss of generality, that $j=0$ and $\lambda=1$. From (24) and (25) we get

$$\psi_{n;1}(z)=\frac{\gamma_{n;1}}{\kappa_n}\phi_n(z)-\frac{\phi_n(\alpha)}{1+K_{n-1}(\alpha,\alpha)}K_{n-1}(z,\alpha).\tag{29}$$

The evaluation at $z=0$ of this last expression yields

$$\frac{\psi_{n;1}(0)}{\gamma_{n;1}}=\frac{\phi_n(0)}{\kappa_n}-\frac{|\phi_n(\alpha)|^2}{1+K_{n-1}(\alpha,\alpha)}\frac{K_{n-1}(0,\alpha)}{\gamma_{n;1}\overline{\phi_n(\alpha)}},$$

and using the Christoffel–Darboux formula (6), we obtain

$$\frac{K_{n-1}(0,\alpha)}{\gamma_{n;1}\overline{\phi_n(\alpha)}}=\frac{\kappa_n}{\gamma_{n;1}}\left(\frac{\overline{\phi_n^*(\alpha)}}{\overline{\phi_n(\alpha)}}-\frac{\phi_n(0)}{\kappa_n}\right).$$

From Corollary 2, under our conditions, the following limit holds $\lim_{n\to\infty} \overline{\frac{\phi_n^*(\alpha)}{\phi_n(\alpha)}} = 0$. Since $K_n(\alpha,\alpha)$ is an increasing sequence and $\lim_{n\to\infty} \frac{1}{\phi_n(\alpha)} = 0$, applying the Stolz–Césaro criterion, we have

$$\lim_{n\to\infty} \frac{|\phi_n(\alpha)|^2}{1+K_n(\alpha,\alpha)} = \left(1-\frac{1}{|\alpha|^2}\right). \tag{30}$$

On the other hand, from (29) we can deduce the following identity

$$\frac{|\phi_n(\alpha)|^2}{1+K_n(\alpha,\alpha)} = 1 - \frac{1+K_{n-1}(\alpha,\alpha)}{1+K_n(\alpha,\alpha)} = 1 - \left(\frac{\gamma_{n;1}}{\kappa_n}\right)^2.$$

Thus,

$$\lim_{n\to\infty} \frac{K_{n-1}(0,\alpha)}{\gamma_{n;1}\,\overline{\phi_n(\alpha)}} = 0$$

and the result follows. □

We are now in a position to summarize the results obtained above. Indeed,

Theorem 6 ([6]). *Let $\{\psi_n\}_{n\geqslant 0}$ be the sequence of monic orthogonal polynomials associated with the inner product* (19)*. Then, uniformly in $\mathbb{C}\setminus\mathbb{D}$,*

$$\lim_{n\to\infty} \frac{\psi_n(z)}{\phi_n(z)} = \prod_{i=1}^{m} B(z;\alpha_i)^{l_i+1}.$$

Proof. First of all, we prove the result for

$$\mathbf{f}(Z) = \mathbf{f}_m(Z) = \left(f^{(l_1)}(\alpha_1),\ldots,f^{(l_m)}(\alpha_m)\right),$$

and $\mathbf{A}_m$ a positive definite Hermitian matrix of order m. Let $\{\psi_{n;m}\}_{n\geqslant 0}$ be the sequence of orthonormal polynomials with respect to (19) for $\mathbf{f}(Z) = \mathbf{f}_m(Z)$. We can assume, without loss of generality, $\mathbf{A}_m = \mathbf{I}_m$ by Lemma 4. Therefore, the relative asymptotics can be written as follows:

$$\lim_{n\to\infty} \frac{\psi_{n;m}(z)}{\phi_n(z)} = \lim_{n\to\infty} \frac{\psi_{n;1}(z)}{\phi_n(z)} \prod_{i=2}^{m} \frac{\psi_{n;i}(z)}{\psi_{n;i-1}(z)},$$

which, using Lemma 6 and Theorem 5, immediately yields

$$\lim_{n\to\infty} \frac{\psi_{n;m}(z)}{\phi_n(z)} = \prod_{i=1}^{m} B(z;\alpha_i).$$

Finally, the proof for a general $\mathbf{f}(Z)$ is a straightforward consequence of the previous analysis. □

3.2 Zeros

In this subsection we study the asymptotic behavior of the zeros of orthogonal polynomials associated with the discrete Sobolev inner product (23). In contrast with the real line case [2, 4, 11, 27, 33], there is not a well-developed theory for zeros of discrete Sobolev OPUC.

The monic version of (27) is

$$\Psi_n(z) = \Phi_n(z) - \frac{\lambda \Phi_n^{(j)}(\alpha)}{1 + \lambda K_{n-1}^{(j,j)}(\alpha, \alpha)} K_{n-1}^{(0,j)}(z, \alpha). \tag{31}$$

Thus,

$$\begin{bmatrix} \Psi_0(z) \\ \Psi_1(z) \\ \vdots \\ \Psi_{n-1}(z) \end{bmatrix} = \mathbf{L}_n \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix},$$

where $\mathbf{L}_n$ is an $n \times n$ lower triangular matrix with 1 as entries in the main diagonal, and the remaining entries are given by (31), i.e.,

$$l_{m,k} = -\frac{1}{\|\Phi_k\|_\sigma^2} \frac{\lambda \Phi_m^{(j)}(\alpha) \overline{\Phi_k^{(j)}(\alpha)}}{(1 + \lambda K_{m-1}^{(j,j)}(\alpha, \alpha))}, \quad 1 \leqslant m \leqslant n, \quad 0 \leqslant k \leqslant m-1.$$

One of our aims is to find a relation between $\mathbf{H}_\Psi$, the Hessenberg matrix associated with the monic orthogonal polynomials $\{\Psi_n\}_{n \geqslant 0}$, and $\mathbf{H}_\sigma$. In particular, we get

$$z \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} = (\mathbf{H}_\sigma)_n \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} + \Phi_n(z) \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

and, on the other hand,

$$z \begin{bmatrix} \Psi_0(z) \\ \Psi_1(z) \\ \vdots \\ \Psi_{n-1}(z) \end{bmatrix} = (\mathbf{H}_\Psi)_n \begin{bmatrix} \Psi_0(z) \\ \Psi_1(z) \\ \vdots \\ \Psi_{n-1}(z) \end{bmatrix} + \Psi_n(z) \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}. \tag{32}$$

Substituting in (32), we obtain

$$z\mathbf{L}_n \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} = (\mathbf{H}_\Psi)_n \mathbf{L}_n \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} + \Phi_n(z) \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} + \mathbf{A}_n \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix},$$

where

$$\mathbf{A}_n = \begin{bmatrix} 0 & \dots\dots & 0 \\ \vdots & & \vdots \\ 0 & \dots\dots & 0 \\ l_{n,0} & \dots\dots & l_{n,n-1} \end{bmatrix}.$$

As a consequence,

$$z \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} = \left(\mathbf{L}_n^{-1}(\mathbf{H}_\Psi)_n \mathbf{L}_n + \mathbf{L}_n^{-1}\mathbf{A}_n\right) \begin{bmatrix} \Phi_0(z) \\ \Phi_1(z) \\ \vdots \\ \Phi_{n-1}(z) \end{bmatrix} + \Phi_n(z) \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

so

$$(\mathbf{H}_\sigma)_n = \mathbf{L}_n^{-1}(\mathbf{H}_\Psi)_n \mathbf{L}_n + \mathbf{L}_n^{-1}\mathbf{A}_n$$

and therefore, since

$$\mathbf{L}_n^{-1}\mathbf{A}_n = \mathbf{A}_n,$$

we have

Theorem 7 ([12]). *Let $(\mathbf{H}_\sigma)_n$ and $(\mathbf{H}_\Psi)_n$ be the $n \times n$ truncated GGT matrices associated with $\{\Phi_n\}_{n\geq 0}$ and $\{\Psi_n\}_{n\geq 0}$, respectively. Then,*

$$(\mathbf{H}_\Psi)_n = \mathbf{L}_n((\mathbf{H}_\sigma)_n - \mathbf{A}_n)\mathbf{L}_n^{-1}.$$

As a consequence, the zeros of Ψ_{n+1} are the eigenvalues of the matrix $(\mathbf{H}_\sigma)_n - \mathbf{A}_n$, a rank one perturbation of the matrix $(\mathbf{H}_\sigma)_n$.

In the previous theorem we have characterized the eigenvalues of the GGT matrix associated with the discrete Sobolev polynomials as the eigenvalues of a rank one perturbation of the GGT matrix associated with the measure.

Notice that $\mathbf{A}_n = (0,\dots,0,1)^T(l_{n,0}, l_{n,1},\dots,l_{n,n-1})$ and, since $l_{n,k} = 0$ for $k < j$, then

$$\mathbf{A}_n = \frac{\lambda \Phi_n^{(j)}(\alpha)}{1+\lambda K_{n-1}^{(j,j)}(\alpha,\alpha)} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \left[0,\dots,0, \overline{\frac{\Phi_j^{(j)}(\alpha)}{\|\Phi_j\|^2}},\dots, \overline{\frac{\Phi_{n-1}^{(j)}(\alpha)}{\|\Phi_{n-1}\|^2}} \right].$$

As an example, if $d\sigma(\theta) = \dfrac{d\theta}{2\pi}$ is the Lebesgue measure, it is not difficult to see that in such a case, if $\alpha = 0$, then $\mathbf{A}_n = 0$, $n \neq j$, and

$$\mathbf{A}_j = \frac{\lambda (j!)^2}{1+\lambda (j!)^2} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \big[0,\dots,0,1,0,\dots,0\big],$$

where in the position j you have 1. On the other hand, if $\alpha = 1$, then for $n \geqslant j$,

$$\mathbf{A}_n = \frac{\lambda \frac{(n)!}{(n-j)!}}{1+\lambda \displaystyle\sum_{k=j}^{n-1} \left(\frac{k!}{(k-j)!}\right)^2} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \left[0,\dots,0,j!,(j+1)!,\dots,\tfrac{(n-1)!}{(n-j-1)!}\right].$$

Denote by $\{\phi_n(\cdot; d\sigma_{j+1})\}_{n\geqslant 0}$ the corresponding sequence of OPUC with respect to

$$d\sigma_j(z) = |z-\alpha|^{2(j+1)} d\sigma(z), \qquad j \geq 0.$$

For any $j \geq 0$, the relation between $\phi_n(\cdot; d\sigma_{j+1})$ and $\phi_n(\cdot, d\sigma)$ is given by Marcellán and Moral [26]

$$(z-\alpha)^{j+1}\phi_{n-j-1}(z, d\sigma_{j+1}) = \frac{\eta_{n-j-1}}{\alpha_n}\left(\phi_n(z) - \sum_{k=0}^{j} \gamma_{n,k} K_{n-1}^{(0,k)}(z,\alpha)\right), \tag{33}$$

where η_n is the leading coefficient of $\phi_n(\cdot, d\sigma_{j+1})$, and $\gamma_{n,k}$ is the kth component of the vector

$$\left[\phi_n(\alpha)\ \phi_n'(\alpha)\ \dots\ \phi_n^{(j)}(\alpha)\right] \begin{bmatrix} K_{n-1}(\alpha,\alpha) & K_{n-1}^{(0,1)}(\alpha,\alpha) & \dots & K_{n-1}^{(0,j)}(\alpha,\alpha) \\ K_{n-1}^{(1,0)}(\alpha,\alpha) & K_{n-1}^{(1,1)}(\alpha,\alpha) & \dots & K_{n-1}^{(1,j)}(\alpha,\alpha) \\ \vdots & \vdots & \ddots & \vdots \\ K_{n-1}^{(j,0)}(\alpha,\alpha) & K_{n-1}^{(j,1)}(\alpha,\alpha) & \dots & K_{n-1}^{(j,j)}(\alpha,\alpha) \end{bmatrix}^{-1}.$$

If $\sigma \in \mathcal{N}$, then [26]

$$\lim_{n\to\infty} \frac{\phi_n(z; d\sigma_{j+1})}{\phi_{n+j+1}(z)} = \left(\frac{\overline{\alpha}}{|\alpha|} \frac{1}{\overline{\alpha} z - 1} \right)^{j+1}, \tag{34}$$

holds uniformly in $|z| > 1$ if $|\alpha| \geqslant 1$, and in $|z| \geqslant 1$ if $|\alpha| > 1$.

On the other hand, by Theorem 5, for $|\alpha| > 1$,

$$\lim_{n\to\infty} \frac{\psi_n(z)}{\phi_n(z)} = \frac{\overline{\alpha}}{|\alpha|} \frac{z - \alpha}{\overline{\alpha} z - 1}, \tag{35}$$

uniformly on every compact subset of $|z| > 1$. From (34) and (35), we have

$$\left(\frac{|\alpha|}{\overline{\alpha}} (\overline{\alpha} z - 1) \right)^j (z - \alpha) \lim_{n\to\infty} \frac{\phi_n(z; d\sigma_{j+1})}{\phi_{n+j+1}(z)} = \lim_{n\to\infty} \frac{\psi_{n+j+1}(z)}{\phi_{n+j+1}(z)}.$$

Hence,

$$\lim_{n\to\infty} \frac{\psi_n(z)}{\phi_{n-j-1}(z; d\sigma_{j+1})} = \left(\frac{|\alpha|}{\overline{\alpha}} (\overline{\alpha} z - 1) \right)^j (z - \alpha),$$

uniformly $|z| \geqslant 1$. The following result is a straightforward consequence of the Hurwitz's Theorem [14].

Theorem 8 ([12]). *There is a positive integer n_0 such that, for $n \geqslant n_0$, the nth Sobolev monic OPUC Ψ_n defined by (23), with $|\alpha| > 1$, has exactly one zero in $\mathbb{C} \setminus \overline{\mathbb{D}}$ accumulating in α, while the remaining zeros belong to $\mathbb{D}$.*

This result is analogous to the well-known result of Meijer [32] for Sobolev OPRL, see also [11]. We now turn our attention to the case when λ tends to infinity. For a fixed n, $j = 0$ and λ tends to infinity, $n - 1$ zeros of ψ_n tend to the zeros of $\phi_{n-1}(z, d\sigma_1)$, and the remaining zero tends to $z = \alpha$. On the other hand, for $j = 1$, the zeros of ψ_n tend to the zeros of a linear combination of $\Phi_n(z)$, $(z - \alpha)\Phi_{n-1}(z, d\sigma_1)$, and $(z - \alpha)^2 \Phi_{n-2}(z, d\sigma_2)$ when $\lambda \to \infty$. This result can be generalized for arbitrary j. Indeed, from (33), notice that

$$-\gamma_{n,j} K_{n-1}^{(0,j)}(z, \alpha) = (z - \alpha)^{j+1} \phi_{n-j-1}(z, d\sigma_{j+1}) - \frac{\eta_{n-j-1}}{\alpha_n} \phi_n(z) + \sum_{k=0}^{j-1} \gamma_{n,k} K_{n-1}^{(0,k)}(z, \alpha).$$

Applying the last formula recursively for $k = 0, 1, \ldots, j - 1$, we obtain

Theorem 9 ([12]). *Let $\{\psi_n\}_{n\geqslant 0}$ be the sequence of orthonormal polynomials with respect to (23), with $j \geq 0$. Then $\psi_n(z)$ is a linear combination of $\phi_n(z)$, $(z - \alpha)\phi_{n-1}(z, d\sigma_1)$, $\ldots$, $(z - \alpha)^{j+1} \phi_{n-j-1}(z, d\sigma_{j+1})$. As a consequence, the zeros of $\psi_n(z)$ tend to the zeros of such a linear combination when $\lambda \to \infty$.*

Acknowledgements The authors wish to express their thanks to Nicholas J. Daras and Michael Th. Rassias for the invitation to participate in this volume. The research of the first author is supported by the Portuguese Government through the Fundação para a Ciência e a Tecnologia (FCT) under the grant SFRH/BPD/ 101139/2014. This author also acknowledges the financial support by the Brazilian Government through the CNPq under the project 470019/2013-1. The research of the first and second author is supported by the Dirección General de Investigación Científica y Técnica, Ministerio de Economía y Competitividad of Spain under the project MTM2012–36732–C03–01.

References

1. Akhiezer, N.I.: The Classical Moment Problem and Some Related Questions in Analysis. Hafner, New York (1965) [Russian Original (1961)]
2. Alfaro, M., López, G., Rezola, M.L.: Some properties of zeros of Sobolev-type orthogonal polynomials. J. Comput. Appl. Math. **69**, 171–179 (1996)
3. Branquinho, A., Foulquié, A., Marcellán, F.: Asymptotic behavior of Sobolev-type orthogonal polynomials on a rectifiable Jordan curve or arc. Constr. Approx. **18**, 161–182 (2002)
4. Bruin, M.G.D.: A tool for locating zeros of orthogonal polynomials in Sobolev inner product spaces. J. Comput. Appl. Math. **49**, 27–35 (1993)
5. Castillo, K.: Monotonicity of zeros for a class of polynomials including hypergeometric polynomials, Appl. Math. Comput. (2015). (in press).
6. Castillo, K.: A new approach to relative asymptotic behavior for discrete Sobolev-type orthogonal polynomials on the unit circle. Appl. Math. Lett. **25**, 1000–1004 (2012)
7. Castillo, K.: On perturbed Szegő recurrence. J. Math. Anal. Appl. **411**, 742–752 (2013)
8. Castillo, K., Cruz-Barroso, R., Perdomo-Pío, F.: On a spectral theorem in the para-orthogonality theory. (Submitted)
9. Castillo, K., Garza, L., Marcellán, F.: A new linear spectral transformation associated with derivatives of Dirac linear functionals. J. Approx. Theory **163**, 1834–1853 (2011)
10. Castillo, K., Garza, L., Marcellán, F.: Perturbations on the subdiagonals of Toeplitz matrices. Linear Algebra Appl. **434**, 1563–1579 (2011)
11. Castillo, K., Mello, M.V., Rafaeli, F.R.: Monotonicity and asymptotics of zeros of Sobolev type orthogonal polynomials: a general case. Appl. Numer. Math. **62**, 1663–1671 (2012)
12. Castillo, K., Garza, L., Marcellán, F.: Zeros of discrete Sobolev orthogonal polynomials on the unit circle. Numer. Algoritm. **60**, 669–681 (2012)
13. Castillo, K., Garza, L.G., Marcellán, F.: On computational aspects of discrete Sobolev inner products. Appl. Math. Comput. **223**, 452–460 (2013)
14. Conway, J.B.: Functions of One Complex Variable I. Springer, New York (1978)
15. Daruis, L., Hernández, J., Marcellán, F.: Spectral transformations for Hermitian Toeplitz matrices. J. Comput. Appl. Math. **202**, 155–176 (2007)
16. Foulquié, A., Marcellán, F., Pan, K.: Asymptotic behavior of Sobolev-type orthogonal polynomials on the unit circle. J. Approx. Theory **100**, 345–363 (1999)
17. Freud, G.: Orthogonal Polynomials. Akadémiai Kiadó, Pergamon Press, Budapest (1971)
18. Geronimus, Y.L.: On polynomials orthogonal on the unit circle, on trigonometric moment problem, and on allied Carathéodory and Schur functions. Rec. Math. [Mat. Sbornik] N.S. **15**(57), 99–130 (1944)
19. Gonchar, A.A.: On the convergence of Padé approximants for some classes of meromorphic functions. Math. USSR Sbornik **26**, 555–575 (1975)
20. Gragg, W.B.: Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle. J. Comput. Appl. Math. **46**, 183–198 (1993) [Russian Original in Numerical Methods of Linear Algebra, pp. 16–32, Moskov Gos. University, Moscow (1982)]

21. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press, Cambridge (1990)
22. Ismail, M.E.H.: Classical and Quantum Orthogonal Polynomials in One Variable. Cambridge University Press, Cambridge (2009) [Encyclopedia in Mathematics and Its Applications, Vol. 98]
23. Levin, E., Lubinsky, D.S.: Universality limits involving orthogonal polynomials on the unit circle. Comput. Methods Funct. Theory **7**, 543–561 (2007)
24. Li, X., Marcellán, F.: On polynomials orthogonal with respect to Sobolev inner products. Pacific J. Math. **175**, 127–146 (1996)
25. López, G., Marcellán, F., Van Assche, W.: Relative asymptotics for orthogonal polynomials with respect to a discrete Sobolev inner product. Constr. Approx. **11**, 107–137 (1995)
26. Marcellán, F., Moral, L.: Sobolev-type orthogonal polynomials on the unit circle. Appl. Math. Comput. **128**, 107–137 (2002)
27. Marcellán, F., Rafaeli, F.R.: Monotonicity and asymptotic of zeros of Laguerre-Sobolev-type orthogonal polynomials of higher derivatives. Proc. Am. Math. Soc. **139**, 3929–3936 (2011)
28. Marcellán, F., Ronveaux, A.: Orthogonal polynomials and Sobolev inner products. A bibliography. Universidad Carlos III de Madrid, Madrid (2014)
29. Marcellán, F., Van Assche, W.: Relative asymptotics for orthogonal polynomials with a Sobolev inner product. J. Approx. Theory **72**, 193–209 (1993)
30. Martínez-Finkelshtein, A., Simon, B.: Asymptotics of the L^2 norm of derivatives of OPUC. J. Approx. Theory **163**, 747–773 (2011)
31. Maté, A., Nevai, P., Totik, V.: Extensions of Szegő's theory of orthogonal polynomials II. Constr. Approx. **3**, 51–72 (1987)
32. Meijer, H.G.: Zero distribution of orthogonal polynomials in a certain discrete sobolev space. J. Math. Anal. Appl. **172**, 520–532 (1993)
33. Pérez, T.E., Piñar, M.: Global properties of zeros for Sobolev-type orthogonal polynomials. J. Comput. Appl. Math. **49**, 225–232 (1993)
34. Rakhmanov, E.A.: On the asymptotics of the ratio of orthogonal polynomials I, II. Sb. Math. **32**, 199–214 (1977), **46**, 105–118 (1983) [Russian original in Mat. Sb. **103**(145), 237–252 (1977); Mat. Sb. **118**(160), 104–117 (1982)]
35. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill, New York (1987)
36. Schur, I.: Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind, I. J. Reine Angew. Math. **147**, 205–232 (1917) [Gohberg, I. (ed.) English translation in I. Schur Methods in Operator Theory and Signal Processing, pp. 31–59. Operator Theory: Advances and Applications 18, Birkhäuser, Basel (1986)]
37. Schur, I.: Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind, II. J. Reine Angew. Math. **148**, 122–145 (1918) [Gohberg, I. (ed.) English translation in I. Schur Methods in Operator Theory and Signal Processing, pp. 66–88. Operator Theory: Advances and Applications 18, Birkhäuser, Basel (1986)]
38. Simon, B.: Orthogonal Polynomials on the Unit Circle. American Mathematical Society, Providence, RI (2005) [American Mathematical Society, Colloquium Publication Series, vol. 54, Part 2]
39. Simon, B.: Szegő's Theorem and Its Descendants: Spectral Theory for L^2 Perturbations of Orthogonal Polynomials. Princeton University Press, Princeton (2011)
40. Stahl, H., Totik, V.: General Orthogonal Polynomials. Encyclopedia of Mathematics and Its Applications, vol. 43. Cambridge University Press, Cambridge (1992)
41. Szegő, G.: Beiträge zur Theorie der Toeplitzschen formen. Math. Z. **6**, 167–202 (1920)
42. Szegő, G.: Beiträge zur Theorie der Toeplitzschen Formen, II. Math. Z. **9**, 167–190 (1921)
43. Szegő, G.: Orthogonal Polynomials, 4th edn. American Mathematical Society, Providence, RI (1975) [American Mathematical Society, Colloquium Publication Series, vol. 24]
44. Tasis, C.: Propiedades de polinomios ortogonales relativos a la circunferencia unidad. Ph.D. Thesis, Departamento de Matemáticas, Universidad de Cantabria (1989, in Spanish)
45. Teplyaev, A.V.: The pure point spectrum of random orthogonal polynomials on the unit circle. Sov. Math. Dokl. **44**, 407–411 (1992) [Russian Original in Dokl. Akad. Nauk SSSR **320**, 49–53 (1991)]

46. Verblunsky, S.: On positive harmonic functions: a contribution to the algebra of Fourier series. Proc. London Math. Soc. **38**(2), 125–157 (1935)
47. Wong, M.L.: Generalized bounded variation and inserting point masses. Constr. Approx. **30**, 1–15 (2009)

On the Unstable Equilibrium Points and System Separations in Electric Power Systems: A Numerical Study

Jinda Cui, Hsiao-Dong Chiang, and Tao Wang

Abstract An equilibrium problem of electric power system is closely associated with the system separations. An effective three-step scheme is developed to compute the system separations subject to different contingencies. For illustrative purposes, the proposed scheme is applied to small-sized power system testing models with promising results. Simulations are performed on the models of electric power systems, which demonstrate the effectiveness of the proposed scheme. This scheme has the potential of being applied to the contingency analysis of large-scale systems.

Keywords: Electric power system • Stability analysis • Equilibrium problem System separation • Numerical study

1 Introduction

Mathematical theory of variational inequalities was initially introduced for studying the partial differential equations [1, 2] with the applications principally drawn from mechanics. It has been extended to treat the equilibrium problems, which are fundamental in various disciplines [3–9], ranging from economics, operations research, to civil and electrical engineering. Different methodologies and approaches [10–15], including systems of algebraic equations, linear and nonlinear optimization, complementary theory and fixed point theory, have been developed to formulate the equilibrium problems and compute the solutions. This study focuses on an equilibrium problem and its application to electric power systems.

Electric power systems are recognized as a class of the largest and physically most complex nonlinear systems in the world. An electric power system is an interconnected system consisting of generating stations that convert fuel energy into

J. Cui
Bigwood Systems Inc., Ithaca, NY 14850, USA
e-mail: jinda@bigwood-systems.com

H.-D. Chiang • T. Wang (✉)
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA
e-mail: chiang@ece.cornell.edu; tw355@cornell.edu

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_5

electricity, primary and secondary distribution substations that distribute power to loads and consumers, and transmission lines that tie the generating stations and distribution substation together. By nature, an electric power system continually experiences disturbances/contingencies, which can be classified into two main categories: event disturbances and load variations. Event disturbances and contingencies refer to the loss of generating units or transmission components (e.g., lines, transformers, substations) due to short-circuits. Such disturbances can occur as a single-equipment outage or as multiple simultaneous outages when the relay actions are taken into account. On the other hand, load variations correspond to the fluctuations of load demands at buses and/or power transfers among buses. Usually the network configuration may remain unchanged after load variations.

To protect power systems from damage due to disturbances, protective relays are placed strategically throughout a power system to detect faults/disturbances and to trigger the opening of circuit breakers necessary to isolate faults [16]. These relays are designed to detect defective lines and apparatus or other power system conditions of an abnormal or dangerous nature and to initiate appropriate control actions. Due to the action of these protective relays, a power system subject to an event disturbance can be viewed as going through network configuration changes in three stages: the pre-fault, the fault-on, and the post-fault systems. More precisely, the *pre-fault* system refers to the undisturbed system; once the system undergoes a fault (an event disturbance), it then moves into the *fault-on* system before the fault is cleared by protective system operations; suppose the fault is cleared at certain time/moment and no additional protective actions occur afterwards, the system then is called the *post-fault* system. Generally speaking, during each stage the system is governed by a dynamical system (i.e., a set of differential and algebraic equations), and the system trajectory is named after the stage. For instance, the fault-on trajectory refers to the system trajectory during the fault-on (system) stage, while the post-fault trajectory is that during the post-fault stage.

In the past decades, researchers and engineers in the power engineering community have growing interests in the theory of nonlinear dynamical systems, to analyze nonlinear problems arising in electric power systems analysis and to develop counter-measures/control schemes for power system instability prevention. Indeed, the power systems nowadays have been pushed closer to the operation limits, due to the increase in load demands and the pressure of economical operations. This trend results in system-wide disturbances, or even worse cascading outages and system blackouts. Furthermore, high penetration of renewable energy can aggravate power systems stability. To maintain the overall stability, a controlled-islanding technique can be adopted to prevent system-wide outage when large disturbance occurs. That is, the entire network is split into a collection of smaller isolated power systems (called power islands) by disconnecting certain transmission lines [17–20]. In general, there can be multiple admissible schemes for splitting the network. It is worthwhile noting that the system separation of power system (i.e., a pattern that the generators lose synchronism due to the faults cleared immediately after the critical clearing time (CCT) [21]) is closely related to the controlling unstable equilibrium point (controlling u.e.p) on the stability boundary [16, 22–25], which can provide an index for evaluating these splitting schemes. Here the controlling

u.e.p associated with a fault/contingency refers to the u.e.p whose stable manifold contains the exit point (i.e., the intersection of the fault-on trajectory with the stability boundary) [16].

The emphasis of the paper lies on the computer simulation of system separations in an electric power system and the development of numerical scheme for computing the number of system separations. For clarity, we review the mathematical preliminaries in the theory of nonlinear dynamical systems, transient stability model for electric power systems, and theoretical results on unstable equilibrium points in Sect. 2. After that, numerical algorithms and simulation results are described in Sect. 3. More precisely, we bring forward a three-step scheme that: line-fitting for the asymptote of the post-fault trajectory at Step 1; generating coarse clusters by the slopes of the fitting lines (for the asymptotes) at Step 2; and further computing the final grouping result at Step 3 based on the coarse clusters obtained in the preceding step. The proposed scheme is tested on different power systems at last, and the simulation results for each step are illustrated and tabulated.

2 Mathematical Preliminary

2.1 Nonlinear Dynamical System and Equilibrium Point

Consider an autonomous nonlinear dynamical system of the general form

$$\dot{x} = f(x), \quad x \in \mathbb{R}^n, \tag{1}$$

where the state vector $x = (x_1, \ldots, x_n)$, and $f(x) = (f_1(x), \ldots, f_n(x))^T$. Naturally the vector field $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is assumed differentiable, and satisfies a sufficient condition for the existence and uniqueness of the solution. The solution of (1) starting from x_0 at time $t = 0$ will be denoted by $\phi(t, x_0)$, or by $x(t)$ when it is clear from the context.

By convention, a point $\hat{x}$ is an asymptotically stable equilibrium point [26] of the dynamical system (1), if it is Lyapunov stable and there exists $\delta > 0$ such that $\phi(t, x_0) \to \hat{x}$ as $t \to \infty$, for all points x_0 satisfying $\|x_0 - \hat{x}\| < \delta$. In other words, there exists a neighborhood of the equilibrium point such that every solution starting in this neighborhood remains in a neighborhood and is attracted to the equilibrium point as time tends to infinity. If δ can be chosen arbitrarily large, then every trajectory is attracted to $\hat{x}$, and thereby $\hat{x}$ is called a *global asymptotically stable equilibrium point*. There are many physical systems containing asymptotically stable equilibrium points, but not globally stable equilibrium points. Alternatively, a useful concept is the *stability region* (also called the *region of attraction or domain of attraction*). The stability region of an asymptotically stable equilibrium point x_s is the set of all points x such that $\phi(t, x) \to x_s$, as $t \to \infty$. We will denote the stability region of x_s by $A(x_s)$, and its closure by $\overline{A(x_s)}$, where

$$A(x_s) \doteq \{x \in \mathbb{R}^n;\ \lim_{t\to\infty} \phi(t, x) = x_s\}. \tag{2}$$

When it is clear from the context, we write A for $A(x_s)$. From a topological point of view, the stability region $A(x_s)$ is an open, invariant, and connected set. The boundary of stability region $A(x_s)$ is called the *stability boundary* (also called *separatrix*) of x_s and will be denoted by $\partial A(x_s)$.

In general, a point $x^* \in \mathbb{R}^n$ is said to be an *equilibrium point* of (1), if $f(x^*) = \mathbf{0}$. Besides, we say that an equilibrium point $x^* \in \mathbb{R}^n$ is *hyperbolic*, if the Jacobian matrix of $f(\cdot)$ at x^* has no eigenvalues with a zero real part. In addition, a *type-k* equilibrium point refers to a hyperbolic equilibrium point at which the Jacobian has exactly k eigenvalues with positive real part. In particular, the type-0 equilibrium points are *asymptotically stable equilibrium points*, while the type-n equilibrium points are called the *source points*. For a type-k equilibrium point x^*, its stable manifold $W^s(x^*)$ and unstable manifold $W^u(x^*)$ are defined, respectively, as

$$W^s(x^*) \doteq \{x \in \mathbb{R}^n : \lim_{t\to\infty} \phi(t,x) = x^*\},$$
$$W^u(x^*) \doteq \{x \in \mathbb{R}^n : \lim_{t\to-\infty} \phi(t,x) = x^*\},$$

where the dimension of $W^u(x^*)$ and $W^s(x^*)$ are k and $(n-k)$, respectively.

However, the structure of the boundary of the stability region $A(x_s)$ can be very complex for a general nonlinear system (1). An alternative study is the quasi-stability boundary. The *quasi-stability boundary* $\partial A_p(x_s)$ of a stable equilibrium point x_s is defined by $\partial\overline{A(x_s)}$ [23], where int$(\cdot)$ refers to the interior of a set and ∂ is the boundary. Clearly, the quasi-stability boundary $\partial A_p(x_s) \subseteq \partial A(x_s)$.

Two stability regions are called *neighboring* to each other, if their closures have nonempty intersection. Given two different stable equilibrium points x_s and x_s', it should be apparent that the stability regions $(A(x_s) \cap A(x_s')) = \emptyset$, and $(\overline{A(x_s)} \cap \overline{A(x_s')}) = (\partial A(x_s) \cap \partial A(x_s'))$. Indeed, $(\partial A_p(x_s) \cap \partial A_p(x_s')) = \overline{A(x_s)} \cap \overline{A(x_s')} = \partial A(x_s) \cap \partial A(x_s')$. This implies that two stability regions are neighboring to each other, if and only if their quasi-stability boundaries have nonempty intersection. In the sequel, a stability boundary thus usually refers to ∂A_p, without causing any confusion.

Recall a set $K \subseteq \mathbb{R}^n$ is *invariant* regarding the dynamics at (1), if every trajectory of (1) starting in K stays in K for all $t \in \mathbb{R}$. Clearly, the stability region is an invariant set. Given two sub-manifolds M_1 and M_2 of a manifold M, we say that they meet the *transversality condition*, if either (1) $(M_1 \cap M_2) = \emptyset$, or (2) at every point $y \in (M_1 \cap M_2)$, the tangent spaces of M_1 and M_2 span the tangent spaces of M at y.

In the subsequent study the conditions below are assumed to be satisfied by (1).

(A1) *All the equilibrium points are hyperbolic, and on a stability boundary they are finite in number.*

(A2) *The stable and unstable manifolds of equilibrium points on the stability boundary satisfy the transversality condition.*

(A3) *Every trajectory approaches an equilibrium point as* $t \to +\infty$.

Here (A1) and (A2) are generic properties for nonlinear dynamical systems while (A3) is not generic, however it is satisfied by a large class of nonlinear dynamical

systems, as the power system dynamics model for transient stability analysis. On the stability boundary, the results below describe the structural characterization.

Theorem 1 (Complete Characterization of Quasi-Stability Boundary [23]). *Consider a stable equilibrium point x_s of the nonlinear dynamical system (1) satisfying the assumptions (A1)–(A3). Let $x_e^i, i \in \mathbb{N}$ be the equilibrium points on the quasi-stability boundary $\partial A_p(x_s)$. Then, the quasi-stability boundary*

$$\partial A_p(x_s) = \bigcup_{x_e^i \in \partial A_p(x_s)} W^s(x_e^i).$$

Theorem 1 asserts that the union of the stable manifolds of the UEPs (unstable equilibrium points) lying on the stability boundary equals the stability boundary. This theorem however provides little information regarding the number and the type of UEPs that can lie on the stability boundary.

Theorem 2 (Characterization of Quasi-Stability Boundary [23]). *Consider a stable equilibrium point x_s of the nonlinear dynamical system (1) satisfying the assumptions (A1)–(A3). Let $\sigma_i, i \in \mathbb{N}$ be the type-one equilibrium points on the quasi-stability boundary $\partial A_p(x_s)$. Then, the quasi-stability boundary*

$$\partial A_p(x_s) = \bigcup_{\sigma_i \in \partial A_p(x_s)} \overline{W^s(\sigma_i)}.$$

Theorem 2 asserts that the union of the closure of stable manifolds of type-one UEPs lying on the stability boundary equals the stability boundary.

2.2 Power System and Transient Stability Model

We consider the classical model for transient stability analysis, for a power system consisting of n generators. The dynamics of the ith generator is described by

$$\begin{cases} \dot{\delta}_i = \omega_i \\ M_i\dot{\omega}_i = -D_i\omega_i + P_{mi} - P_{ei}(\delta) \end{cases} \tag{3}$$

for $1 \le i \le n$, where D_i and M_i are damping ratio and inertia constant of machine i. Here the loads have been modeled as constant impedances, and

$$P_{ei}(\delta) \doteq \sum_{j=1}^{n} E_i E_j \cdot \left\{ B_{ij} \sin(\delta_i - \delta_j) + G_{ij} \cos(\delta_i - \delta_j) \right\} \tag{4}$$

refers to the electrical power at machine i, while E_i is the constant voltage behind direct axis transient reactance, P_{mi} is the machine power. More detailed and sophisticated models for power system transient stability analysis can be found, for example, in the book [16].

The equations in (3) can be written in a number of reference frames, as in terms of absolute angles, relative angles between machines, or relative to the center of inertia. Here we use the relative angles between machines by fixing δ_i (say $i = 1$) as the reference angle, and define the set

$$\mathscr{H} \doteq \{(\delta, \omega);\ \delta \in \mathbb{R}^n, \delta_1 \equiv 0,\ \omega = \mathbf{0}\},$$

which is an $(n-1)$-dimensional hyperplane of $\mathbb{R}^{2n}$. Clearly, if $x_* = (\delta_*, \omega_*)$ is an equilibrium point of the system (3), one must have the point $x_* \in \mathscr{H}$.

Indeed, restricted to the hyperplane $\mathscr{H}$, the system (3) is spatially periodic, in the following sense. Let $x \doteq (\delta, \omega) = (\delta_1, \ldots, \delta_n, \omega_1, \ldots, \omega_n) \in \mathbb{R}^{2n}$, with

$$F_i(x) \doteq \omega_i; \qquad F_{n+i}(x) \doteq (-D_i\omega_i + P_{mi} - P_{ei}(\delta))/M_i$$

for $1 \le i \le n$. The system (3) can be reformulated as

$$\dot{x} = F(x) = (F_1(x), \ldots, F_{2n}(x))^T,\ x \in \mathbb{R}^{2n}.$$

The system is spatially periodic restricted to $\mathscr{H}$, if there exist $(n-1)$ constants $p_k > 0$ for $2 \le k \le n$, such that $F_j(x) = F_j(x + p_k e_k)$ for all $x \in \mathbb{R}^{2n}$ and $1 \le j \le 2n$. Here e_k denotes the vector in $\mathbb{R}^{2n}$ with 1 in the kth coordinate and 0's elsewhere. In addition, an $(n-1)$-tuple $p_* = (p_2^*, \ldots, p_n^*)$ is called the spatial periods, if each $p_k^* > 0$ is the minimum positive number p_k s.t. $F_j(x) = F_j(x + p_k e_k)$ for all x, j.

On the system (3), the following conditions are made.

(A4) *The spatial-period $p_k^* = 2\pi$, for all $2 \le k \le n$. Moreover, there is at most one stable equilibrium point in each region of the form $\{(\delta, \omega);\ z_i \le \delta_i < z_i + 2\pi, 2 \le i \le n\} \subseteq \mathscr{H}$, for all $z_i \in \mathbb{R}$.*

It is worthwhile noting that, by applying a linear transformation on δ, any system (3) always can be transformed to a spatially periodic system (restricted to $\mathscr{H}$) with $p_k^* = 2\pi$ for all $2 \le k \le n$. The remainder of the assumption ensures that, if x_s is a stable equilibrium point of (3), then another stable equilibrium point $\tilde{x}_s$ can always be represented by

$$\tilde{x}_s = x_s + (0, \alpha_2 p_2^*, \ldots, \alpha_n p_n^*, 0, \ldots, 0) = x_s + 2\pi\,(0, \alpha_2, \ldots, \alpha_n, 0, \ldots, 0) \tag{5}$$

for some integer α_k, $2 \le k \le n$. Let $\mathscr{P} \doteq \{2\pi\,(0, \alpha_2, \ldots, \alpha_n, 0, \ldots, 0) \in \mathbb{R}^{2n};\ \alpha_k \in \mathbb{Z},\ 2 \le k \le n\}$, where $\mathbb{Z}$ is the set of all integers. Then, two stable equilibrium points of (3) always differ by a vector in $\mathscr{P}$.

(A5) *The intersection of the hyperplane $\mathscr{H}$ with every stability region is bounded. Moreover, the closures of two neighboring stability regions share a boundary of dimension $(2n-1)$.*

In general, a stability region of (3) is unbounded [16, 24, 25], but the intersection with the hyperplane $\mathscr{H}$ is bounded as assumed in (A5). The second part assumes that two neighboring stability regions are contiguous, instead of sharing a corner.

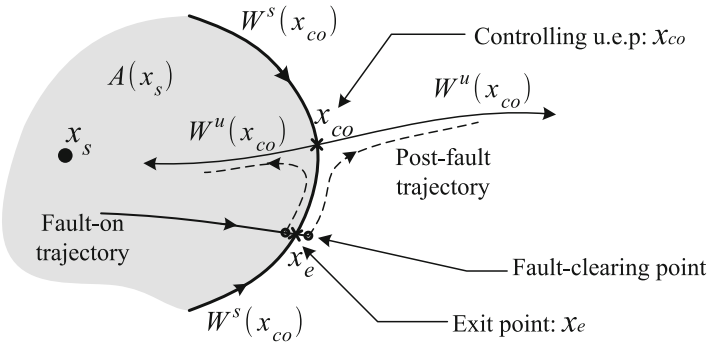


Fig. 1 An illustration of the inclination lemma. The point x_s indicates the stable equilibrium point/state, and x_{co} refers to the controlling unstable equilibrium point (controlling u.e.p) on the stability boundary, where the fault-clearing point (f.c.p) is close to the exit point x_e. If the fault-clearing point lies inside the stability region, then the post-fault trajectory moves along the stable manifold of the controlling u.e.p and then along the portion of unstable manifold of the controlling u.e.p inside the stability region, and converges to the post-fault stable equilibrium point x_s eventually. On the other hand, if the f.c.p lies outside of the stability region, then the post-fault trajectory moves along the stable manifold of the controlling u.e.p and then approach the portion of unstable manifold (of the controlling u.e.p) outside the stability region

The aforesaid equilibrium problem refers to the problem of identifying the number of controlling unstable equilibrium points on a stability boundary. This is pertinent to the estimate of system separations [22]. Indeed, the system separation of an electric power system indicates a pattern that the system deviates from the initial state. As suggested by the inclination lemma [27], several facts are summarized [22] for the connection between the critical unstable trajectory (i.e., the post-fault trajectory with the fault being cleared right after CCT) and the controlling unstable equilibrium point on the stability boundary (see Fig. 1), where the stable manifold of the controlling unstable equilibrium point contains the point of intersection (i.e., exit point) of the fault-on trajectory with the stability boundary.

Fact 1: If the portion of unstable manifold of the controlling unstable equilibrium point which lies outside of the stability region converges to a stable equilibrium point, then the critical unstable trajectory converges to the same stable equilibrium point.

Fact 2: If the portion of unstable manifold of the controlling unstable equilibrium point which lies outside of the stability region becomes unbounded, then the critical unstable trajectory also becomes unbounded.

Fact 3: All the critical unstable trajectories (of the same post-fault system) with the same controlling unstable equilibrium point have the same asymptotical behavior.

Fact 4: All the critical unstable trajectories (of the same post-fault system) with the same controlling unstable equilibrium point which is of type-one have the same system separation.

In view of these facts, the number of controlling unstable equilibrium points provides an upper bound for the number of different system separations. On the system model (3) it is evident that if (δ^*, ω^*) is an equilibrium point of (3), one must have $\omega^* = 0$ and $P_{mi} - P_{ei}(\delta^*) = 0$ for all i. In view of the expression (4) for P_{ei}, a system of $2n$ polynomial equations can be obtained for the system of equations $\{P_{mi} - P_{ei}(\delta) = 0;\ i = 1, \ldots, n\}$ by trigonometric substitutions, say

$$\begin{cases} P_{m1} = \sum_{j=1}^{n} E_1 E_j \{-B_{1j} x_j + G_{1j} y_j\}, & \text{(6a)} \\ P_{mi} = \sum_{j=1}^{n} E_i E_j \{B_{ij}(x_i y_j - y_i x_j) + G_{ij}(y_i y_j + x_i x_j)\},\ 2 \le i \le n-1, & \text{(6b)} \\ 1 = x_i^2 + y_i^2,\ 2 \le i \le n, & \text{(6c)} \end{cases}$$

where the substitutions $x_i \doteq \sin(\delta_i - \delta_1)$ and $y_i \doteq \cos(\delta_i - \delta_1)$ are used, especially $x_1 = 0$ and $y_1 = 1$. By Theorem 4.1 [28] and Theorem 3.1 [29], there are exactly $\binom{2n-2}{n-1}$ complex solutions to the system of equations consisting of (6a)–(6c). Together with the multiplicity of stability regions sharing an equilibrium point, we have the following estimate on the controlling unstable equilibrium points.

Theorem 3 ([30]). *Consider a system (3) satisfying assumptions (A1)–(A5), then there are totally no more than* $2 \cdot \binom{2n-2}{n-1}$ *controlling unstable equilibrium points on the stability boundary of a stable equilibrium point.*

As mentioned earlier, the number of system separations is bounded from above by the number of controlling or type-one unstable equilibrium points on the stability boundary. Thus, Theorem 3 shows that there are at most $2 \cdot \binom{2n-2}{n-1}$ system separations for the power system (3) of n generators. For instance, for $n = 3$ there are at most $12 = 2 \cdot \binom{4}{2}$ system separations, while for $n = 4$ the number of system separations cannot exceed $40 = 2 \cdot \binom{6}{3}$.

It is worthwhile noting that the frequency/probability of occurrence varies with system separation. The theoretical results, as Theorem 3, provide bounds on the number of possible system separations, but give little information about the frequency. This part will be also addressed by the numerical study in the next section.

3 Numerical Scheme and Simulation Results

We propose a numerical scheme for computing the number of possible system separations due to different contingencies. Two contingencies are said to yield a same system separation, if their post-fault trajectories converge asymptotically, where the fault-on trajectory refers to the trajectory $\phi(t, x_0)$ following the dynamics of the power system when a fault occurs, while the post-fault trajectory is that for the system with the fault being cleared immediately after the CCT. Since there can be numerous contingencies yielding a same system separation. The number of system separations can be far less than that of contingencies. In this regard, a set

of 'N-1' and 'N-2' contingencies are simulated, and a numerical scheme is devised to compute the number of system separations. As described below, the numerical scheme consists of three steps: line-fitting, pre-grouping, and sub-grouping.

Step 1: (line-fitting) Find the best-fitting line to the post-fault trajectory (due to a contingency), which gives an approximation for the asymptote of trajectory.
Step 2: (pre-grouping) Group the contingencies by the slope of fitting line for the post-fault trajectories, and generate the coarse clusters of contingencies.
Step 3: (sub-grouping) Generate the (final) groups by measuring Hausdorff distance between any couple of post-fault trajectories for each cluster of contingencies generated at Step 2. The number of groups indicates the number of system separations, and each group contains the contingencies yielding a same system separation.

The rationale for Step 2 is that asymptotically convergent post-fault trajectories must have near same slopes for their asymptotes. Concerning the above three-step scheme, some questions naturally arise.

- What is the importance of Steps 1 and 2?
- Or to ask, why not compute directly the pairwise Hausdorff distance for all post-fault trajectories as Step 3 and use it to generate the groups?

The questions are answered using the following observation. First of all, the number of contingencies can grow significantly, as the system size increases. Moreover, for a time-domain simulation up to 40 s, the post-fault trajectory can consist of more than 5000 points, and it consumes roughly four and half hours (using a PC with i7 2.4 GHz Quad-Core CPU and 16G RAM), to compute the pairwise Hausdorff distance for a collection of 93 post-fault trajectories that are generated by simulating the contingencies for an electric power system having 4 generators and 57 buses. On the other hand, by taking the proposed three-step scheme, the computing time can be drastically reduced, approximately by 88.9 %. That is, the proposed scheme only takes about half an hour to produce the same result, which motivates the design of above scheme adopting an overarching mechanism: screening at Step 1, ranking/preprocessing at Step 2 and detailed analysis at Step 3.

Indeed, towards testing the proposed three-step scheme, one key prerequisite is the time-domain simulation under contingencies and the computation of the CCT, which can be performed according to the procedure illustrated in Fig. 2. Specifically, CCT is computed below by Steps I–IV.

Step I: When a contingency in the contingency list occurs, the program first invokes the power flow solver to obtain the initial state. Initially, a small value near zero is assigned to the *fault clearing time* (FCT), to examine the validity of the scenario. By the time-domain simulation that follows (using the incumbent FCT), if the system appears unstable, the program outputs a warning message and continues to simulate the subsequent contingency. Otherwise, if the system turns out to be stable, we move on to the next step.

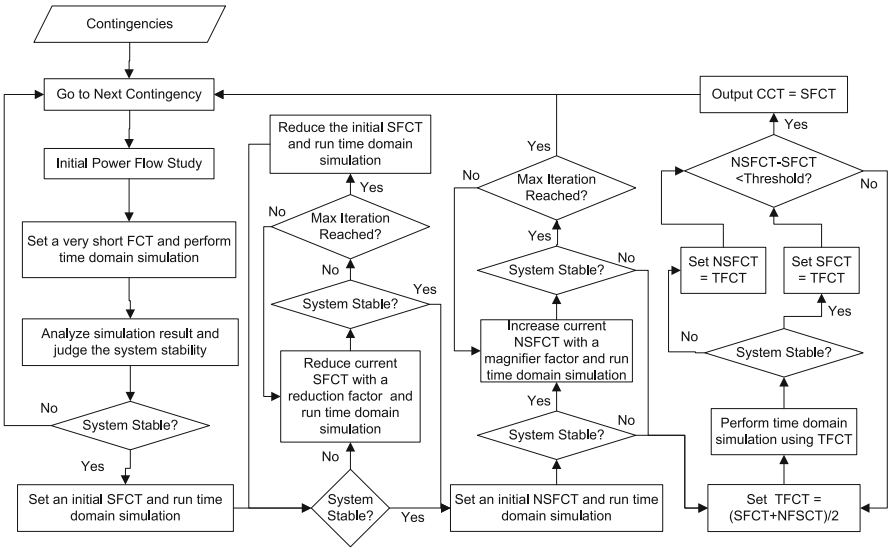


Fig. 2 Flow chart for the time-domain simulation and the computation of CCT

Step II: To capture the stable mode of the post-fault system, an initial guess is assigned to the *stable fault clearing time* (SFCT). The time-domain simulation is performed again using the incumbent SFCT.

- If the system is unstable, then the incumbent SFCT is decreased by multiplying it with a reduction factor, and a time-domain simulation is performed for another round with the updated incumbent SFCT.
- Otherwise, if it hits the maximum number of iterations but the system is still unstable, then we reduce the incumbent SFCT (e.g., assigning the half value of the sum of the incumbent SFCT and FCT, as the new SFCT) and start the procedure again.

Eventually, the program will find a stable mode, and output the SFCT.

Step III: To capture the unstable mode of the system, an initial guess is assigned to the *unstable fault clearing time* (NSFCT). The time-domain simulation is performed using the incumbent NSFCT.

- If the system appears stable, the value of incumbent NSFCT is increased by multiplying it by an amplification factor and a time-domain simulation is carried out again with the updated incumbent NSFCT.
- Otherwise, if the maximum number of iterations is reached but the system is still stable, this suggests that the effect of the fault is so insignificant that the system can always restore the equilibrium state. Then, the program terminates the search for NSFCT and moves on to the simulation of the next contingency.

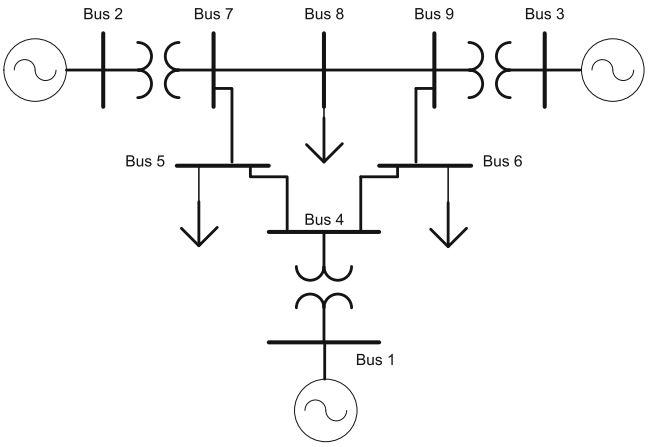


Fig. 3 The one-line diagram of 3-generator 9-bus power system

Following above procedure, the valid SFCT and NSFCT will be output finally, as soon as the simulation of the present contingency is not aborted due to hitting the maximum number of iterations.

Step IV: Observe that CCT must be a value confined between SFCT (stable fault clearing time) and NSFCT (unstable fault clearing time). So, a bisection method can be utilized to find CCT.

To examine the effectiveness of the proposed three-step scheme, it was applied to simulate two electric power systems: one has 3 generators and 9 buses, while the other consists of 4 generators and 57 buses. The one-line diagrams are shown in Figs. 3 and 4. Both systems are simulated subject to various contingencies, as tabulated in Tables 1, 2, 3, and 4, where the simulation adopts the trapezoidal method to compute the pre-fault trajectories ($0 \sim 1.0$ s), fault-on trajectories (1.0 s $\sim$ FCT), and post-fault trajectories (FCT $\sim$ end) of the model (3), with fixed-time step 0.01 s.

For the 3-generator 9-bus model, the system under heavy and/or light-load condition is simulated up to 40 s, and the time-domain simulation subject to certain contingency is shown in Fig. 5. For clarity, the time-domain simulations for all contingencies are presented together and illustrated in Fig. 6. Without applying the three-step scheme, one can directly observe that, there are six different system separations for the power system bearing heavy load, and nine system separations for the lightly loaded system. The actual numbers of system separations are both less than the upper bound $12 = 2 \cdot \binom{2\times 3-2}{3-1}$ in Theorem 3.

For the power system having 4 generators and 57 buses, the post-fault trajectories are shown in Figs. 7 and 8. Above all, for the power system under heavy-load condition, the proposed three-step scheme provides more insights and achieves significant efficiency. The results yielded from Steps 1 and 2 of the proposed scheme are shown in Fig. 9, subject to the list of contingencies in Table 3. Indeed, there are four coarse clusters in Table 5, by pre-grouping the contingencies according

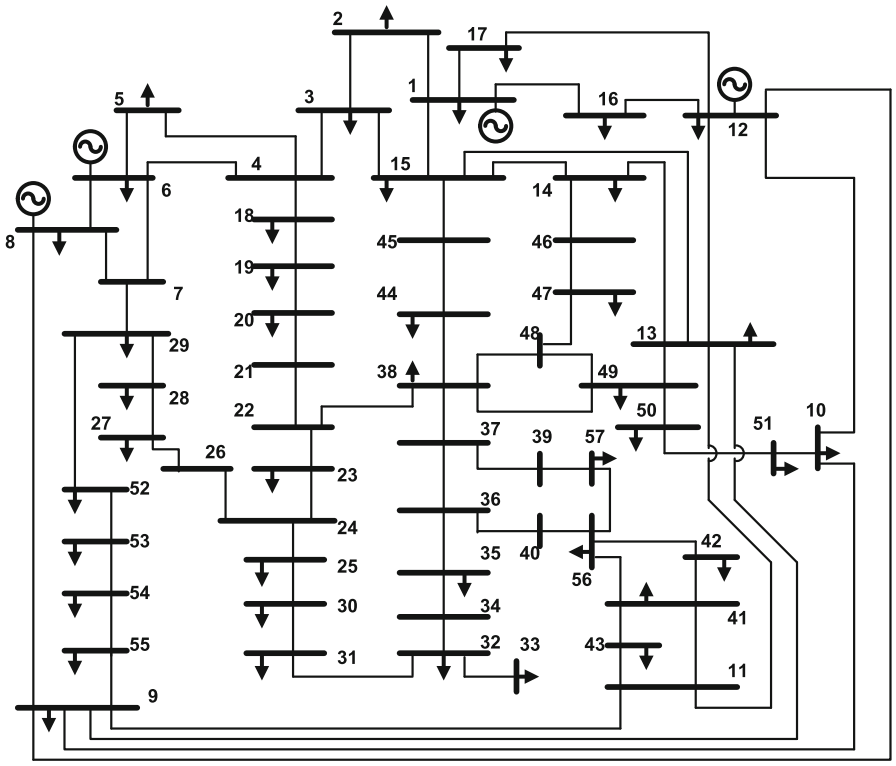


Fig. 4 The one-line diagram of 4-generator 57-bus power system

Table 1 Contingency list for 3-generator 9-bus system bearing heavy load, where FCT refers to the fault clearing time and "*i*-*k*" means the line joining bus *i* and bus *k*

'N-1' contingency					'N-2' contingency				
#	Fault bus	CCT (s)	FCT (s)	Tripped line	#	Fault buses	CCT (s)	FCT (s)	Tripped lines
1	9	1.086685	1.086735	9-8	7	7, 5	1.041648	1.041698	7-5, 5-4
2	8	1.066602	1.066652	9-8	8	7, 4	1.031230	1.031280	7-5, 5-4
3	5	1.103224	1.103274	5-4	9	5, 4	1.047266	1.047316	7-5, 5-4
4	4	1.084212	1.084262	5-4	10	9, 6	1.046012	1.046062	9-6, 6-4
5	6	1.140343	1.140393	6-4	11	9, 4	1.041648	1.041698	9-6, 6-4
6	4	1.115959	1.116009	6-4	12	6, 4	1.059215	1.059265	9-6, 6-4

to the slopes of asymptotes, where the elements in different cluster are indicated by a unique marker, e.g. square '□', diamond '◇', 'x', and '+' in Fig. 9 right. The clusters are further grouped, and eventually it leads to 25 different groups of contingencies as summarized in Table 5. In the simulation, two contingencies are considered to yield a same system separation or belong to a same (final) group, if the Hausdorff distance between the last portion of trajectories (i.e., the portion of last 0.5 s before the end) is less than 2.0, where the system is simulated more

Table 2 Contingency list for 3-generator 9-bus system under light-load condition

'N-1' contingency					'N-2' contingency				
#	Fault bus	CCT (s)	FCT (s)	Tripped line	#	Fault buses	CCT (s)	FCT (s)	Tripped lines
1	9	1.192853	1.192903	9-8	13	9, 7	1.155462	1.155512	9-8, 7-8
2	8	1.288580	1.288630	9-8	14	8, 7	1.190524	1.190574	9-8, 7-8
3	7	1.181565	1.181615	7-8	15	9, 8	1.194004	1.194054	9-8, 7-8
4	8	1.258949	1.258999	7-8	16	7, 5	1.202435	1.202485	7-5, 5-4
5	9	1.201368	1.201418	9-6	17	7, 4	1.193863	1.193913	7-5, 5-4
6	6	1.245604	1.245654	9-6	18	5, 4	1.231970	1.232020	7-5, 5-4
7	7	1.162426	1.162476	7-5	19	9, 6	1.227864	1.227914	9-6, 6-4
8	5	1.238862	1.238912	7-5	20	9, 4	1.180034	1.180084	9-6, 6-4
9	5	1.356447	1.356497	5-4	21	6, 4	1.220925	1.220975	9-6, 6-4
10	4	1.268502	1.268552	5-4					
11	6	1.418502	1.418552	6-4					
12	4	1.255854	1.255904	6-4					

than 40 s and the trajectory of system dynamics usually has length greater than 1000. The number of system separations (which is 25) indeed is less than the upper bound $40 = 2 \cdot \binom{2\times 4-2}{4-1}$ as asserted in Theorem 3. In addition, the system separation corresponding to Group XXV has the highest frequency of occurrence (i.e., 68 out of 93), while the frequency of occurrence for each remaining one does not exceed one thirtieth of that of Group XXV (see Table 5).

For the 4-generator 57-bus power system bearing light load, the three-step scheme is utilized for the contingency analysis as well. As shown in Fig. 10, by Steps 1–2 of the proposed scheme the contingencies in Table 4 are classified into four coarse clusters, and the contingencies in each cluster are tabulated in Table 6. The grouping result obtained at Step 3 is same as that at Step 2, tabulated in Table 6. Thus there are four different system separations for the system under light-load condition, which clearly meets the upper bound (40) presented in Theorem 3.

4 Conclusions and Final Remarks

We have presented an equilibrium problem in the electric power engineering, and linked it to the system separations. Transient stability model and theoretical results on the characterization of stability region and the unstable equilibrium point have been reviewed. In particular, an upper bound for the number of system separations has been summarized in Theorem 3. The main contribution of the present study lies in bringing forward a three-step scheme to compute the system separations yielded from different contingencies. The proposed scheme has been simulated on the electric power systems of three/four generators, which can achieve better efficiency and lead to roughly 90 % reduction in computing time (as tested on a 4-generator

Table 3 Contingency list for 4-generator 57-bus system bearing heavy load

'N-1' contingency					'N-1' contingency				
#	Fault bus	CCT (s)	FCT (s)	Tripped line	#	Fault bus	CCT (s)	FCT (s)	Tripped ine
1	1	1.100514	1.100564	1-2	48	18	1.758292	1.758342	18-19
2	2	1.172091	1.172141	1-2	49	20	2.860644	2.860694	19-20
3	1	1.096465	1.096515	1-15	50	23	1.410856	1.410906	22-23
4	15	1.128139	1.128189	1-15	51	22	1.303229	1.303279	22-38
5	1	1.114229	1.114279	1-16	52	24	2.149893	2.149943	23-24
6	16	1.271685	1.271735	1-16	53	26	2.834199	2.834249	26-27
7	1	1.120106	1.120156	1-17	54	27	2.072748	2.072798	27-28
8	17	1.329678	1.329728	1-17	55	28	1.693964	1.694014	28-29
9	2	1.166736	1.166786	2-3	56	29	1.538924	1.538974	28-29
10	3	1.144597	1.144647	3-4	57	52	3.152941	3.152991	29-52
11	15	1.172026	1.172076	3-15	58	35	2.115851	2.115901	35-36
12	4	1.256730	1.256780	4-6	59	40	1.846049	1.846099	36-40
13	6	1.256730	1.256780	4-6	60	49	1.306102	1.306152	38-49
14	5	1.401321	1.401371	5-6	61	41	1.451020	1.451070	41-42
15	6	1.193904	1.193954	5-6	62	42	3.354345	3.354395	41-42
16	6	1.182148	1.182198	6-7	63	41	1.454416	1.454466	41-43
17	7	1.191422	1.191472	6-7	64	43	1.524360	1.524410	41-43
18	6	1.167911	1.167961	6-8	65	41	1.451216	1.451266	56-41
19	8	1.095028	1.095078	6-8	66	42	3.587159	3.587209	56-42
20	7	1.166213	1.166263	7-8	67	45	1.392504	1.392554	44-45
21	8	1.089216	1.089266	7-8	68	46	1.366969	1.367019	46-47
22	8	1.056962	1.057012	8-9	69	49	1.306690	1.306740	48-49
23	9	1.081640	1.081690	8-9	70	49	1.308780	1.308830	49-50
24	9	1.145380	1.145430	9-10	71	50	1.431885	1.431935	49-50
25	10	1.265480	1.265530	9-10	72	51	1.518417	1.518467	50-51
26	9	1.148058	1.148108	9-11	73	55	1.507642	1.507692	54-55
27	11	1.230346	1.230396	9-11	74	57	2.164995	2.165045	57-56
28	9	1.139035	1.139085	9-12	75	4	1.248762	1.248812	4-18
29	12	1.080203	1.080253	9-12	76	18	1.717148	1.717198	4-18
30	9	1.147274	1.147324	9-13	77	7	1.193969	1.194019	7-29
31	13	1.149168	1.149218	9-13	78	29	1.369843	1.369893	7-29
32	10	1.239815	1.239865	10-12	79	9	1.137347	1.137397	9-55
33	12	1.074979	1.075029	10-12	80	55	1.440897	1.440947	9-55
34	11	1.227668	1.227718	11-13	81	10	1.242166	1.242216	10-51
35	13	1.148123	1.148173	11-13	82	51	1.473094	1.473144	10-51
36	12	1.049027	1.049077	12-13	83	11	1.214737	1.214787	11-41
37	13	1.087518	1.087568	12-13	84	41	1.424766	1.424816	11-41
38	12	1.084187	1.084237	12-16	85	11	1.219831	1.219881	11-43
39	16	1.269987	1.270037	12-16	86	43	1.501372	1.501422	11-43

(continued)

Table 3 (continued)

40	12	1.085363	1.085413	12-17	87	13	1.133102	1.133152	13-49
41	17	1.334576	1.334626	12-17	88	14	1.187700	1.187750	14-46
42	13	1.149691	1.149741	13-14	89	46	1.349858	1.349908	14-46
43	14	1.205398	1.205448	13-14	90	15	1.157397	1.157447	15-45
44	13	1.151127	1.151177	13-15	91	45	1.375328	1.375378	15-45
45	15	1.174573	1.174623	13-15	92	24	2.358469	2.358519	24-26
46	14	1.206835	1.206885	14-15	93	57	2.136995	2.137045	39-57
47	15	1.174050	1.174100	14-15					

Table 4 Contingency list for 4-generator 57-bus system under light-load condition

'N-1' contingency					'N-1' contingency				
#	Fault bus	CCT (s)	FCT (s)	Tripped line	#	Fault bus	CCT (s)	FCT (s)	Tripped line
1	1	1.491576	1.491626	1-2	24	9	1.685148	1.685198	9-11
2	2	3.312544	3.312594	1-2	25	11	3.102454	3.102504	9-11
3	1	1.313221	1.313271	1-15	26	9	1.677474	1.677524	9-12
4	15	1.438024	1.438074	1-15	27	12	1.298135	1.298185	9-12
5	1	1.287163	1.287213	1-16	28	9	1.682862	1.682912	9-13
6	1	1.298461	1.298511	1-17	29	13	1.749149	1.749199	9-13
7	2	3.616085	3.616135	2-3	30	12	1.318054	1.318104	10-12
8	3	2.282794	2.282844	3-4	31	11	3.102686	3.102736	11-13
9	4	1.309302	1.309352	3-4	32	13	1.809722	1.809772	11-13
10	3	1.443836	1.443886	3-15	33	12	1.329352	1.329402	12-13
11	15	1.901479	1.901529	3-15	34	12	1.311639	1.311689	12-16
12	4	2.474043	2.474093	4-5	35	12	1.296759	1.296809	12-17
13	4	2.298550	2.298600	4-6	36	13	1.809395	1.809445	13-14
14	6	1.791028	1.791078	4-6	37	14	2.354959	2.355009	13-14
15	6	1.792497	1.792547	5-6	38	13	1.810293	1.810343	13-15
16	6	1.796252	1.796302	6-7	39	15	1.846049	1.846099	13-15
17	7	2.066463	2.066513	6-7	40	14	2.358796	2.358846	14-15
18	6	1.788497	1.788547	6-8	41	15	1.909071	1.909121	14-15
19	8	1.388129	1.388179	6-8	42	9	1.668984	1.669034	9-55
20	7	1.719679	1.719729	7-8	43	11	3.097335	3.097385	11-41
21	8	1.413142	1.413192	7-8	44	11	3.102143	3.102193	11-43
22	8	1.342022	1.342072	8-9	45	13	1.791109	1.791159	13-49
23	9	1.680821	1.680871	9-10	46	14	2.346877	2.346927	14-46
					47	15	1.853478	1.853528	15-45

57-bus power system) by adopting the mechanism: screening, ranking/preprocessing, and detailed analysis. This scheme demonstrates great potential in the contingency analysis (especially for large-scale electric power systems), and shows that the upper bounds in Theorem 3 that 12 at $n = 3$ and 40 at $n = 4$, are quite tight for the 3-generator 9-bus system and the 4-generator 57-bus system.

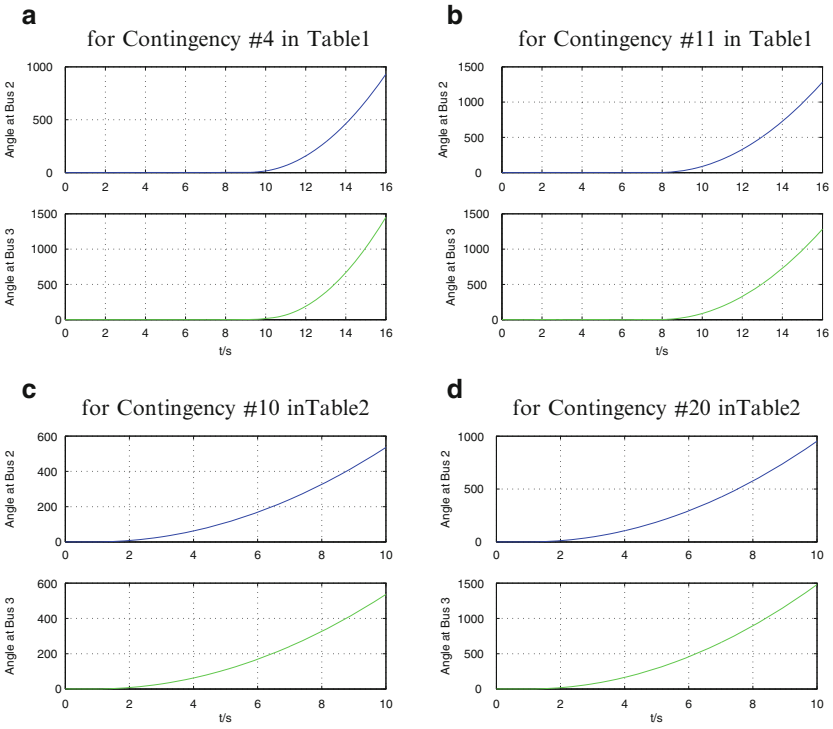


Fig. 5 Time-domain simulation for the 3-generator 9-bus system. Here the *top plots* the post-fault trajectories (i.e. the machine angle at Bus-k over time t) for contingencies in Table 1 and the *bottom ones* are for contingencies in Table 2. (**a**) for Contingency #4 in Table 1, (**b**) for Contingency #11 in Table 1, (**c**) for Contingency #10 in Table 2, (**d**) for Contingency #20 in Table 2

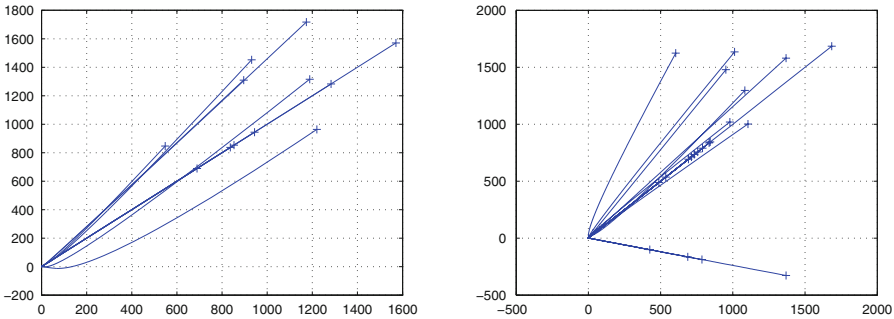


Fig. 6 Time-domain simulations for 3-generator 9-bus system, where the *left plots* all post-fault trajectories subject to the contingencies in Table 1 for the system under heavy-load condition, while the *right plots* the post-fault trajectories subject to the contingencies in Table 2 for the lightly loaded system. The *horizontal* and the *vertical axes* correspond to the machine angle of generator at bus #2, bus #3, respectively

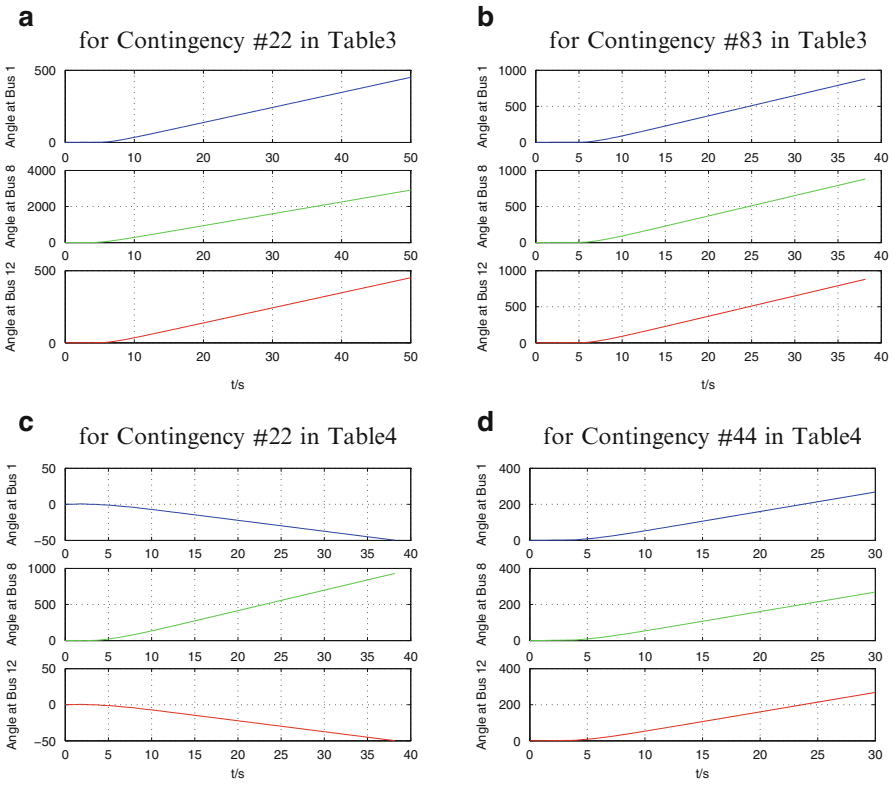


Fig. 7 Time-domain simulation for generators in 4-generator 57-bus system (i.e., the machine angle at Bus-k over time t). Here the *top plots* are simulations for contingencies in Table 3 for the system under heavy-load condition and the *bottom ones* are for contingencies in Table 4 for the system under light-load condition. (**a**) for Contingency #22 in Table 3, (**b**) for Contingency #83 in Table 3, (**c**) for Contingency #22 in Table 4, (**d**) for Contingency #44 in Table 4

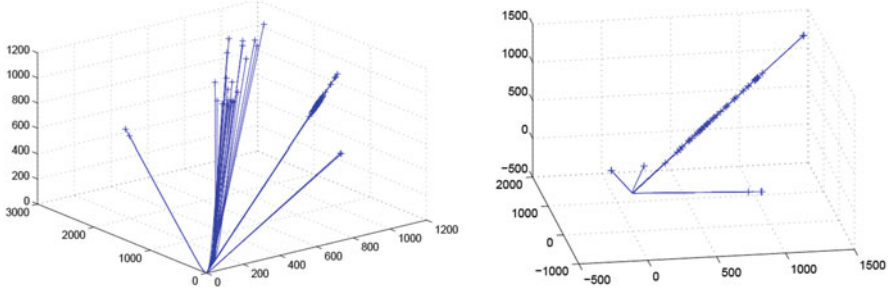


Fig. 8 Time-domain simulations for 4-generator 57-bus system, where the *left plots* all post-fault trajectories subject to the contingencies in Table 3 for the system bearing heavy load, while the *right plots* the post-fault trajectories subject to the contingencies in Table 4 for the lightly loaded system. The x-y-z axes correspond the machine angle of generator at bus #1, bus #8, and bus #12, respectively

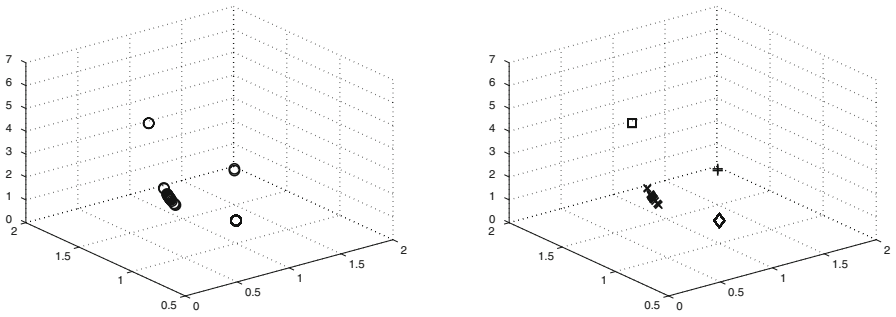


Fig. 9 An illustration for Step 1 (*left plot*, the value of slopes of asymptotes) and Step 2 (*right plot*, the coarse clusters generated by pre-grouping) of proposed scheme applied to 4-generator 57-bus system (under heavy-load condition) in Fig. 8 *left*

Table 5 Pre-grouping result for 4-generator 57-bus system under heavy-load condition, and the elements in the coarse cluster marked by the symbol of square '□', plus-sign '+', cross 'x' and diamond '◇', respectively, in Fig. 9 right

Cluster marker	Contingency #	Cluster marker	Contingency #
□	9, 10	+	49, 53
x	1, 2, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 25, 26, 39, 69, 74, 91, 92, 93	◇	3, 4, 5, 6, 7, 8, 19, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 70, 71, 72, 73, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90

Table 6 Grouping result at Step 3 for 4-generator 57-bus system bearing heavy load

Group No.	Contingency #	Group No.	Contingency #
I	9, 10	II	49
III	53	IV	1
V	2	VI	11
VII	12	VIII	13
IX	14	X	15
XI	16	XII	17
XIII	18	XIV	20
XV	21	XVI	22
XVII	25	XVIII	26
XIX	39	XX	69
XXI	74	XXII	91
XXIII	92	XXIV	93
XXV	3, 4, 5, 6, 7, 8, 19, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 70, 71, 72, 73, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90		

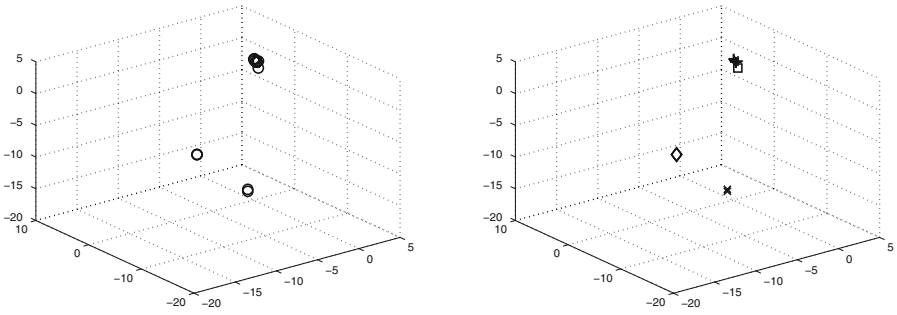


Fig. 10 An illustration for Step 1 (*left plot*, the value of slopes of asymptotes) and Step 2 (*right plot*, the coarse clusters generated by pre-grouping) of proposed scheme applied to 4-generator 57-bus system (under light-load condition) in Fig. 8 *right*

Table 7 Pre-grouping result for 4-generator 57-bus system under light-load condition, and the elements in the coarse cluster marked by the symbol of square '□', plus-sign '+', cross 'x' and diamond '◇', respectively, in Fig. 10 right

Cluster marker	Contingency #	Cluster marker	Contingency #
□	12	+	1, 2, 3, 4, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47
x	5, 8	◇	30, 32, 33

Acknowledgements This work was partially supported by the CERT through the National Energy Technology Laboratory Cooperative Agreement No. DE-FC26-09NT43321, and partially supported by the National Science Foundation, USA, under Award #1225682.

References

1. Hartman, P., Stampacchia, G.: On some non-linear elliptic differential-functional equations. Acta Math. **115**, 271–310 (1966)
2. Kinderlehrer, D.: The coincidence set of solutions of certain variational inequalities. Arch. Ration. Mech. Anal. **40**, 231–250 (1971)
3. Konnov, I.: Equilibrium Models and Variational Inequalities. Elsevier, Amsterdam (2007)
4. Cavazzuti, E., Pappalardo, M., Passacantando, M.: Nash equilibria, variational inequalities, and dynamical systems. J. Optim. Theory Appl. **114**, 491–506 (2002)
5. Smith, M.J.: The existence, uniqueness and stability of traffic equilibria. Transp. Res. B Methodol. **13**, 295–304 (1979)
6. Dafermos, S.: Traffic equilibrium and variational inequalities. Transp. Sci. **14**, 42–54 (1980)
7. Harker, P.T.: A variational inequality approach for the determination of oligopolistic market equilibrium. Math. Program. **30**, 105–111 (1984)
8. Friesz, T.L., et al.: A variational inequality formulation of the dynamic network user equilibrium problem. Oper. Res. **41**, 179–191 (1993)
9. Ferris, M.C., Pang, J.S.: Engineering and economic applications of complementarity problems. SIAM Rev. **39**, 669–713 (1997)

10. Cardell, J.B., Carrie, C.H., Hogan, W.W.: Market power and strategic interaction in electricity networks. Resour. Energy Econ. **19**, 109–137 (1997)
11. Pappalardo, M., Passacantando, M.: Stability for equilibrium problems: from variational inequalities to dynamical systems. J. Optim. Theory Appl. **113**, 567–582 (2002)
12. Dupuis, P., Nagurney, A.: Dynamical systems and variational inequalities. Ann. Oper. Res. **44**, 7–42 (1993)
13. Hantoute, A., Mazade, M.: Lyapunov functions for evolution variational inequalities with locally prox-regular sets. hal.archives-ouvertes.fr (2013)
14. Chang, S.S., et al.: A new method for solving a system of generalized nonlinear variational inequalities in Banach spaces. Appl. Math. Comput. **217**, 6830–6837 (2011)
15. Camlibel, M.K., Pang, J.S., Shen, J.L.: Lyapunov stability of complementarity and extended systems. SIAM J. Optim. **17**, 1056–1101 (2006)
16. Chiang, H.D.: Direct Methods for Stability Analysis of Electric Power Systems. Wiley, New Jersey (2011)
17. Xu, G., Vittal, V.: Slow coherency based cutset determination algorithm for large power systems. IEEE Trans. Power Syst. **25**(2), 877–884 (2010)
18. Hashiesh, F., et al.: An intelligent wide area synchrophasor based system for predicting and mitigating transient instabilities. IEEE Trans. Smart Grid **3**(2), 645–652 (2012)
19. Sun, K., Zheng, D.Z., Lu, Q.: Splitting strategies for islanding operation of large-scale power systems using OBDD-based methods. IEEE Trans. Power Syst. **18**(2), 912–923 (2003)
20. Kasem Alaboudy, A., et al.: Microgrid stability characterization subsequent to fault-triggered islanding incidents. IEEE Trans. Power Syst. **27**(2), 658–669 (2012)
21. Yorino, N., Priyadi, A., Kakui, H., Takeshita, M.: A new method for obtaining critical clearing time for transient stability. IEEE Trans. Power Syst. **25**(3), 1620–1626 (2010)
22. Chiang, H.D., Tong, J., Miu, K.N.: Predicting unstable modes in power systems: theory and computations. IEEE Trans. Power Syst. **8**(4), 1429–1437 (1993)
23. Chiang, H.D., Fekih-Ahmed, L.: Quasi-stability regions of nonlinear dynamical systems: theory. IEEE Trans. Circuits Syst. **43**, 627–635 (1996)
24. Chiang, H.D., Wu, F., Varaiya, P.: Foundations of direct methods for power system transient stability analysis. IEEE Trans. Circuits Syst. **34**(2), 160–173 (1987)
25. Chiang, H.D., Wu, F.F.: Stability of nonlinear systems described by a second-order vector differential equation. IEEE Trans. Circuits Syst. **35**(6), 703–711 (1988)
26. Perko, L.: Differential Equations and Dynamical Systems. Springer, Berlin (2000)
27. Araújo, V., Pacifico, M.J., Viana, M.: Three-Dimensional Flows. Springer, Berlin (2010)
28. Baillieul, J., Byrnes, C.I.: Geometric critical point analysis of lossless power system models. IEEE Trans. Circuits Syst. **29**, 724–737 (1982)
29. Baillieul, J.: The critical point analysis of electric power systems. In: 23rd IEEE Conference on Decision and Control, pp. 154–159, December 1984
30. Chiang, H.D., Wang, T.: On the number of system separations in power system. In: 2015 IEEE International Symposium on Circuits and Systems (ISCAS) (2015)

Security and Formation of Network-Centric Operations

Nicholas J. Daras

Abstract After giving definitions and background information of key terms, we report and analyze the multi-layer graph model of network centric operations (NCOs), and we mention its advantages. Moreover, we investigate security problem of NCOs by applying methods of vertex pursuit games. Finally, in Sect. 6 we take up with the problem of network centric warfare strategic formation.

Keywords: Network centric operations • Multi-layer graph model • Vertex pursuit game • Network centric warfare strategic formation • Network centric operations-graphs • Operational utility function

1 Introduction

This paper explores various concepts related to the Network Centric Warfare framework and investigates security and formation aspects of network centric operations (NCOs). It is divided into six sections. Section 2 deals with definitions and background information of key terms such as Cyber Warfare, Information Warfare, C4ISR, and Network Centric. Special emphasis is given to NCOs Conceptual Framework. Section 3 briefly reports and analyzes the three main thematic NCO-pillars: Net Centric Theoretical Foundations/Mathematical Modeling, Net Centric Technologies and Related Issues and Operational Experiences. Next, in Sect. 4 we apply graph theory concepts to NCO. To do so, we consider Wong-Jiru's multi-layer graph model of NCO and we describe interlayer relationships. Our analysis proceeds with definitions and implications of several NCO-layered metrics (out-degree, in-degree, density, reachability, point connectivity, distance, number of geodesics, maximum flow, network centrality, Freeman degree centrality, betweenness centrality, closeness centrality, edge betweenness, flow betweenness). The section ends with the mention of key advantages of the multi-layer NCO model. Section 5 investigates the security problem of NCOs by applying methods of vertex

N.J. Daras (✉)
Department of Mathematics and Engineering Sciences, Hellenic Military Academy, 16673 Vari Attikis, Greece
e-mail: darasn@sse.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_6

pursuit games. Specifically, we suppose an intruder (or attacker) has invaded into the complex process of a NCO with the intention to destroy or cause sabotage at the vertices of one or more of its five layers (Processes, People, Applications, Systems, Physical Network). The intruder could represent virus or hacker, or other malicious agents intent on avoiding capture. A set of searchers are attempting to capture the intruders. Although placing a searcher on each vertex of a layer guarantees the capture of the intruders, we discuss and investigate the more interesting (and more difficult) problem to find the minimum number of searchers required capturing the intruders. A motivation for minimizing the number of searchers comes from the fact that fewer searchers require fewer resources. NCOs that require a smaller number of searchers may be viewed as more secure than those where many searchers are needed. Finally, in Sect. 6 we take up with the problem of network centric warfare strategic formation. After introducing distance-based operational utility functions, we keep to the study of two layer distance-based operational utilities and of best response NCO-graphs. Then, we consider pairwise operational stability in the NCOs and we conclude with a study of the NCOs formation with arbitrary operational utility functions.

2 Background Information

This section deals with definition and background information of key terms such as Cyber Warfare, Information Warfare, C4ISR, and Network Centric Warfare.

2.1 Cyber Warfare

Definition 1.

i. **Cyberspace** is the notional environment in which digitized information is communicated over computer networks.
ii. **Cyber Warfare** is the use of existing and emerging internet-based technologies to conduct warfare in cyberspace with the aim of attacking and disrupting information systems and communication networks.

2.2 Information Warfare

The term Information Warfare or IW is similar in meaning to Cyber warfare though with a more streamlined goal of achieving competitive advantage. As per the Institute for Advanced Study of Information Warfare (IASW),

Definition 2 ([19]). Information warfare is the offensive and defensive use of information and information systems to deny, exploit, corrupt, or destroy an adversary's information, information-based processes, information systems, and computer-based networks while protecting one's own. Such actions are designed to achieve advantages over military, political, or business adversaries.

Information Warfare is generally subdivided into Information Assurance and Information Denial:

Definition 3.

i. **Information Assurance** focuses on assuring the flow of mission critical information in the event of any attack on the information infrastructure.
ii. **Information Denial** is the offensive part of Information Warfare wherein the focus is to disrupt the adversary's mission critical operations to get a competitive advantage.

Remark 1. Information Assurance is not limited to assuring the availability of information but deals with all the information security goals of not only preserving the CIA (confidentiality, integrity, and availability) of information systems but also ensuring proper authentication and non-repudiation of critical information.

Based on the target audience, Information Warfare can be classified into three classes [25, 26]:

Definitions 4.

i. **Personal Information Warfare** is known as **Class I Information Warfare** and is aimed against individual privacy involving attacks on personal and confidential data.
ii. **Commercial Information Warfare** is known as **Class II Information Warfare** and involves industrial espionage and broadcasting of false information against business rivals using the internet.
iii. **Global Information Warfare** is known as **Class III Information Warfare** and is aimed at countries, political alliances/spheres of influence, global economic forces, sensitive national information systems and infrastructure.

2.3 C4ISR Concept of Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance

Definition 5 ([20]). C4ISR is a term used for effective interfacing of Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance technologies and procedures to deliver a decisive war fighting advantage (Fig. 1).

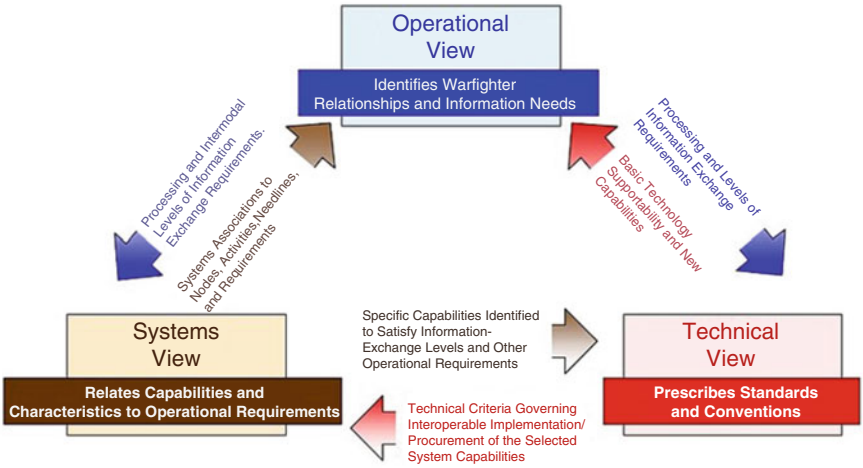


Fig. 1 The C4ISR framework

2.4 Network Centric Warfare

Definition 6 ([15]). Network Centric Warfare (or **Net Centric Warfare** or **NCW**) is a term which broadly describes the combination of emerging tactics, techniques, and procedures that a fully or even partially networked force can employ to create a decisive war fighting advantage.

Remark 2. Network Centric Warfare is also referred to as ***Network Centric Operations*** or ***NCOs***.

Remark 3. The C4ISR framework refers specifically to the military's implementation of a Network Centric Warfare framework.

The objectives of Network Centric Warfare include [16, 17, 27]:

1. Better synchronization of geographically dispersed combat units
2. More effective combat power by networking sensors, weapons, and decision makers
3. Increased speed of executing command and control procedures
4. Seamless interoperability between coalition forces
5. Access to real-time information at every echelon of the military hierarchy
6. Increased survivability and greater lethality in combat operations.

The integration of various weapons and sensors and other combat systems in the three dimensions of land, sea, and air warfare (including support by space-based satellite communication and surveillance) is depicted schematically in Fig. 2 below. This NCW integration includes not only conventional military systems but also specialized systems.

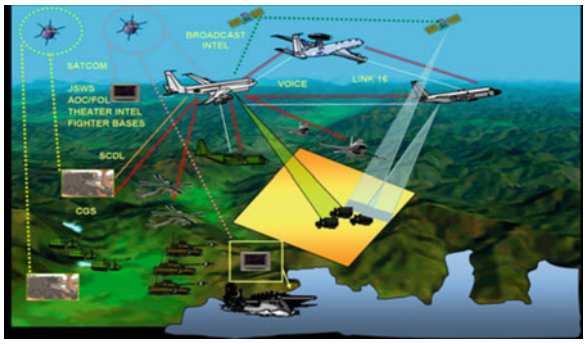


Fig. 2 NCW implementation [17]

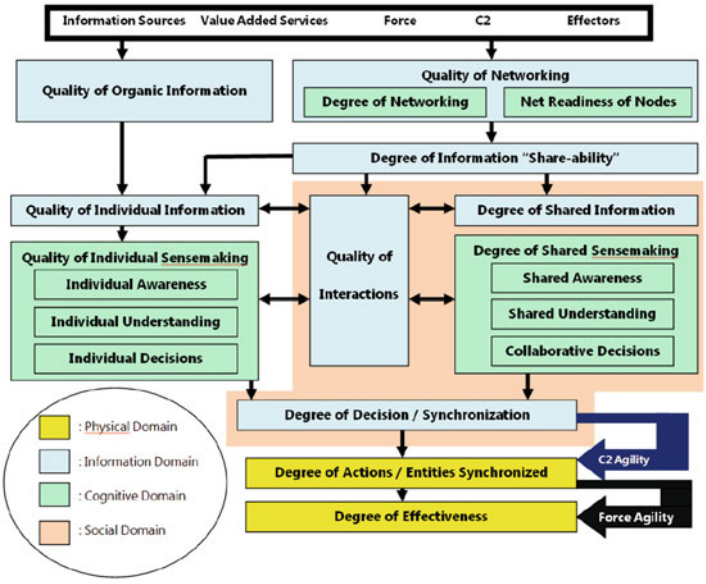


Fig. 3 NCO conceptual framework [2, 21]

Figure 3 above depicts the Conceptual Framework as found in NCO-CF Version 2.0.

3 The Main Thematic Pillars of NCO Approach

There are three central thematic NCO-pillars: **Net Centric Theoretical Foundations/Mathematical Modeling, Net Centric Technologies and Related Issues and Operational Experiences**. Each of these thematic pillars can be further analyzed in individual major sub-issues as follows [28] (Tables 1, 2, and 3).

Table 1 The first thematic NCO-pillar

Net centric theoretical foundations/mathematical modeling issues
(i) NCO scientific theory and tests
(ii) NCO architecture formation
(iii) Overconfidence about the effectiveness of NCO
(iv) Reduced effectiveness for urban counter-insurgency operations
(v) Underestimating our adversaries
(vi) Overreliance on information
(vii) Management of information overload
(viii) Increasing complexity of military systems
(ix) NCO security problem, that is the problem of secure protection from vulnerabilities of military software and data (common Internet threats and vulnerabilities and attacks aimed at destroying the operational capability of critical infrastructure and espionage) and form vulnerabilities of military equipment to electronic warfare

Table 2 The second thematic NCO-Pillar

Net centric technologies and related issues
(i) Command, control, communications, computers, and intelligence
(ii) Interoperability
(iii) Space dominance
(iv) Networked weapons
(v) Bandwidth limitations
(vi) Unmanned robotic vehicles (UVs)
(vii) Sensor technology
(viii) Software design

Table 3 Third thematic NCO-pillar

Operational experiences' issues
(i) Network communications
(ii) Sensors
(iii) Satellites
(iv) Bandwidth and latency
(v) Air dominance
(vi) Operations with coalition forces

In what follows we will restrict ourselves to the first central thematic NCO-pillar and investigate the NCO Architecture Formation (the problem of network formation among a set of nodes where each node forms links with other nodes in the network to maximize some operational utility), as well as the NCO Security problem (secure protection from Vulnerabilities of Military Software and Data and Vulnerabilities of Military Equipment to Electronic Warfare).

4 Applying Graph Theory Concepts to NCO

4.1 The Multi-Layer Graph Model of NCOs

4.1.1 Foundations of the General Theory

If the entities (people, processes, technology) that enable NCO can be considered nodes (or vertices) in a network (resp. graph) or a series of networks (resp. graphs), then one has to understand the **network structures**, **characteristics**, and **dynamics** of those networks (graphs).

How a network is structured and its characteristics are fundamental to understanding what the network is potentially capable of.

To begin, a multi-layer model of NCO, as depicted in Fig. 4, is proposed in [29].

Definition 7. In a multi-layer NCO model:

1. **Each family of contributor to NCO is designated as a layer**. (Thus, People, Applications, Systems, etc., each plays a part in the success of some Process that supports NCO.)
2. **At each layer, the family of contributor is represented graphically as a network**.
3. **The nodes represent individual contributors**.
4. **The edges between the nodes represent a layer-specific relationship**.

Table 4 defines the nodes and edges representation of each layer [29].

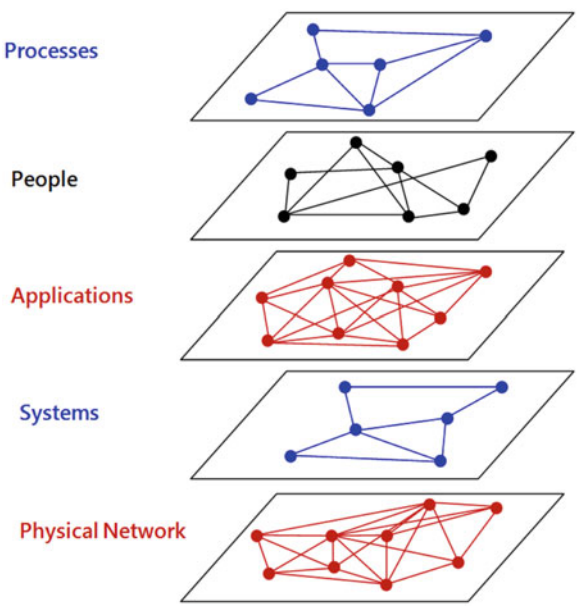


Fig. 4 Layered model of network centric operations

Table 4 Layer, node, and edge definitions for the multi-layer NCO model

Layer name	Layer definition	Node definition	Edge definition
Processes	*Series of tasks in the process of interest that lead to a mission objective.* These processes are based on higher level guidance, such as doctrine or ROEs	*Each node represents one task in the series of tasks*	*Edge between tasks represents the transition of one task to another.* By default, the edge also represents the order in which the tasks are accomplished. A node can have multiple edges if tasks are accomplished concurrently
People	*Actors that perform tasks*	*Each node represents a person or a group of persons*	Edges between persons represent working relationships where specific information is sent or received. A "human network"
Applications	*Tools that send, receive, and/or process information*. These tools may be automated or require an operator interface	*Each node represents an application*. A separate node may be used to designate one copy of an application if multiple copies exist in the network of interest	*Edges between applications represent data-specific interoperability between systems*. The edge is specific to the data that is passed, since systems may be partially interoperable
Systems	*Platform which houses the application(s)* (i.e., an aircraft platform could be grounded but its applications may still function)	*Each node represents a system*	*Edges between systems represent communications interoperability*
Physical network	*Communications infrastructure*	*Each node represents routers, servers, radios, etc.*	*Edges between nodes represent communications pathways*. These edges include both wired and wireless pathways

A summary of the interlayer relationships is shown in Fig. 5.

For this model to be useful for analysis the interlayer relationships must be further defined and a representation for these relationships must be established. Like the node and edge definitions in Table 4, the interlayer relations also have definitions relevant to NCO. These definitions are found in Table 5.

4.1.2 Further Description of Layer and Interlayer Relationships

A. **Task Allocation in the Processes Layer.** Using graph theory, *we can represent any allocating sequential process with several tasks*. For instance, in Fig. 6, the tasks are performed in the following order: first "*a*", then "*b*", etc., ending with task "*e*". The arrows help depict this order. Hence, ***such any process layer is a directed graph***.

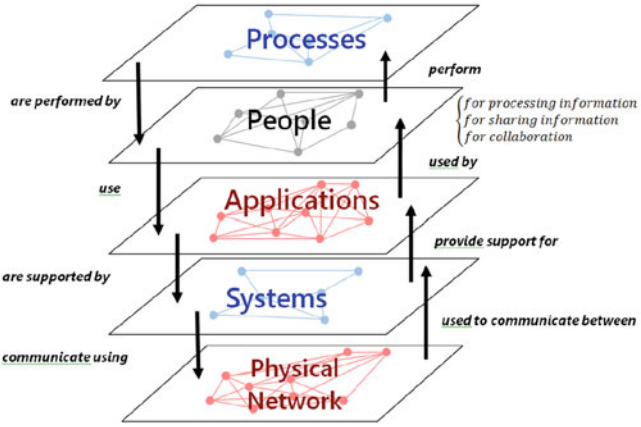
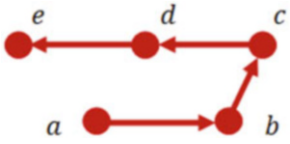


Fig. 5 Interlayer relationships of the multi-layer NCO model [29]

Table 5 Definitions of interlayer relationships of the multi-layer NCO model [29]

Mapping	Node to node mapping	Edge to edge mapping
Process–people	Allocates task to person(s)	Order or route of process tasks through people
People–applications	Identifies the applications used by person(s)	Route of information transactions through applications
Applications–systems	Identifies which systems support which applications. For some, the system and application are the same	Route of information from application to application through supporting systems. For cases where multiple applications are supported by one system, there may be edges from the application layer that "roll-up" into a system node and do not exist on the mapping
Systems–physical network	Identifies which entry points into the communications infrastructure is accessed by which system	Route of communications from one system to another. From a wireless communications perspective, this could represent the route of data transmitted from an aircraft via a radio to a ground node to a radio and back to another aircraft's radio through the physical infrastructure

Fig. 6 A process layer



B. **Working Relationships in the People Layer.** Using graph theory, *we can describe people's engagements in working relationships*. For instance, in Fig. 7, the people layer shows four people who engage in a working relationship.

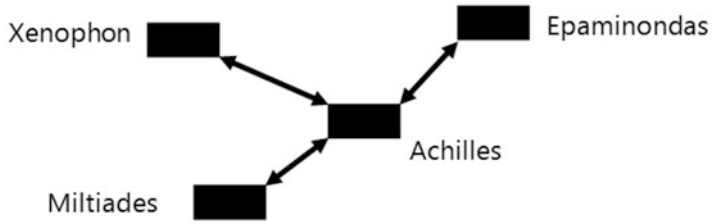


Fig. 7 A people layer

This graph is bi-directed—the edges can be traversed in either direction. In the context of human behavior in a working environment, if Miltiades works with Achilles, then it is assumed that Achilles also works with Miltiades. This graph also depicts that certain persons do not work with each other. For instance, Epaminondas and Xenophon do not work together directly.

C. **The Process–People Mapping.**

Definition 8.

i. In a Multi-layer NCO Model, any process of "allocating" will be termed "**mapping**."

ii. The intermediary layer between the process and people layers is termed the "**process–people mapping**."

Remark 4. The process–people map (allocation) will reflect "**who did what**" and "**when**."

Example 1 ([29]). Having regard to the data of Figs. 6 and 8, the process–people mapping in Fig. 8, below, shows the nodal mapping as well as edge mapping. From a nodal perspective, the graph shows that Miltiades is responsible for task "a", Achilles does task "b", and so on. As depicted, Achilles is actually responsible for two tasks, "*b*" and "*d*". The edge mapping shows the order or "route" of the process as it progresses through the responsible persons. While each layer provides information about each homogenous entity, the mapping provides a graphical representation of the interaction and relationships *between* the entities.

To perform the analysis, each layer and mapping is represented by a series of matrices. Both adjacency and incidence matrices are used.

To accomplish the mapping, two matrices are used.

(a) **In the first matrix, one correlates the vertices of one layer to desired vertices of another layer**.

(b) Likewise, **in the second matrix the edges of both layers are correlated**.

Definition 9. These two comprehensive mapping matrices then served as the **model-wide mapping**, up and down the stack of layers.

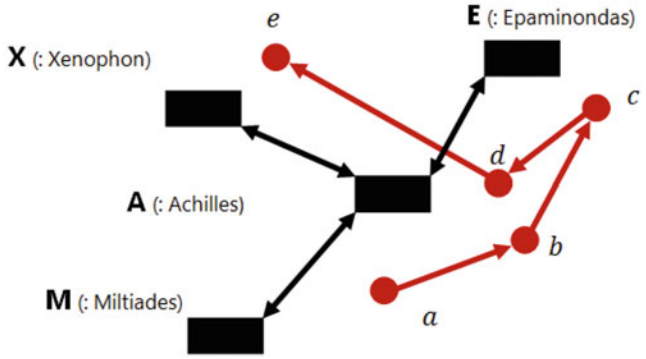


Fig. 8 A process–people mapping

The comprehensive mapping vertices matrix is then used to trace the all vertices associated with one vertex throughout the entire model. The same occurs for the comprehensive edge matrix. Thus, a vertex or edge at any layer of the model may be altered and the effects of that alteration may be traced throughout the other layers. The column and row labels are duplicated because this allows traceability of each node or vertex.

Example 2 (Suite). Continuing with the data of Example 1, as it is represented in Figs. 6, 7 and 8, let us see how the comprehensive vertex and edge matrix for the given two layers would be constructed. Edge labels have been added (Fig. 9).

4.2 Definition of NCO-Layered Graph Metrics

The following quantifiable characteristics or graph metrics will be used in the NCO analysis. Following each definition is a discussion of the possible implication of that metric to NCO analysis, tying the definitions' theoretical meaning to the practical application. These metrics will be used to objectively extend the current NCO Conceptual Framework. While the term "**vertex**" is used in the formal, graph theoretical definitions, the term "**node**" will be used throughout the majority of the analysis because of its common use in network analysis. Further discussion of each graph metrics can be found in [19].

4.2.1 Out-Degree

Definition 10. For a directed graph with vertex v, the **out-degree** $\mathbf{d^-(v)}$ of v is the number of edges with tail v (see page 58 in [27]).

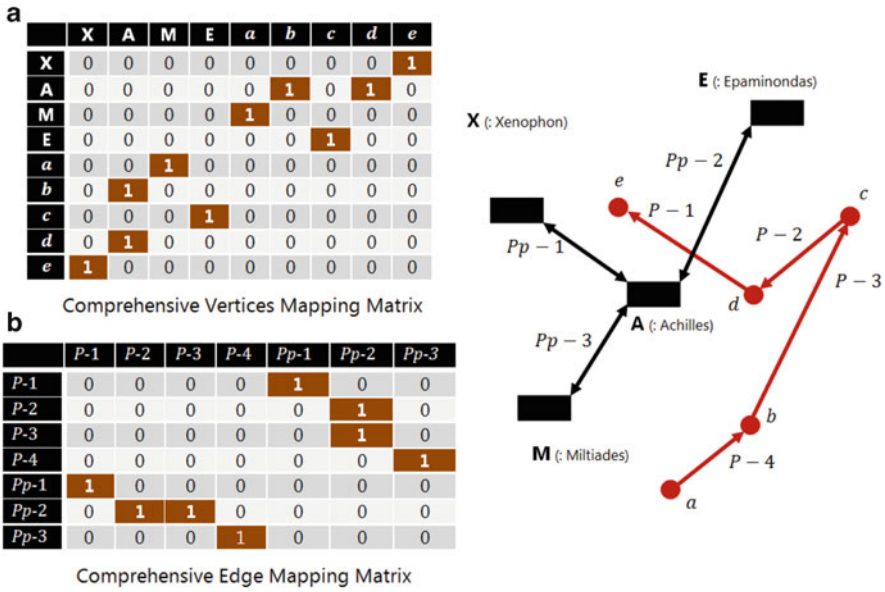


Fig. 9 Comprehensive vertices and edge mapping matrices; (**a**) shows mapping of vertices to vertices, (**b**) shows mapping of edges to edges, and (**c**) the graphical depiction of the mappings

NCW Implication ([29]) A vertex serves as an information source within a network. From a collaboration viewpoint, a vertex with a high $d^-(v)$ may indicate network node with a high level of collaboration with the nodes around them and carries a greater potential to influence its neighbors and the rest of the network. For example, should this node pass inaccurate data into network, more nodes would be affected than if a node with lower $d^-(v)$ had passed on that data. Referencing the layered NCW model, **at the People layer, this node may characterize a commander's position as a commander may be giving orders to multiple supporting commanders**. **At the Process layer, such a node may indicate that many processes rely on this task in order to proceed**.

4.2.2 In-Degree

Definition 11. For a directed graph with vertex v, the **in-degree $\mathbf{d}^+(\mathbf{v})$** of v is the number of edges with head v[1] (see page 58 in [27]).

NCW Implication ([29]) A vertex serves as an information sink within a network. A vertex with high $d^+(v)$ may be a critical convergence point for some activity.

[1] Douglas B. West, Introduction to Graph Theory. Upper Saddle River, NJ: Prentice Hall, 2001.

A high $d^+(v)$ may also be a sign of potential information overload or, since it receives many different inputs, it may be a potential point of conflict. **At the Application layer, a node with high $d^+(v)$ may indicate an application that may benefit from an improvement to its data processing functions to increase efficiency**.

Remark 5. How does one interpret a vertex with both high in- and out-degree? This vertex may be a bottleneck to the overall operations or could benefit from improvements to increase efficiency. For example, these improvements could be increased manning at the People layer, increased automation at the Applications layer. On the other hand, perhaps certain routing or switching systems may also exhibit these characteristics by design.

4.2.3 Density

Definition 12. For a graph G with n vertices and e edges, **the density d** $(\mathscr{G})$ of $\mathscr{G}$ is the ratio of the number of edges to number of vertices *(see pages 435 and 519 in [27]):*

$$d(\mathscr{G}) := e(\mathscr{G})/n(\mathscr{G}).$$

NCW Implication ([29]) For the System layer, the measure of an N-node network's "N^2-connectedness" would be the density. In "Power to the Edge," the N^2 approach is an ill-fated solution to system interoperability in which system $\mathscr{A}$ can be interoperable with all other systems if system $\mathscr{A}$ understands the same language, protocol, etc., with every other system. This scheme would hold true for every system in the network, resulting in a network where every system must be connected to every other system to communicate across the network. This approach results in a very unsustainable network. Measuring density at the System layer would provide a quantifiable network characteristic.

4.2.4 Reachability

Definition 13. The ***reachability*** of a vertex pair (u, v) in a graph $\mathscr{G}$ is a value 1 or 0, provided that there exists or not a path from u to v. The value is 1 if a path exists, 0 if it does not (see page 142 in [9]).

NCW Implication ([29]) Reachability at all layers indicates if there are any unconnected nodes in the network. **Reachability at the Applications, Systems and Physical Network layers can also be an indication of the level of interoperability**. A fully interoperable network would earn value 1 between every pair of nodes. A network may be considered "***weak***" if few nodes are reachable from few others, "***strong***" if many nodes are reachable from many others.

4.2.5 Point Connectivity

Definition 14. The size of the vertex cut of graph $\mathscr{G}$ between two nonadjacent vertices, u and v is **point connectivity**. The **vertex cut** is defined as the smallest set $S \subseteq V(\mathscr{G})$, such that $(\mathscr{G} \backslash S)$ has more than one component (see page 149 in [27]). The point connectivity is another term for the size of the vertex cut.

NCW Implication ([29]) For any layer in the multi-layer graph model of NCO, point connectivity indicates vulnerability of a network between two nodes of interest. For instance, if the point connectivity of node A to node B is three, there are three nodes whose removal would completely disrupt the communication between A and B. Inspection of the network graph would reveal those three specific nodes and a course of action can be developed to prevent any disruptions. Though point connectivity does not explicitly indicate the nodes that compose the vertex cut, it can point to potential problem areas.

4.2.6 Distance

Definition 15. The **distance** $d(u, v)$, also known as the **geodesic distance**, between two vertices in a graph is defined to be minimum distance of the path from vertex u to vertex v *(pages 70 and 520 in [27]).*

Remark 6. Distance is measured by summing the value of each edge connecting each internal vertex along the (u, v) path. The value used here is 1, but may be weighted with other values depending on the context of the graph. For instance, if the edges represented physical distance, the value of each edge may represent mileage between vertices. This metric is an important macrocharacteristic, because it analyzes each possible path across the network between each u and v for all u and v.

NCW Implication ([29]) At the People layer, a high distance $d(u, v)$ may reflect the reach of the circle of influence or social network of a person A [19]. If person B is $d(u, v) = 2$ away from person A, he/she is "*someone who knows someone who knows person A*." Person A's influence is further diluted as $d(u, v)$ increases. The exertion of influence is important both to passing on commander's intent and collaboration between persons. **At the Application layer, a high $d(u, v)$ may indicate that data originating from application A is undergoing d transformations before it is finally usable by application B. At the System layer, long distances may indicate a lack of communications interoperability, if the message being sent from system A must be translated by protocol gateways at each intermediate platform before the destination system can accept it. At the Physical layer, a long distance between nodes may indicate a longer overall network delay as data packets travel through the infrastructure**. In all these cases, the distance metric may be used to streamline for increased efficiency at each layer.

4.2.7 Number of Geodesics

Definition 16. The **number of geodesics** in a graph is the number of shortest paths connecting any pairs of vertices in the graph *(see page 141 of [9]).*

NCW Implication ([29]) This metric is a measure of redundancy at any layer of the NCW model. Multiple paths indicate that two nodes have several ways of reaching each other.

4.2.8 Maximum Flow

Definition 17. In a graph $\mathcal{G}$, the **value of each edge** can represent a capacity. Let $c(x)$ denote the capacity of each edge x of a graph $\mathcal{G}$. A **flow** in $\mathcal{G}$ between two nodes s and t is a function f such that

$$0 \leq f(x) \leq c(x) \text{ for every edge } x.$$

The **maximum flow** between s and t is the sum of the flow along all paths leaving s and arriving at t *(see page 143 in [9]).*

NCW Implication ([29]) For every layer, maximum flow reflects the network wide connectivity, or strength of overall connections, between two nodes. However, the meaning of that connectivity varies for each layer. **At the People layer, maximum flow contributes to the maximum collaborative reach between two persons**. For example, if person A issues an order to all his/her neighbors and those neighbors pass that order on, the maximum flow at person B will be the sum of all the previous connections that order passed through before it reached person B. The greater the value of the maximum flow, the greater the number of persons across the entire network that received that order. **For the Application, System, and Physical layers, the maximum flow is a network-wide snapshot of how widely information could be disseminated throughout the network**.

4.2.9 Network Centrality

The concept of network centrality comes from the study of network structure and the desire to understand how the relative placement of a node in a network may inherently constrain or aid the node's behavior. There are three basic facets of centrality, or network placement: degree centrality, closeness centrality, and betweenness centrality.

4.2.10 Freeman Degree Centrality

As it is already stated in Definitions 7 and 8, for a vertex v in a directed graph $\mathscr{G}$, the ***in-degree centrality*** is $d^+(v)$ and the ***out-degree centrality*** is $d^-(v)$.

Definitions 18.

i. For a vertex v in a directed graph $\mathscr{G}$, the ***in-degree centrality*** of v is simply the number $d^+(v)$, while the ***out-degree centrality*** of v is simply $d^-(v)$.

ii. *For a bidirectional graph, the* ***degree centrality*** *of a vertex v is simply the degree of v:*

$$d(v) := d^+(v) \equiv d^-(v).$$

Remark 7. The degree centrality reflects the direct relationships of a node with others in a graph adjacent to it (see page 167 of [9]). It measures the relative importance of a node within the graph.

The network centrality based on degree is also a useful metric. It provides ***the measure of variability of the degree centrality across the entire network as measured against an ideal star network of the same size***.

Definitions 19.

i. The **Freeman degree centrality**

$$c(v_i)$$

is the degree centrality divided by the maximum possible degree centrality $c_{\max}$, expressed as a percentage.

ii. For a given network with vertices

$$v_1, \ldots, v_n$$

and maximum degree centrality $c_{\max}$, the **network degree centralization measure**, defined for any vertex i, is

$$\sum_{i=1}^{n} [c_{\max} - c(v_i)]$$

divided by the maximum value possible

$$p := \max\{[c_{\max} - c(v_1)], \ldots, [c_{\max} - c(v_n)]\}$$

(see page 167 in [9]).

NCW Implication See in-degree and out-degree.

4.2.11 Betweenness Centrality

Definition 20.

i. Let $b_{x,z} = g_{xyz}/g_{xz}$ be the proportion of all geodesics g, linking vertex x and vertex z which pass through vertex y. The ***betweenness*** b_y of vertex y is the sum of all $b_{x,z}$ where x, y, and z are distinct, i.e.,

$$b_y = \sum_{x,z} b_{x,z}.$$

Betweenness is therefore a measure of the number of times vertex y occurs on a geodesic.

ii. The ***betweenness centrality*** $c(v_i)$ is the betweenness divided by the maximum possible betweenness expressed as a percentage.

iii. For a given network with vertices $v_1, v_2, \ldots, v_n$ and maximum betweenness centrality $c_{\max}$, the ***network betweenness centralization measure*** is the sum

$$\sum_{i=1}^{n} (c_{\max} - c(v_i))$$

divided by the maximum value possible $c_{\max}$ *(see page 171 of [9]).*

NCW Implication ([29]) In general, for all layers, if node A has a high betweenness centrality, it has the greater capacity to facilitate or limit interaction between the nodes it links than other nodes. The criticality of node A is based on which other nodes must use the path that upon which node A lies. From an NCO viewpoint, such a node could become a roadblock or single point of failure. Based upon this criticality, the design of the network at a layer, particularly the Applications, Systems, and Physical Network layers, may require adjustments to address such issues.

4.2.12 Closeness Centrality

Definition 21. The **closeness centrality** $c(u)$ of vertex u is the sum of geodesic distances to all other nodes in graph G:

$$c(u) = \left(\sum_{v} d(u, v) \right)^{-1} \text{ (see page 169 of [9]).}$$

NCW Implications ([29]) In general, for all layers, closeness centrality measures the ability for nodes to access all nodes in the network more quickly than anyone else. The nodes with highest closeness centrality scores would have the shortest paths to the other nodes. For the People layer, this person may be best positioned in the network to disseminate data quickly to others, assuming the applications layer is optimal. These persons may also be best to monitor others in the network

most efficiently. For the Application and System layer, a node with low closeness centrality may signal a node that has low interoperability with other applications and systems.

4.2.13 Edge Betweenness

Definition 22. Let $b_{i,j,k}$ be the proportion of all geodesics linking vertex j and vertex k which pass through edge i. The ***betweenness of edge*** i is the sum of all $b_{i,j,k}$ where j and k are distinct. Betweenness is therefore a measure of the number of times an edge occurs on a geodesic (see page 173 of [9]).

NCW Implication ([29]) In general, for all layers, an edge with a high edge betweenness indicates a critical relationship since many paths contain this edge. For the People layer, this measure will indicate a very important relationship, perhaps one that has a high collaboration potential. For the Application layer, this measure will highlight a critical interoperability link. For the System and Physical Network layer, an edge with high edge betweenness could indicate a more heavily used communications and infrastructure link, respectively.

4.2.14 Flow Betweenness

Definition 23. Let $m_{i,j,k}$ be the amount of flow between vertex j and vertex k which must pass through i for any maximum flow. The ***flow betweenness*** of vertex i is the sum of all $m_{i,j,k}$ where i, j, and k are distinct and $j < k$. The flow betweenness is therefore a measure of the contribution of a vertex to all possible maximum flows (see page 177 of [9]).

Remark 8. The flow betweenness centrality $c(v_i)$ of a vertex i is the flow betweenness of i divided by the total flow through all pairs of points where i is not a source or sink. For a given network with vertices $v_1, v_2, \ldots, v_n$ and maximum flow betweenness centrality $c_{\max}$, the network flow betweenness centralization measure is the sum $\sum_{i=1}^{n} (c_{\max} - c(v_i))$ divided by the maximum value possible $c_{\max}$, where $c(v_i)$ is the flow betweenness centrality of vertex v_i (see page 177 of [9]).

NCW Implication ([29]) For all layers, the flow betweenness is a measure of the possible workload performed by each node if all maximum flows were utilized.

4.3 Advantages of the Multi-Layer NCO Model

The advantages of this layered model are the following [29].

1. **Network analysis metrics may be applied at any level**, allowing each layer to be analyzed

2. **The mapping between layers allows the traceability of cause-and-effect** from either bottom-up (i.e., effect of loss of people on the completion of the process) or top-down (i.e., consolidation of application on the type of platform supporting it.
3. **Upholds and provides additional insight** to the concepts in the NCO-CF (see pages 63 and 64 of [9])
4. Integrally accounts for the accomplishment of commander's intent via the processes layer into the model. Thus, **objective operational effectiveness measures on completion of processes can be made to support assessments**.
5. **Allows flexibility for the audience to determine the amount of detail at each layer**. Layers, vertices, and edges may be defined to suit the level of analysis desired.
6. When vertices and edges are specifically labeled, **commanders can trace the specific effect to a cause in the NCO system as a whole**.
7. **The layered model, coupled with the above metrics, produces a holistic view of the networks** involved for the successful execution of a mission objective at the Process layer.

   (a) **Individual nodes/edges.** For the applied metrics, nodes and edges produce individual characteristics, allowing a detailed look at each contributor to the network.
   (b) **Individual layer.** The network at each layer produces characteristics which can be collected into a view depicted in a radar chart. Each layer is then assigned a composite network score, which is calculated by normalizing the area under the curve of the radar graph.
   (c) **Network Centricity Score.** The network centricity score, *NC*, provides a holistic score for all the layers. For i layers, $NC = \prod_i N_i$. The initial *NC* score may be used as a baseline. When changes are made to any layer(s), the recalculated *NC* score will indicate the relative merit of those changes.
   (d) **Mission Effectiveness.** The measure of mission effectiveness resides at the Process layer, since the lower layers support the completion of a process. The Process layer consists of tasks (nodes) and transitions (edges). Both the task and transition must be accounted for in this measure, because a task may be completed but not successfully transitioned to the next task. Therefore, the degree of mission effectiveness could be expressed as the sum of the ratio of tasks and edges completed.

5 Security of NCOs

5.1 Vertex Pursuit Games in NCO Security Modeling

Suppose a number of intruders (or attackers) have invaded into the complex process of a NCO with the intention to destroy or cause sabotage at the vertices of one

or more of its five layers (Processes, People, Applications, Systems, Physical Network). The intruders could represent viruses or hackers, or some other malicious agents intent on avoiding capture. A set of **searchers** are attempting to capture the **intruders**. Although placing a searcher on each vertex of a layer guarantees the capture of the intruders, it is a more interesting (and more difficult) problem to **find the minimum number of searchers required to capture the intruders**. A motivation for minimizing the number of searchers comes from the fact that fewer searchers require fewer resources. NCOs that require a smaller number of searchers may be viewed as more secure than those where many searchers are needed.

In this paper, we assume that *the number of intruders has been limited in number one at each layer* and that *the invasion has taken place in at least one layer of the NCO*. Of course, the same approach can be applied when an intruder penetrates every layer. Then, the solution will be given below can be applied easily to each layer separately. However, the general problem in which several intruders are loose on the vertices of one or more layers is open and can be the subject of other scientific investigations.

Vertex pursuit games would be a suitable model for such simplified network security problems with only one-layer-intruder [23, 24]. To see this, observe that the five layers of an NCO (Processes' layer, People's layer, Applications' layer, Systems' layer and Physical Network's layer) may be viewed as five undirected, simple, and finite graphs G_1, G_2, G_3, G_4, and G_5. The k_i NCO-searchers begin by occupying a set of k_i vertices in G_i ($i = 1, 2, 3, 4, 5$). The intruder in the i NCO-layer then chooses a vertex of the G_i, and the k_i NCO-searchers and intruder in the i NCO-layer move in alternate rounds. The NCO-controllers (or NCO-supervisors) use edges to move the k_i NCO-searchers from vertex to vertex in the G_i. More than one NCO-searcher is allowed to occupy such a vertex, and the NCO-controllers may remain on their current vertex. The NCO-controllers know each other's current locations and can remember all the previous moves. The k_i NCO-searchers win if at least one of the k_i NCO-searchers can eventually occupy the same vertex as the intruder; otherwise, the intruder wins. It is understood that the whole process should last a predetermined duration. As placing a k_i NCO-searcher on each vertex of the G_i guarantees that the k_i NCO-searchers win, we may define the NCO-*search number in* G_i, written

$$s\left(G_i\right),$$

which is the minimum number of k_i NCO-searchers needed to win on G_i. Such a number was first introduced by Aigner and Fromme [1] who proved (among other things) that if G_i were planar, then $s\left(G_i\right) \leq 3$. For a survey of results on vertex pursuit games, the reader is directed to the surveys [3, 15, 18]. While searchers and intruders have been extensively studied in highly structured deterministic graphs such as graph products (see [22]), the work [8] is the first to consider such vertex pursuit games in random models of complex networks [5, 10], such as NCO.

We will consider random NCO-graphs (in the sense of Erdös-Rényi) and their generalizations used to model complex networks. The ***random NCO-graph*** $G_i(\mathbb{N}_i; p_i)$ in the i-layer ($i = 1,\ 2,\ 3,\ 4,\ 5$) consists of the probability space

$$(\Omega_i, \mathscr{F}_i, \mathbb{P}),$$

where Ω_i is the set of all NCO-graphs in the i-layer with vertex set $\mathbb{N}_i$ (with $|\mathbb{N}_i| = n_i$), $\mathscr{F}_i$ is the family of all subsets of Ω_i, and, for every $G_i \in \Omega_i$,

$$\mathbb{P}(G_i) = p_i^{|E(G_i)|}(1-p_i)^{\binom{n_i}{2}-|E(G_i)|}.$$

This space may be viewed as $\binom{n_i}{2}$ independent coin flips, one for each pair of vertices, where the probability of success (that is, drawing an edge) is equal to p. Note that $p_i = p_i(n_i)$ can tend to zero with n_i.

All asymptotics throughout are as $n_i \to \infty$. We say that an event in a probability space holds ***asymptotically almost surely*** (***a.a.s.***) if the probability that it holds tends to 1 as n_i goes to infinity. For $p \in (0, 1)$ or $p = p(n_i)$ tending to 0 with n_i, define

$$\mathbb{L}_i n_i := \log\frac{1}{1-p_i} n_i.$$

According to [6], the following result holds.

Theorem 1. *Let $i = 1,\ 2,\ 3,\ 4,\ 5$ and $0 < p_i < 1$ be fixed. For every real $\epsilon > 0$ a.a.s. for $G_i \in G_i(\mathbb{N}_i; p_i)$*

$$(1-\epsilon)\,\mathbb{L}_i n_i \le s(G_i) \le (1+\epsilon)\,\mathbb{L}_i n_i.$$

The problem of determining the NCO-search number of $G_i(\mathbb{N}_i; p_i)$ where $p_i = p_i(n_i)$ is a function of n was left open in [6]. However, it can be showed that *the NCO-search number of $G_i(\mathbb{N}_i; p_i)$ is always bounded from above by $n_i^{(1/2)+o(1)}$ and this bound is achieved for sparse random graphs*. More precisely, it can be showed that

$$s(G_i(\mathbb{N}_i; p_i)) \le 160,000\sqrt{n_i}\log n_i, \text{ whenever } n_i p_i \le 2.1\log n_i \text{ and}$$

$$s(G_i(\mathbb{N}_i; p_i)) \ge \frac{1}{(n_i p_i)^2}\ n_i^{\frac{1}{2}\ \frac{\log\log(n_i p_i)-2}{\log\log(n_i p_i)}},\ \text{for } n_i p_i \to \infty.$$

Since "*if either $n_i p_i = n_i^{o(1)}$ or $n_i p_i = n_i^{(1/2)+o(1)}$, then a.a.s. $s(G_i(\mathbb{N}_i; p_i)) = n_i^{(1/2)+o(1)}$*," it would be natural to "*assume that the NCO-searcher number of $G_i(\mathbb{N}_i; p_i)$ is close to $\sqrt{n_i}$ also for $n_i p_i = n_i^{\alpha+o(1)}$, where $0 < \alpha < 1$.*"

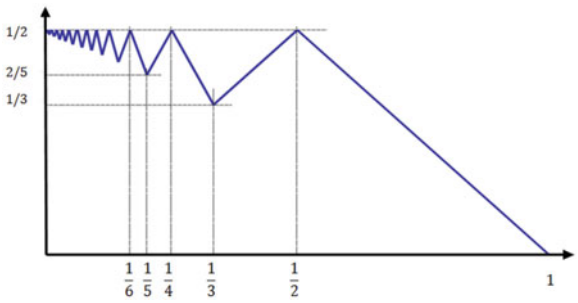


Fig. 10 The "zigzag" function f

It can be shown that the actual behavior of $s\left(G_i\left(\mathbb{N}_i; p_i\right)\right)$ is more complicated. In fact, the function

$$f : (0,1) > \mathbb{R} : x \mapsto f(x) = \frac{\log \mathbb{E}\left(s\left(G_i\left(\mathbb{N}_i; \mathbb{N}_i{}^{x-1}\right)\right)\right)}{\log n_i};$$

f has an unexpected zigzag shape; see Fig. 10 above. Here

$$\mathbb{E}\left(s\left(G_i\left(\mathbb{N}_i; \mathbb{N}_i{}^{x-1}\right)\right)\right)$$

denotes the expected value of the NCO-searcher number for $G_i\left(\mathbb{N}_i; p_i\right)$. A main result is that

In the next subsection, we will show that *if* $n_i p_i = {n_i}^{\alpha+o(1)}$, *where* $(1/2) < \alpha \leq 1$, *then*
$s\left(G_i\left(\mathbb{N}_i; p_i\right)\right) = \left(\log\left(n_i/p_i\right)\right) = {n_i}^{1-\alpha+o(1)}$ *and* $s\left(G_i\left(\mathbb{N}_i; {n_i}^{-(1/2)+o(1)}\right)\right) = {n_i}^{(1/2)+o(1)}$ *a.a.s.*

Recent work by Chung and Lu [10, 11] supplies an extension of the $G_i\left(\mathbb{N}_i; p_i\right)$ random NCO-graphs to random NCO-graphs $G_i\left(\mathbf{w}^{(i)}\right)$ in the i-layer with given expected degree sequence $\mathbf{w}^{(i)}$: For example, if $\mathbf{w}^{(i)}$ follows a power law distribution in the i-layer, then $G_i\left(\mathbf{w}^{(i)}\right)$ supplies a model for NCO. We determine bounds on the NCO-searcher number of random power law NCO-graphs as discussed in the next subsection.

5.2 Results

We now consider the NCO-searcher number $s\left(G_i\right)$ of a classical random NCO-graph $G_i\left(\mathbb{N}_i; p_i\left(n_i\right)\right)$ in the i-layer, when $p_i\left(n_i\right)$ is a function of $n_i = |\mathbb{N}_i|$. We will abuse notation and refer to p_i rather than $p_i\left(n_i\right)$. The main results are summarized as follows (see also [7]).

Theorem 2 ([7]).

(1) *In an NCO, suppose that* $i = 1, 2, 3, 4, 5$ *and* $p_i \geq p_i^{(0)}$ *where* $p_i^{(0)}$ *is the smallest* p_i *for which*

$$\left(p_i^2/40\right) \geq \frac{\log\left(\left(\log^2 n_i\right)/p_i\right)}{\log n_i}$$

holds. Then a.a.s. the graph $G_i \in G_i\left(\mathbb{N}_i; p_i\right)$ *in the i-layer of the NCO satisfies*

$$\mathbb{L}_i n_i - \mathbb{L}_i\left(\left(p_i^{-1}\mathbb{L}_i n_i\right)\left(\log n_i\right)\right) \leq s\left(G_i\right) \leq \mathbb{L}_i n_i - \mathbb{L}_i\left(\left(\mathbb{L}_i n_i\right)\left(\log n_i\right)\right) + 2.$$

(2) *If* $\left((2\ \log\ n_i)/\sqrt{n_i}\right) \leq p_i = \mathrm{o}(1)$ *and* $\omega\left(n_i\right)$ *is any function tending to infinity, then a.a.s the graph* $G_i \in G_i\left(\mathbb{N}_i; p_i\right)$ *in the i-layer of satisfies*

$$\mathbb{L}_i n_i - \mathbb{L}_i\left(\left(p_i^{-1}\mathbb{L}_i n_i\right)\left(\log n_i\right)\right) \leq s\left(G_i\right) \leq \mathbb{L}_i n_i + \mathbb{L}_i\left(\omega\left(n_i\right)\right).$$

By Theorem 2, we have the following corollary.

Corollary 1 ([7]). *If* $i = 1, 2, 3, 4, 5$ *and* $p_i = {n_i}^{-\mathrm{o}(1)} < 1$, *then a.a.s.* $G_i \in G_i\left(\mathbb{N}_i; p_i\right)$ *satisfies*

$$s\left(G_i\right) = (1 + \mathrm{o}(1))\,\mathbb{L}_i n_i.$$

Indeed, from part (1) it follows that if p_i is a constant, then

$$s\left(G_i\right) = \mathbb{L}_i n_i - 2\mathbb{L}_i \log n_i + (1) = (1 + \mathrm{o}(1))\,\mathbb{L}_i n_i.$$

From part (2), for $p_i = {n_i}^{-\mathrm{o}(1)}$ tending to zero with n_i, the lower bound is

$$\begin{aligned}\mathbb{L}_i n_i - \mathbb{L}\left(\left(p_i^{-1}\mathbb{L}_i n_i\right)\left(\log n_i\right)\right) &= \mathbb{L}_i n_i - 2\mathbb{L}_i\left((1 + \mathrm{o}(1))\,p_i^{-1}\log n_i\right) \\ &= \mathbb{L}_i n_i - 2\mathbb{L}_i\left({n_i}^{\mathrm{o}(1)}\right) = (1 + \mathrm{o}(1))\,\mathbb{L}_i n_i.\end{aligned}$$

Note that for $p_i = {n_i}^{-\alpha(1+\mathrm{o}(1))}$ $(0 < a < (1/2))$ we do not have a concentration for $s\left(G_i\right)$ but the following bounds hold

$$(1 + \mathrm{o}(1))\,(1 - 2a)\mathbb{L}_i n_i \leq s\left(G_i\right) \leq (1 + \mathrm{o}(1))\,\mathbb{L}_i n_i :$$

Let us finally describe results for the NCO-searcher number $s\left(G_i\right)$ of random power law NCO-graphs in the i-layer $(i = 1, 2, 3, 4, 5)$. Let $\mathbf{w}^{(i)} = \left(w_1^{(i)}, \ldots, w_\ell^{(i)}\right)$ be any finite sequence of ℓ nonnegative real numbers. We define a random NCO-graph model in the i-layer, written $G_i\left(\mathbf{w}^{(i)}\right)$, as follows. Typically, vertices are integers in the vertex set $\mathbb{N}_i$. Each potential edge between a and b is chosen independently with probability $p_{a,b}^{(i,\ell)} = w_a^{(i)} w_b^{(i)} \rho_{i,\ell}$, where

$$\rho_{i,\ell} := 1/\sum_{\lambda=1}^{\ell} w_\lambda^{(i)}.$$

We will always assume that

$$\max_{\lambda}\left[w_{\lambda}^{(i)}\right]^2 < \sum_{\lambda=1}^{\ell} w_{\lambda}^{(i)}$$

which implies that $p_{a,b}^{(i,\ell)} \in [0,1[$. The model $G_i\left(\mathbf{w}^{(i)}\right)$ is referred to as *random NCO-graphs in the i-layer with given expected degree sequence* $\mathbf{w}^{(i)}$. Observe that $G_i\left(\mathbb{N}_i; p_i\right)$ may be viewed as a special case of $G_i\left(\mathbf{w}^{(i)}\right)$ by taking $\mathbf{w}^{(i)}$ to be equal the constant ℓ-sequence $(p_i n_i,\ p_i n_i, \ldots,\ p_i n_i)$.

Given $\beta > 2$, $d > 0$, and a function $M = M(n_i) = \mathrm{o}\left(p_i n_i\right)$ (with M tending to infinity with n_i), let us consider the random graph in the i-layer with given expected degrees $w_{\lambda}^{(i)} > 0$, where

$$w_{\lambda}^{(i)} = \mathbf{c}_i\, \lambda^{-1/(\beta-1)} \tag{1}$$

for λ satisfying $\lambda_0 \leq \lambda < n_i + \lambda_0$. The term $\mathbf{c}_i$ depends on i, β and d, and λ_0 depends also on M; namely,

$$\mathbf{c}_i = \left(\frac{\beta-2}{\beta-1}\right) d\, {n_i}^{1/(\beta-1)} \quad \textit{and} \quad \lambda_0 = n_i\left(\frac{d}{M}\left(\frac{\beta-2}{\beta-1}\right)\right)^{\beta-1}. \tag{2}$$

It is not hard to show (see [10, 11]) that

Proposition 1. *Asymptotically almost surely, the random NCO-graphs in the i-layer with the expected degrees satisfying* (1) *and* (2) *follow a power law degree distribution with exponent* β*, average degree* $(1 + \mathrm{o}(1))\, d$*,* and *maximum degree* $(1 + \mathrm{o}(1))\, M$.

The next theorem shows that the NCO-searcher number of random power law graphs in the i-layer is a.a.s. (n_i), and so is of much larger order than the logarithmic NCO-searcher number of $G_i\left(N_i; p_i\right)$ random NCO-graphs in the i-layer. Hence, these results are suggestive that in power law NCO-graphs in the i-layer, on average a large number of NCO-searcher are needed to secure the network.

Theorem 3 ([7]). *Let* $i = 1, 2, 3, 4, 5$. *For a random power law NCO-graph* $G_i \in G_i\left(\mathbf{w}^{(i)}\right)$ *in the i-layer with exponent* $\beta > 2$ *and average degree d, a.a.s. the following hold.*

(1) *If X is the random variable denoting the number of isolated vertices in* $G_i\left(\mathbf{w}^{(i)}\right)$ *and* $\Gamma(\cdot,\cdot)$ *is the incomplete gamma function, then*

$$\begin{aligned} s\left(G_i\right) \geq X &= (1 + \mathrm{o}(1))\, n_i \int_0^1 \exp\left(-d\,\frac{\beta-2}{\beta-1}\, x^{-1/(\beta-1)}\right) dx \\ &= (1 + \mathrm{o}(1))\, (d\,(\beta-2))^{\beta-1} (\beta-1)^{2-\beta}\, n_i\, \Gamma\left(1-\beta,\ d\frac{\beta-2}{\beta-1}\right). \end{aligned}$$

Table 6 Bonato et al. [7]: upper and lower bounds for the NCO-searcher number of $G_i(\mathbf{w})$ for various values of d (top row) and β (left column)

	10	20
2.1	0.1806/0.2940	$0.5112 \cdot 10^{-1}/0.1265$
2.7	$0.4270 \cdot 10^{-2}/0.1895$	$0.4205 \cdot 10^{-4}/0.8261 \cdot 10^{-1}$

(2) *For* $u \in]0, 1[$, *define*

$$f(u) := u + \int_u^1 \exp\left(-d\,\frac{\beta-2}{\beta-1}\,u^{(\beta-2)/(\beta-1)}\,x^{-1/(\beta-1)}\right)dx:$$

Then

$$s(G_i) \leq (1 + o(1))\; n_i\;\; \min_{0<u<1} f(u).$$

We note that integrals in the statement of Theorem 3 do not possess closed-form solutions in general. As in [7], numerical values may be supplied for lower/upper bounds of the NCO-searcher number of $G(\mathbf{w})$ when $d = 10,\ 20$ and $\beta = 2.1,\ 2.7$ (Table 6) (note that the values of $d = 10$ and $\beta = 2.1$ coincide with earlier experimental results found in the web graph; see, for example, the survey [5]).

While Theorem 3 suggests a large number of k_i NCO-searchers are needed to secure complex networks against intruders, by item (1) it is the abundance of isolated vertices in G_i that makes the cop number equal to (n_i). To overcome the issue with isolated vertices, we consider restricting the movements of the cops and robber to the subgraph induced by sufficiently high degree vertices.

Fix $\beta \in]2, 3[$. Define the ***core*** of the NCO-graph G_i, written

$$\widehat{G}_i,$$

as the subgraph induced by the set of vertices of degree at least ${n_i}^{1/\log\log n_i}$. Random power law graphs with $\beta \in]2, 3[$ are referred to as *octopus* graphs in [5], since the core is dense with small diameter $(\log\log n_i)$ and the overall diameter is $(\log n_i)$. For $G_i \in G_i\left(\mathbf{w}^{(i)}\right)$, since the expected degree of vertex λ in G_i is

$$w_\lambda^{(i)} = \frac{\beta-2}{\beta-1}\,d\,{n_i}^{1/(\beta-1)}\lambda^{-1/(\beta-1)},$$

vertices with expected degree at least ${n_i}^{1/\log\log n_i}$ have labels at most

$$\lambda_{\Lambda_i} = \left(\frac{\beta-2}{\beta-1}\,d\right)^{\beta-1}\,{n_i}^{1-(\beta-1)/\log\log n_i}.$$

The order of the core is written Λ_i. By the Chernoff's bound,

$$\Lambda_i = (1 + \mathrm{o}(1))\, i_{\Lambda_i} - i_0 = (1 + \mathrm{o}(1))\, i_{\Lambda_i} = \left(n_i^{1-(\beta-1)/\log\log n_i}\right),$$

provided that

$$\log M \gg \frac{\log n_i}{\log\log n_i}.$$

Thus,

$$n_i = {\Lambda_i}^{1+(\beta-1)/\log\log n_i + (1)/\log^2\log \Lambda_i}. \tag{3}$$

We consider the NCO-searcher number of the k_i searchers of random power law NCO-graphs in the i-layer (so the cop and robber are restricted to movements within the core). As vertices in the core informally represent the *hubs* of the network, one would suspect that the NCO-searcher number of the core is of smaller order than the core itself. This intuition is made precise by the following theorem, which provides a sublinear upper bound for the NCO-searcher number of the core in the i-layer.

Theorem 4 ([7]). *Let $i = 1, 2, 3, 4, 5$. For a random power law NCO-graph $G_i \in G_i(\mathbf{w})$ in the i-layer with power law exponent $\beta \in]2, 3[$ a.a.s. the NCO-searcher number of the core $\widehat{G}_i$ of G_i satisfies*

$${\Lambda_i}^{(1+\mathrm{o}(1))(3-\beta)/\log\log\Lambda_i} \leq s\left(\widehat{G}_i\right) \leq {\Lambda_i}^{1-(1+\mathrm{o}(1))(\beta-1)(3-\beta)/(\beta-2)\log\log\Lambda_i}.$$

As the asymptotic bounds in Theorem 4 are not tight, it is an interesting open problem to determine the asymptotic value of the cop number of the core of random power law graphs.

6 Network Centric Strategic Formation

6.1 Distance-Based Operational Product Utility of NCO

As usual, the five layers of an NCO (Processes' layer, People's layer, Applications' layer, Systems' layer, and Physical Network's layer) may be viewed as five undirected, simple, and finite graphs

$$G_1 = (\mathbb{N}_1; \mathbb{E}_1), G_2 = (\mathbb{N}_2; \mathbb{E}_2), G_3 = (\mathbb{N}_3; \mathbb{E}_3), G_4 = (\mathbb{N}_4; \mathbb{E}_4), \text{ and } G_5 = (\mathbb{N}_5; \mathbb{E}_5).$$

A canonical problem in NCO formation involves distance-based utilities in each layer [12]. In this multi-layer case, there is a net benefit of

$$b_i(k)$$

to the central NCO-designer for each pair of vertices in the i-layer that are k hops away from each other in the i-layer of NCO, where $b_i(\cdot)$ is a decreasing nonnegative function with $b_i(1) = 0$ (i.e., *vertices in the i-layer that are further away provide smaller benefits*). Let

$$\mathbb{N}_i \ \ (\text{with } |\mathbb{N}_i| = n_i)$$

denote the set of NCO-vertices in the i-layer and let

$$\mathfrak{G}^{n_i}$$

be the set of all NCO-graphs on n_i nodes. The outcome of the NCO-formation process in the i-layer is a graph

$$G_i = (\mathbb{N}_i; \mathbb{E}_i) \in \mathfrak{G}^{n_i}.$$

An NCO-graph G_i in the i-layer has an associated (or **operational value function**) $u_i : \mathfrak{G}^{n_i} \to \mathbb{R}$ given by

$$u_i(G_i) = \sum_{a,b \in \mathbb{N}: i \neq j} b_i(d_{G_i}(a,b)) - \mathfrak{c}_i |\mathbb{E}_i|, \tag{4}$$

where

$$\mathfrak{c}_i > 0$$

is a uniform cost for each edge in the i-layer of NCO.

With this formulation, there is an inherent trade-off faced by the NCO-designer: *adding edges to a larger number of the i-layer's NCO-vertices yields a larger benefit* (by reducing the distances between *i-layer's* NCO-vertices), but *also a larger cost invested in edges*. To this end, we may introduce the following definition.

Definition 24. An NCO

$$\mathscr{G} = (\mathscr{G}_1, \mathscr{G}_2, \mathscr{G}_3, \mathscr{G}_4, \mathscr{G}_5) \in \prod_{n_1,n_2,n_3,n_4,n_5} (\mathfrak{G}^{n_1} \times \mathfrak{G}^{n_2} \times \mathfrak{G}^{n_3} \times \mathfrak{G}^{n_4} \times \mathfrak{G}^{n_5})$$

is **efficient with respect to the operational product utility**

$$(u_1, u_2, u_3, u_4, u_5),$$

if it has the highest operational utility in all of its layers separately. In other words, if

$$u_i(\mathscr{G}_i) \geq u_i(G_i), \quad \forall G_i = (\mathbb{N}_i; \mathbb{E}_i) \in \mathfrak{G}^{n_i}$$

whenever $n_i = 0, 1, 2, \ldots$ and $i = 1, 2, 3, 4, 5$.

A representative result in this setting is that there are only a few different kinds of efficient NCO, depending on the relative values of the *i*-layer edge costs and connection benefits: the empty NCO (for high edge costs), the star NCO (for medium edge costs), and the fully connected NCO (for low edge costs) (see [14] and [12]).

6.2 *Two-Layer Distance-Based Operational Utilities: Best Response NCO-Graphs*

We will now design operationally compatible NCO-layer graphs. To this end we will investigate the distance-based NCO-formation according to Shahrivar-Sundaram's multi-layer setting in [26]. Specifically, we will suppose $G_i = (\mathbb{N}_i; \mathbb{E}_i)$ is a given graph in the *i*-layer of an NCO, where the edge set $\mathbb{E}_i$ specifies a type of relationship between the vertices in $\mathbb{N}_i$. In order to simplify the situation, we will assume that *any other j-layer graph* $G_j = (\mathbb{N}_j; \mathbb{E}_j)$ *has the same number of vertices with the given graph* $G_i = (\mathbb{N}_i; \mathbb{E}_i)$, i.e., $|\mathbb{N}_i| = \left|\mathbb{N}_j\right|$. This allows us to *identify the vertices of the new graphs with the vertices of the given one*:

$$\mathbb{N} := \mathbb{N}_i \equiv \mathbb{N}_j, \text{ whenever } j = 1, 2, 3, 4, 5.$$

In particular, *the set of all the edges* $(a, b)_i \in \mathbb{E}_i$ *of the given i-layer graph* $G_i = (\mathbb{N}; \mathbb{E}_i)$ *can be imbedded into the set of all the edges* $(a, b)_j \in \mathbb{E}_j$ *of any other j-layer graph* $G_j = (\mathbb{N}; \mathbb{E}_j)$. Following this, our objective consists in designing another *j*-layer graph $G_j = (\mathbb{N}; \mathbb{E}_j)$ on the same set of vertices $\mathbb{N}$, with operational utility

$$u_j\left(G_j/G_i\right) = \sum_{(a,b)\in\mathbb{E}_i} b_j\left(d_{G_j}(a,b)\right) - \mathfrak{c}_j\left|\mathbb{E}_j\right|. \tag{5}$$

Here $d_{G_j}(a, b)$ denotes the shortest path between vertices a and b in graph G_j, when the edge $(a, b) \in \mathbb{E}_i$ is viewed as an edge lying in $\mathbb{E}_j$.

This operational utility function captures the idea that only distances between certain pairs of vertices (specified by the edge set $\mathbb{E}_i$) matter in the new graph G_j, as opposed to the distances between all pairs of vertices, as in the traditional distance based network formation model described in Eq. (4). Indeed the traditional distance-based operational utility function in (4) is obtained when G_i is the complete graph.

Optimal design of $\mathbb{E}_j$ with respect to $\mathbb{E}_i$ is a maximization problem for the operational utility function in relation (5).

Definition 25. In an NCO, if $G_j^* = \left(\mathbb{N}; \mathbb{E}_j^*\right)$ is the *j*-layer that maximizes (5), then G_j^* is called the **best response *j*-layer graph** to G_i or, equivalently, **the efficient *j*-layer graph** with respect to the operational utility function (5).

Remark 9. It is justifiable to speculate that *the best response j-layer graph* G_j^* *relatively to the layer graph* G_i *is always a subgraph of* G_i. This is trivially true when G_i is the complete graph, but, in general, the conjecture fails [26].

We will now characterize certain properties of best response layer graphs. We start with the following useful result.

Lemma 1 ([26]). *In an NCO, if* G_j^* *is the best response j-layer graph to* G_i*, then the number of edges in* G_j^* *is less than or equal to the number of edges in* G_i*, with equality if and only if* $G_j^* = G_i$.

The next lemma provides the best response layer graphs relatively to the subgraphs of G_i.

Lemma 2 ([26]). *In an NCO, suppose* G_j^* *is the best response j-layer graph to* G_i *and* G_j^* *is not connected. Let* $G_{j_k}^* = \left(\mathbb{N}_k,\ \mathbb{E}_{j_k}^*\right)$, $k = 1, \ldots, K$, *be the components of* G_j^*. *Let* $G_{i_k} = (\mathbb{N}_k, \mathbb{E}_{i_k})$, $k = 1, \ldots, K$, *be the subgraphs induced by vertex sets* $\mathbb{N}_k$ *on* G_i. *Then* $G_{j_k}^*$ *is the best response* j_k*-layer graph to* G_{i_k} *for all* $k = 1, \ldots, K$.

In general, the solution to the optimization problem (5) depends on both G_i and the relative sizes of the operational utility function and edge cost. In the following, we characterize the best response NCO-layer graph in certain cases.

Proposition 2 ([26]). *In an NCO, suppose the graph* $G_i = (\mathbb{N}; \mathbb{E}_i)$ *is connected. If*

$$b_j(1) > \mathfrak{c}_j,$$

then the best response j-layer graph to G_i *is also connected.*

Proof. Assume the best response j-layer graph $G_j^* = \left(\mathbb{N}; \mathbb{E}_j^*\right)$ is *not* connected. Then $\exists (a, b) \in \mathbb{E}_i$ such that $d_{G_j^*}(a, b) = \infty$. For $G_j^{*\prime} = \left(\mathbb{N}; \mathbb{E}_j^{*\prime}\right)$ with $\mathbb{E}_j^{*\prime} = \mathbb{E}_j^* \cup \{(a, b)\}$, $u_j\left(G_j^{*\prime}/G_i\right) - u_j\left(G_j^*/G_i\right) \geq b_j(1) - \mathfrak{c}_j > 0$. This contradicts the assumption that G_j^* is the best response j-layer graph to G_i.

The next two propositions describe the best response network with respect to a general network in two specific cases (see [26]).

Proposition 3. *In an NCO, suppose* $G_i = (\mathbb{N}; \mathbb{E}_i)$ *be an arbitrary graph. If*

$$b_j(1) - \mathfrak{c}_j > b_j(2),$$

then the best response j-layer graph to G_i *is* $G_j^* = G_i$.

Proof. Suppose that $G_j^* = \left(\mathbb{N}; \mathbb{E}_j^*\right)$ is the best response j-layer graph and $G_j^* \neq G_i$. By Lemma 1, we know that the number of edges in G_j^* is less G_i. So, there are two vertices a and b such that $(a, b) \in \mathbb{E}_i$ and $d_{G_j^*}(a, b) > 1$. Adding the edge (a, b)

to $\mathbb{E}_j^*$ increases the utility by at least $b_j(1) - \mathfrak{c}_j - b_j(2) > 0$ which contradicts the assumption that $G_j^* \neq G_i$ is the best response j-layer graph. Therefore, the best response j-layer graph must be equal to G_i.

Proposition 4. *In an NCO, if*

$$b_j(1) < \mathfrak{c}_j,$$

then the best response j-layer graph to the i-layer graph $G_i = (\mathbb{N}; \mathbb{E}_i)$ is not G_i, unless G_i is empty.

Proof. If $G_j^* = G_i \neq \emptyset$, then $u_j\left(G_j^*/G_i\right) = |\mathbb{E}_i|\left(b_j(1) - \mathfrak{c}_j\right) < 0$ due to the assumption that $b_j(1) < \mathfrak{c}_j$. Thus it must be the case that $G_j^* \neq G_i$, or G_i is the empty network.

The above results lead to the following complete characterizations of the best response i-layer graph when G_i is a tree or a forest.

Corollary 2. *Suppose $G_i = (\mathbb{N}; \mathbb{E}_i)$ is a tree in an NCO. If*

$$b_j(1) > \mathfrak{c}_j,$$

the best response j-layer graph $G_j^* = \left(\mathbb{N}; \mathbb{E}_j^*\right)$ with respect to G_i is $G_j^* = G_i$. Otherwise if

$$b_j(1) < \mathfrak{c}_j,$$

G_j^* is the empty graph.

Proof. For $b_j(1) > \mathfrak{c}_j$, by Proposition 1, G_j^* must be a connected graph, and thus has at least $|\mathbb{N}| - 1$ edges. By Lemma 1 and the fact that G_i is a tree and has $|\mathbb{N}| - 1$ edges [13], G_j^* has exactly $|\mathbb{N}| - 1$ edges, and thus $G_j^* = G_i$. If $b_j(1) < \mathfrak{c}_j$, then, by Proposition 3, the best response j-layer graph G_j^* is not equal to G_i. By Lemma 1, G_j^* cannot be a connected graph and thus has components $G_{j_k}^* = \left(\mathbb{N}_k, \mathbb{E}_{j_k}^*\right)$. Denote the induced subgraphs of G_i on vertex sets $\mathbb{N}_k$ by $G_{i_k} = (\mathbb{N}_k, \mathbb{E}_{i_k})$. By Lemma 2, $G_{j_k}^*$ is the best response j_k-layer graph to G_{i_k}. If G_j^* is not the empty graph, $|\mathbb{N}_k| > 2$ for some component $G_{j_k}^*$. Since G_i is a tree, G_{i_k} has $|\mathbb{N}_k| - 1$ or fewer edges. Thus, by Lemma 1 and Proposition 3, $G_{j_k}^*$ must have fewer than $|\mathbb{N}_k| - 1$ edges. This contradicts the fact that $G_{j_k}^*$ is a component, and thus G_j^* must be the empty graph whenever $b_j(1) < \mathfrak{c}_j$.

Corollary 3. *Let G_1 be a forest in an NCO. Then $G_j^* = G_i$ is the best response j-layer graph to G_i if*

$$b_j(1) > \mathfrak{c}_j.$$

Otherwise, if

$$b_j(1) < \mathfrak{c}_j,$$

the empty network is the best response to G_i.

Proof. Denote the induced subgraphs of G_i on vertex sets $\mathbb{N}_k$ by $G_{i_k} = (\mathbb{N}_k, \mathbb{E}_{i_k})$ for $k = 1, \ldots, \mathscr{K}$, where $K \geq 2$. By Lemma 1, G_j^* is a disconnected graph. It suffices to show that the induced subgraphs of G_j^* are the same as the induced subgraphs of G_i if $b_j(1) > \mathfrak{c}_j$. Denote the induced subgraphs of G_j^* on vertex sets $\mathbb{N}_\ell$ by $G_{j_\ell}^* = \left(\mathbb{N}_\ell, \mathbb{E}_{j_\ell}^*\right)$ for $\ell = 1, \ldots, \mathscr{L}$. Suppose that there exists a $\mathbb{N}_\ell$ consisting of vertices from multiple $\mathbb{N}_k$. Denote the induced subgraph of G_i on $\mathbb{N}_\ell$ by $\widehat{G_{i_k}}$. From Lemma 2, it follows that $G_{j_k}^*$ is the best response j_k-layer graph to $\widehat{G_{i_k}}$. But we know that $\widehat{G_{i_k}}$ is a forest with more than one tree and, consequently, $G_{j_k}^*$ is not a connected graph, contradicting our assumption that $G_{j_k}^*$ is a component of G_j^*. Next, we show that $\mathbb{N}_\ell$ cannot be a strict subset of $\mathbb{N}_k$. By way of contradiction, suppose that $\mathbb{N}_\ell$ is a strict subset of $\mathbb{N}_k$. Then there must exist an $N_{\ell'}$ that is also a strict subset of $\mathbb{N}_k$, such that there is an edge in $\mathbb{E}_i$ between some vertex in $\mathbb{N}_\ell$ and some vertex in $N_{\ell'}$ (since the graph G_{i_k} is connected). Since $b_j(1) > \mathfrak{c}_j$, as in the proof of the Proposition 1, adding this edge to G_j^* increases the operational utility $u_j\left(G_j^*/G_i\right)$, contradicting the assumption that G_j^* is the best response j-layer graph. Thus the vertex sets of the components of G_j^* are the same as the vertex sets of the components of G_i. By Lemma 2 and Corollary 1, each component of G_j^* is equal to the corresponding component of G_i, and thus $G_j^* = G_i$.

If $b_j(1) < \mathfrak{c}_j$, the same argument as in Corollary 1 shows that each $\mathbb{N}_\ell$ is a single vertex and as a result, the best response j-layer graph to G_i is the empty graph, which proves the claim. In both of the above cases, G_j^* is equal to the union of the best response NCO-layer graphs to each of the G_{i_k}.

6.3 *Pairwise Operational Stability in NCO*

In the previous section, we assumed that a centralized NCO-designer chooses the best response NCO-layer graph to a given NCO-layer graph. In this section, we study the satisfaction of the individual vertices in the network with the decision of the central NCO-designer. Specifically, let $G_i = (\mathbb{N}; \mathbb{E}_i)$ be a given NCO-layer graph, and let $\mathscr{U}_{j/i}$ denote the set of all possible operational utility functions $u_j\left(G_j/G_i\right)$ for the NCO-layer graph $G_j = (\mathbb{N}; \mathbb{E}_j)$ based on the j-layer graph G_i. As in the previous section we suppose $\mathbb{N} := \mathbb{N}_i \equiv \mathbb{N}_j$. Let $n = |\mathbb{N}|$. For each vertex $a \in \mathbb{N}$, define the allocation rule

$$\mathfrak{A}_{a,j}\left(G_j,\ G_i,\ u_j\right) : \mathfrak{G}^n \times \mathfrak{G}^n \times \mathscr{U}_{j/i} \to \mathbb{R}$$

specifying the amount of operational utility that is allocated to the a-vertex from the overall operational utility generated by the formed NCO-layer graph G_j. For simplicity, we will use the notation $\mathfrak{A}_a\left(G_j\right)$ when G_i and $\mathscr{U}_{j/i}$ are fixed.

For a given best response j-layer graph G_j and individual operational utility functions $\mathfrak{A}_a$, it may be the case that a certain vertex can improve its own operational utility by removing one or more of its incident edges in G_j, or by adding additional edges from itself to other vertices.

As in [8], it is assumed that any vertex can remove any of its incident edges unilaterally, but that adding an edge to another vertex requires the consent of that vertex. This motivates the following definition of pairwise stability of a given NCO [14]. In this definition, when $[a,b] \notin G$, $G+[a,b]$ denotes the NCO-layer graph obtained by adding an edge between a and b in G. Similarly, $G-[a,b]$ represents the NCO-layer graph obtained by deleting the edge $[a,b]$ when $[a,b] \in G$.

Definition 26 ([14]). An NCO-layer graph $G = (\mathbb{N};\mathbb{E})$ is said to be **pairwise stable** if

$$\forall\, [a,b] \in \mathbb{E},\ \mathfrak{A}_a\left(G\right) \geq \mathfrak{A}_a\left(G-[a,b]\right) \text{ and } \mathfrak{A}_a\left(G\right) \geq \mathfrak{A}_a\left(G-[a,b]\right),$$

and

$$\forall\, [a,b] \notin \mathbb{E},\ \text{if } \mathfrak{A}_a\left(G+[a,b]\right) > \mathfrak{A}_a\left(G\right) \text{ then } \mathfrak{A}_a\left(G+[a,b]\right) < \mathfrak{A}_a\left(G\right).$$

The graph is **pairwise unstable** if it is not pairwise stable.

In words, pairwise stability of an NCO-layer graph corresponds to the situation where no vertex has any incentive to change any (one) of its connections in the NCO-layer graph. This is a modification of the notion of a Nash equilibrium in network formation, capturing the concept of negotiation and agreement between the endpoints prior to forming the edge. Various versions of this notion have been studied in the network formation literature (see, for example, [4, 12, 13]).

We now investigate the pairwise stability properties of the best response NCO-layer graphs. Consider the allocation rule

$$\mathfrak{A}_a\left(G_j/G_i\right) = \frac{1}{2}\sum_{[a,b]\in\mathbb{E}_i} b_j\left(d_{G_j}\left(a,b\right)\right) - \frac{\mathfrak{c}_j}{2}\deg_a\left(G_j\right), \tag{6}$$

where $\deg_a\left(G_j\right)$ is the degree of vertex i in the NCO-layer graph G_j. Note that the total utility in (5) satisfies

$$u_j\left(G_j/G_i\right) = \sum_{a\in\mathbb{N}} \mathfrak{A}_a\left(G_j/G_i\right).$$

It is not hard to show that for any $a,b \in \mathbb{N}$ where $G_j = (\mathbb{N};\mathbb{E}_j)$, if $[a,b] \notin \mathbb{E}_j$, then it would not be beneficial for at least one of the vertices a or b to add the edge

(a, b) to the NCO-layer graph G_j. By way of contradiction assume that

$$\mathfrak{A}_a\left(G_j + [a, b]\right) \geq \mathfrak{A}_a\left(G_j\right) \text{ and } \mathfrak{A}_b\left(G_j + [a, b]\right) \geq \mathfrak{A}_b\left(G_j\right)$$

with one of the inequalities strict. Then,

$$\mathfrak{A}_a\left(G_j + [a, b]\right) + \mathfrak{A}_b\left(G_j + [a, b]\right) > \mathfrak{A}_a\left(G_j\right) + \mathfrak{A}_b\left(G_j\right).$$

However, this means that $u_j\left(G_j + [a, b]/G_i\right) > u_j\left(G_j/G_i\right)$ which contradicts the assumption that G_j is the optimal NCO-layer graph. This immediately implies that if the empty NCO-layer graph is the best response to an NCO-layer graph, it is pairwise stable. For a general NCO-layer graph, to conclude that the best response NCO-layer graph G_2 is pairwise stable, we also need to show that removing any of the edges from NCO-layer graph G_2 is not beneficial for any of its endpoints. However, this is *not* true in general.

Proposition 5. *In an NCO, if the best response j-layer graph with respect to a i-layer graph G_i is $G_j = G_i$, then G_j is pairwise stable.*

Proof. As argued above, adding an edge is not beneficial to any vertex. Thus it suffices to show that removing any of the edges is unrewarding for both of its endpoints. Since $G_j = G_i$, edge $[a, b]$ is only useful for the connection between vertices a and b. Consequently, removing the edge $[a, b]$ increases the utility of G_j. This contradicts the fact that $G_j = G_i$ is the best response to G_i.

Remark 10. Note that the above result encompasses cases where G_i is an arbitrary NCO-layer graph and $b_j(1) - \mathfrak{c}_j > b_j(2)$ (by Proposition 2), and where G_i is a tree (by Corollary 3).

6.4 NCO-Formation with Arbitrary Operational Utility Functions

So far we have discussed the construction of one NCO-layer graph with respect to another layer graph of the same NCO based on the distance operational utility function, and how vertices of the NCO-layer graphs evaluate decisions made by the central NCO-designer. In this section, we consider the scenario where there is no central NCO-designer and vertices themselves establish multiple different types of relationships with other vertices over time; each type of relationship corresponds to a different layer (or edge set) on the set of N vertices. The operational utility of the vertices is a function of their status in both NCO-layer graphs G_1 and G_2. We will start by briefly reviewing the concept of an improving path [12] for single-layer network formation. Note that we are not assuming distance-based utility functions in this section and the analysis is applicable to any operational utility function.

A five layer NCO $G \in (\mathfrak{S}^n)^5$ is represented as $G = (G_1, G_2, G_3, G_4, G_5)$, where

$$G_i = (\mathbb{N}, \mathbb{E}_i)\,(n = |\mathbb{N}|)$$

is the graph of the i-layer. With this notation, G_1 is the graph of the Processes' layer, G_2 is the graph of the People's layer, G_3 is the graph of the Applications' layer, G_4 is the graph of the Systems' layer, and G_5 is the graph of the Physical Network's layer. The NCO-designer may decide about the connections in different layers based on the operational utility which is a function of connections in the NCO layers.

For any $s \in \{1, 2, 3, 4, 5\}$, let us consider the operational utility function

$$v\left(G_{q_1}, \ldots, G_{q_s}\right) : (\mathfrak{S}^n)^s \equiv \underbrace{\mathfrak{S}^n \times \cdots \times \mathfrak{S}^n}_{s-\text{times}} \to \mathbb{R}$$

that evaluates $\left(G_{q_1}, \ldots, G_{q_s}\right)$ and assigns to it a (positive) real number.

The allocation rule in $(\mathfrak{S}^n)^s$ at a vertex $a \in \mathbb{N}$ will be denoted by

$$\mathfrak{Y}_a\left(G_{q_1}, \ldots, G_{q_s}; v\right) : (\mathfrak{S}^n)^s \times \mathscr{U} \equiv \underbrace{\mathfrak{S}^n \times \cdots \times \mathfrak{S}^n}_{s-\text{times}} \times \mathscr{U} \to \mathbb{R}.$$

For simplicity, we will omit the argument v and simply denote it as

$$\mathfrak{Y}_a\left(G_{q_1}, \ldots, G_{q_s}\right).$$

Definition 27. Let $s \in \{1, 2, 3, 4, 5\}$. In an NCO, the two subsets of s-layer graphs $\left(G_{q_1}, \ldots, G_{q_s}\right)$ and $\left(G'_{q_1}, \ldots, G'_{q_s}\right)$ are said to be **adjacent** if one could reach the new set of s NCO-graphs $\left(G''_{q_1}, \ldots, G''_{q_s}\right)$ with at most one change in one of the graphs $G_{q_1}, \ldots, G_{q_s}$. In other words,

$$G'_{q_r} = G_{q_r} \pm [a, b] \text{ for only one } r \in \{1, \ldots, s\}.$$

We also use the notation

$$\left(G'_{q_1}, \ldots, G'_{q_s}\right) = \left(G_{q_1}, \ldots, G_{q_s}\right) \pm [a, b]_i \text{ for} i \in \{1, 2, 3, 4, 5\},$$

depending on the i-layer from which we add or remove the edge $[a, b]$.

Definition 28. Let $s \in \{1, 2, 3, 4, 5\}$. In an NCO, the set of s layer graphs $\left(G'_{q_1}, \ldots, G'_{q_s}\right)$ **defeats** the set of s layer graphs $\left(G_{q_1}, \ldots, G_{q_s}\right)$ if they are adjacent and one of the following conditions holds.

1. $\left(G'_{q_1}, \ldots, G'_{q_s}\right) = \left(G_{q_1}, \ldots, G_{q_s}\right) + [a, b]_i$ *for some* $i \in \{1, 2, 3, 4, 5\}$, *with* $\mathfrak{Y}_a\left(G'_{q_1}, \ldots, G'_{q_s}\right) \geq \mathfrak{Y}_a\left(G_{q_1}, \ldots, G_{q_s}\right)$ *and* $\mathfrak{Y}_a\left(G'_{q_1}, \ldots, G'_{q_s}\right) \geq \mathfrak{Y}_a\left(G_{q_1}, \ldots, G_{q_s}\right)$,
   where at least one of the above inequalities is strict.
2. $\left(G'_{q_1}, \ldots, G'_{q_s}\right) = \left(G_{q_1}, \ldots, G_{q_s}\right) - [a, b]_i$ *for some* $i \in \{1, 2, 3, 4, 5\}$, *with*

$\mathfrak{Y}_a\left(G'_{q_1},\ldots,G'_{q_s}\right) > \mathfrak{Y}_a\left(G_{q_1},\ldots,G_{q_s}\right)$ *and* $\mathfrak{Y}_a\left(G'_{q_1},\ldots,G'_{q_s}\right) > \mathfrak{Y}_a\left(G_{q_1},\ldots,G_{q_s}\right)$.

If the set of s layer graphs $\left(G_{q_1},\ldots,G_{q_s}\right)$ is not defeated by any other adjacent set of five NCO-graphs, we say it is **intra-layer pairwise stable**.

Let us now consider a directed graph Γ^s with the *s layer graphs* $\left(G_{q_1},\ldots,G_{q_s}\right) \in (\mathfrak{S}^n)^s$ as its vertices. There is an edge from vertex $\left(G_{q_1},\ldots,G_{q_s}\right)$ to vertex $\left(G'_{q_1},\ldots,G'_{q_s}\right)$, if they are adjacent (i.e., *if one could reach the new set of s layer graphs* $\left(G''_{q_1},\ldots,G''_{q_s}\right)$ *with at most one change in one of the graphs* $G_{q_1},\ldots,G_{q_s}$) and the former set of five NCO-graphs is defeated by the latter.

Definition 29. In an NCO, an **improving path** in Γ^s is a chain of sets of s layer graphs

$$\left(G_{q_1}^{(1)},\ldots,G_{q_s}^{(1)}\right),\ldots,\left(G_{q_1}^{(k)},\ldots,G_{q_s}^{(k)}\right),$$

where there is an edge from vertex $\left(G_{q_1}^{(i)},\ldots,G_{q_s}^{(i)}\right)$ to vertex $\left(G_{q_1}^{(i+1)},\ldots,G_{q_s}^{(i+1)}\right)$.

We have the following general result.

Lemma 3 ([13]). *In an NCO, for any value function v and allocation rule* $\mathfrak{Y}_a$*, there exists at least one pairwise stable layer graph or a closed set of layer graphs.*

Applying Lemma 3 we infer

Proposition 6. *In an NCO, for any v and* $\mathfrak{Y}_a$ *there exists at least one intra-layer pairwise stable graph or a closed set of s layer graphs in* Γ^s.

The following lemma relates intra-layer pairwise stable networks to stability properties of the individual layers when the utility function has a special form. We will use the following definition.

Definition 30 ([26]). A function $f : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is said to be **increasing in its arguments (IA)** if and only if $\forall x,y,z,u\in\mathbb{R}$,

1. $f(x,y) > f(z,y) \Longleftrightarrow x > z$.
2. $f(x,y) > f(x,u) \Longleftrightarrow y > u$.

Proposition 7. *In an NCO, suppose that the operational utility function of each vertex a has the form*

$$\mathfrak{Y}_a\left(G_{q_1},\ldots,G_{q_s}\right) = f\left(\mathfrak{Y}_a^{(q_1)},\ldots,\mathfrak{Y}_a^{(q_s)}\right)$$

where

1. $\mathfrak{Y}_a^{(q_r)} : \mathfrak{S}^n\times\mathscr{U}\to\mathbb{R}$ *is the operational utility function of vertex a in layer* q_r *and*
2. f *is an IA function.*

Then the following statements are equivalent.

1. *The layer graphs* $G^\star_{q_1} = \left(\mathbb{N},\mathbb{E}^\star_{q_1}\right),\ldots,G^\star_{q_s} = \left(\mathbb{N},\mathbb{E}^\star_{q_s}\right)$ *are pairwise stable.*

2. $\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right)$ *is intra-layer pairwise stable.*

Proof. Assume that statement 1 is true. By way of contradiction suppose that $\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right)$ is not intra-layer pairwise stable. Then one of the following cases must happen.

- $\exists\, [a,b] \in \mathbb{E}^{\star}_{q_r}$ with $r \in \{1, \ , \ldots, s\}$ such that

$$\mathfrak{Y}_a\left(\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right) - [a,b]_{q_r}\right) > \mathfrak{Y}_a\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right).$$

  Since we are making a change in only one of the layers and the amount of utility that vertices receive from each of their layers is a function of the edge set in only that layer and f is an increasing in its arguments function, we can conclude that

$$\mathfrak{Y}^{(q_r)}_a\left(G^{\star}_{q_r} - [a,b]\right) > \mathfrak{Y}^{(q_r)}_a\left(G^{\star}_{q_r}\right).$$

  However, this is impossible due to the assumption that $G^{*}_{q_r}$ is pairwise stable.
- $\exists\, [a,b] \notin \mathbb{E}^{\star}_{q_r}$ such that

$$\mathfrak{Y}_a\left(\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right) + [a,b]_{q_k}\right) \geq \mathfrak{Y}_a\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right)$$

  and

$$\mathfrak{Y}_a\left(\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right) + [a,b]_{q_\ell}\right) \geq \mathfrak{Y}_a\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right)$$

  with one of the inequalities strict. Again since the change is in only one of the layers and using the IA property, we can conclude that

$$\mathfrak{Y}^{(q_k)}_a\left(G^{\star}_{q_k} + [a,b]\right) > \mathfrak{Y}^{(q_k)}_a\left(G^{\star}_{q_k}\right) \text{ and } \mathfrak{Y}^{(q_k)}_b\left(G^{\star}_{q_k} + [a,b]\right) > \mathfrak{Y}^{(q_k)}_b\left(G^{\star}_{q_k}\right)$$

  with one of the inequalities strict. However, this contradicts the assumption that $G^{\star}_{q_k}$ is pairwise stable.

  Next we show that if statement 2 is true, statement 1 is also true. Again by way of contradiction suppose that $\left(G^{\star}_{q_1}, \ldots, G^{\star}_{q_s}\right)$ is intra-layer pairwise stable but $G^{\star}_{q_1}$ is not pairwise stable. Then one of the following cases must happen.
- $\exists\, [a,b] \in \mathbb{E}^{\star}_{q_1}$ such that

$$\mathfrak{Y}^{(q_1)}_a\left(G^{\star}_{q_1} - [a,b]\right) > \mathfrak{Y}^{(q_1)}_a\left(G^{\star}_{q_1}\right).$$

  Then, based on the IA of f, we must have

$$f\left(\mathfrak{Y}^{(q_1)}_a\left(G^{\star}_{q_1} - [a,b]\right), \ \ldots, \mathfrak{Y}^{(q_k)}_a\left(G^{\star}_{q_k}\right)\right) > f\left(\mathfrak{Y}^{(q_1)}_a\left(G^{\star}_{q_1}\right), \ \ldots, \mathfrak{Y}^{(q_k)}_a\left(G^{\star}_{q_k}\right)\right)$$

$$\Rightarrow$$

$$\mathfrak{Y}_a\left(\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)-[a,b]_{q_1}\right)>\mathfrak{Y}_a\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)$$

which contradicts the intra-layer pairwise stability assumption of $\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)$.

- $\exists\,[a,b]\notin\mathbb{E}^\star_{q_1}$ such that

$$\mathfrak{Y}^{(q_1)}_a\left(G^\star_{q_1}+[a,b]\right)\geq\mathfrak{Y}^{(q_1)}_a\left(G^\star_{q_1}\right).$$

and

$$\mathfrak{Y}^{(q_1)}_b\left(G^\star_{q_1}+[a,b]\right)>\mathfrak{Y}^{(q_1)}_b\left(G^\star_{q_1}\right).$$

with one of the inequalities strict. Then, using the IA property, we can conclude that

$$\mathfrak{Y}^{(q_1)}_a\left(\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)+[a,b]_{q_1}\right)\geq\mathfrak{Y}^{(q_1)}_a\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)$$

and

$$\mathfrak{Y}^{(q_1)}_b\left(\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)+[a,b]_{q_1}\right)\geq\mathfrak{Y}^{(q_1)}_b\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)$$

with one of the inequalities strict, which again contradicts the intra-layer pairwise stability assumption of $\left(G^\star_{q_1},\ldots,G^\star_{q_s}\right)$.
The same holds for $G^\star_{q_2},\ldots,G^\star_{q_s}$ and therefore the proof is complete.

References

1. Aigner, M., Fromme, M.: A game of cops and robbers. Discrete Appl. Math. **8**, 1–12 (1984)
2. Alberts, D.S., Garstka, J.J., Stein, F.P.: Network Centric Warfare: Developing and Leveraging Information Superiority, 2nd edn. (Revised). CCRP Publication Series, Washington (2002)
3. Alspach, B.: Sweeping and searching in graphs: a brief survey. Matematiche **59**, 5–37 (2006)
4. Blume, L., Easley, D., Kleinberg, J., Kleinberg, R., Tardos, E.: Network formation in the presence of contagious risk. In: Proceedings of the 12th ACM Conference on Electronic Commerce, pp. 1–10 (2011)
5. Bonato, A.: A Course on the Web Graph. American Mathematical Society Graduate Studies Series in Mathematics. American Mathematical Society, Providence (2008)
6. Bonato, A., Hahn, G., Wang, C.: The cop density of a graph. Contrib. Discret. Math. **2**, 133–144 (2007)
7. Bonato, A., PraΓat, P., Wang, C.: Pursuit-evasion in models of complex networks. Internet Math. **4**(4), 299–436 (2007)
8. Bonato, A., PraΓlat, P., Wang, C.: Vertex pursuit games in stochastic network models. In: Proceedings of the 4th Workshop on Combinatorial and Algorithmic Aspects of Networking (2007)
9. Borgatti, S.P., Everett M.G., Freeman, L.C.: UCINET user's manual. Analytic Technologies, Harvard (2006)
10. Chung, F.R.K., Lu, L.: Complex Graphs and Networks. American Mathematical Society, Providence (2004)

11. Chung, F.R.K., Lu, L.: The average distance in a random graph with given expected degrees. Internet Math. **1**, 91–114 (2006)
12. Jackson, M.O.: Social and Economic Networks. Princeton University Press, Princeton (2008)
13. Jackson, M.O., Watts, A.: The evolution of social and economic networks. J. Econ. Theory **106**, 265–295 (2002)
14. Jackson, M.O., Wolinsky, A.: A strategic model of social and economic networks. J. Econ. Theory **71**(0108), 44–74 (1996)
15. Fomin, F.V., Thilikos, D.: An annotated bibliography on guaranteed graph searching. Theor. Comput. Sci. **399**, 236–245 (2008)
16. Gao, J., Buldyrev, S.V., Stanley, H.E., Havlin, S.: Networks formed from interdependent networks. Nat. Phys. **8**, 40–48 (2012)
17. Goyal S., Vigier A.: Robust networks. Working Paper (2011)
18. Hahn, G.: Cops, robbers and graphs. Tatra Mt. Math. Publ. **36**, 163–176 (2007)
19. Hanneman R.A., Riddle, M.: Introduction to Social Network Methods. University of California, Riverside. http://faculty.ucr.edu/~hanneman/ (2005). Accessed 9 July 2006
20. Kearns, M., Judd, S., Vorobeychik, Y.: Behavioral experiments on a network formation game. In: Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 690–704 (2012)
21. Network Centric Operations Conceptual Framework. Version 1.0, November 2003, http://www.dtic.mil/dtic/tr/fulltext/u2/a457620.pdf
22. Neufeld, S., Nowakowski, R.: A game of cops and robbers played on products of graphs. Discret. Math. **186**, 253–268 (1998)
23. Nowakowski, R., Winkler, P.: Vertex to vertex pursuit in a graph. Discret. Math. **43**, 230–239 (1983)
24. Quilliot, A.: Jeux et pointes fixes sur les graphes. Ph.D. Dissertation, Universit*ι* de Paris VI (1978)
25. Schneider, C.M., Araujo, N.A.M., Havlin, S., Herrmann, H.J.: Toward designing robust coupled networks, arXiv:1106.3234 [condmat. stat-mech] (2011)
26. Shahrivar, E.M., Sundaram, S.: Strategic multi-layer network formation. In: 52nd IEEE Conference on Decision and Control, Florence, pp. 582–587, 10–13 December 2013
27. West, D.B.: Introduction to Graph Theory. Prentice Hall, Upper Saddle River (2001)
28. Wilson, C.: Network centric warfare: background and oversight issues for Congress. Congressional Research Service, The Library of Congress. http://www.fas.org/sgp/crs/natsec/RL32411.pdf (2007)
29. Wong-Jiru, A.: Major, USAF, graph theoretical analysis of network centric operations using multi-layer models. Thesis, Air Force Institute of Technology, Department of the Air Force, Air University, Wright-Patterson Air Force Base, Ohio (2006)

A Bio-Inspired Hybrid Artificial Intelligence Framework for Cyber Security

Konstantinos Demertzis and Lazaros Iliadis

Abstract Confidentiality, Integrity, and Availability of Military information is a crucial and critical factor for a country's national security. The security of military information systems (MIS) and Networks (MNET) is a subject of continuous research and design, due to the fact that they manage, store, manipulate, and distribute the information. This study presents a bio-inspired hybrid artificial intelligence framework for cyber security (bioHAIFCS). This framework combines timely and bio-inspired Machine Learning methods suitable for the protection of critical network applications, namely military information systems, applications and networks. More specifically, it combines (a) the hybrid evolving spiking anomaly detection model (HESADM), which is used in order to prevent in time and accurately, cyber-attacks, which cannot be avoided by using passive security measures, namely: Firewalls, (b) the evolving computational intelligence system for malware detection (ECISMD) that spots and isolates malwares located in packed executables untraceable by antivirus, and (c) the evolutionary prevention system from SQL injection (ePSSQLI) attacks, which early and smartly forecasts the attacks using SQL Injections methods.

1 Introduction

Application of high protection level measures by the army, in order to secure its information systems (IS), can offer a serious advantage in the evolution of a crisis, in tactical and operational level. It is a fact that the necessity to ensure secrecy of military IS and Confidentiality of information control and management systems (C4I) is a critical stabilization factor between opposite forces and a matter of honor for each side. The opposite can have serious consequences difficult to estimate in terms of material or moral cost. Thus, the development of network security systems following military specifications and demands is absolutely necessary. They could combine smart techniques capable of preventing attacks of zero-day nature.

K. Demertzis (✉) • L. Iliadis
Department of Forestry & Management of the Environment & Natural Resources, Democritus University of Thrace, Xanthi 671 00, Greece
e-mail: kdemertz@fmenr.duth.gr; liliadis@fmenr.duth.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_7

The most popular attack techniques aiming to gain access in important or sensitive data use one of the methods below:

- Direct invasion to the system with attacks of DoS,
- Dispersion and installation of malware software
- Exploitation of potential weaknesses in the security of the system and mainly in the security of the network applications with attacks of SQL Injections type.

In the case of direct attack in a network, the usual security measures are the installation of a Firewall, in order to prevent non-authorized access in certain services and the installation of an intrusion detection system (IDS). The IDS are network and event monitoring and analysis systems. The target is to spot indications of potential intrusion efforts or efforts aiming to deviate the security mechanisms by external non-authorized users or users with limited authorization. The protection in this case is based on passive measures that use statistical analysis of events. There are network based (NIDS) and host based (HIDS) IDSs. Some of them are looking for specific signatures of known threats, whereas others are spotting anomalies by comparing traffic patterns against a baseline [1].

There are three basic approaches for designing and building IDS, namely: the Statistical, the Knowledge based, and the Machine Learning one which has been employed in this research effort. The concept of the statistical-based systems (SBID) is simple: it determines "normal" network activity and then all traffic that falls outside the scope of normal is flagged as anomalous (abnormal). These systems attempt to learn network traffic patterns on a particular network. This process of traffic analysis continues as long as the system is active, so, assuming network traffic patterns remain constant, the longer the system is on the network, the more accurate it becomes. The knowledge based intrusion detection systems (KBIDES) classify the data vectors based on a carefully designed Rule Set or they use models obtained from past experience in a heuristic mode. The Machine Learning approach automates the analysis of the data vectors, and they result in the implementation of systems that have the capacity to improve their performance as time passes.

This research effort aims in the development and application of an innovative hybrid evolving spiking anomaly detection model (HESADM) [2], which employs classification performed by evolving spiking neural networks (eSNN), in order to properly label a potential anomaly (PAN) in the net. On the other hand, it uses a multi-layer feed forward (MLFF) ANN to classify the exact type of the intrusion.

The second attack approach is the dispersion and installation of malwares which are untraceable by the usual antivirus systems. Malware is a kind of software used to disrupt computer operation, gather sensitive information, or gain access to private computer systems. To identify already known malware, existing commercial security applications search a computer's binary files for predefined signatures. However, obfuscated viruses use software packers to protect their internal code and data structures from detection. Antivirus scanners act like file filters, inspecting suspicious file loading and storing activities. Malicious programs with obfuscated content can bypass antivirus scanners. Eventually, they are unpacked and executed in the victim's system [3].

Code packing is the dominant technique used to obfuscate malicious code, to hinder an analyst's understanding of the malware's intent and to evade detection by Antivirus systems. Malware developers transform executable code into data, at a post-processing stage in the whole implementation cycle. This transformation uses static analysis and it may perform compression or encryption, hindering an analyst's understanding. At runtime, the data or hidden code is restored to its original executable form, through dynamic code generation using an associated restoration routine. Execution then resumes as normal to the original entry point, which marks the entry point of the original malware, before the code packing transformation is applied. Finally, execution becomes transparent, as both code packing and restoration have been performed. After the restoration of one packing, control may transfer another packed layer. The original entry point is derived from the last such layer [4].

Code packing provides compression and software protection of the intellectual properties contained in a program. It is not necessarily advantageous to flag all occurrences of code packing as indicative of malicious activity. It is advisable to determine if the packed contents are malicious, rather than identifying only the fact that unknown contents are packed. Unpacking is the process of stripping the packer layers off packed executables to restore the original contents in order to inspect and analyze the original executable signatures. Universal unpackers, introduce a high computational overhead, low convergence speed, and computational resource requirements. The processing time may vary from tens of seconds to several minutes per executable. This hinders virus detection significantly, since without a priori knowledge on the nature of the executables to be checked for malicious code all of them would need to be run through the unpacker. Scanning large collections of executables may take hours or days. This research effort aims in the development and application of an innovative, fast, and accurate evolving computational intelligence system for malware detection (ECISMD) [5] approach for the identification of packed executables and detection of malware by employing eSNN. A multilayer evolving classification function (ECF) model has been employed for malware detection, which is based on fuzzy clustering. Finally, an evolutionary genetic algorithm (GA) has been applied to optimize the ECF network and to perform feature extraction on the training and testing datasets. A main advantage of ECISMD is the fact that it reduces overhead and overall analysis time, by classifying packed or not packed executables.

The third way widely used to overcome the security measures by exploiting the gaps in the control systems is the SQL injections one. This approach tries to exploit vulnerabilities in the security of network applications. SQL injection is a code injection technique, used to attack data driven applications, in which malicious SQL statements area SQL injection attack consists of insertion or "injection" of a SQL query via the input data from the client to the application. A successful SQL injection exploit can read sensitive data from the database, modify database data (Insert/Update/Delete), execute administration operations on the database (such as shutdown the DBMS), recover the content of a given file present on the DBMS file system, and in some cases issue commands to the operating system. SQL injection

attacks are a type of injection attack, in which SQL commands are injected into data-plane input in order to effect the execution of predefined SQL commands [6]. This study proposes a bio-inspired Artificial Intelligence model named evolutionary prevention system from SQL injection (ePSSQLI) Attacks. It combines the use of MLFF ANN with optimization techniques of genetic algorithms (evolutionary optimization), in order to handle the potential intrusion attacks, based on SQL injection type.

2 Literature Review

Artificial Intelligence and data mining algorithms have been applied as intrusion detection methods in finding new intrusion patterns [7–10], such as clustering (unsupervised learning) [11–13] or classification (supervised learning) [14–17]. Also, a few hybrid techniques were proposed like Neural Networks with Genetic Algorithms [18] or Radial Based Function Neural Networks with Multilayer Perceptron [19, 20]. Besides, other very effective methods exist such as Sequential Detection [21], State Space [22], Spectral Methods [23], and combinations of those.

Dynamic unpacking approaches monitor the execution of a binary in order to extract its actual code. These methods execute the samples inside an isolated environment that can be deployed as a virtual machine or an emulator [24]. The execution is traced and stopped when certain events occur. Several dynamic unpackers use heuristics to determine the exact point where the execution jumps from the unpacking routine to the original code. Once this point is reached, the memory content is bulk to obtain an unpacked version of the malicious code. Other approaches for generic dynamic unpacking have been proposed that are not highly based on heuristics such as PolyUnpack [25] Renovo [26], OmniUnpack [27], or Eureka [28].

However, these methods are very tedious and time consuming, and cannot counter conditional execution of unpacking routines, a technique used for anti-debugging and anti-monitoring defense [29]. Another common approach is using the structural information of the executables to train supervised machine-learning classifiers to determine if the sample under analysis is packed or if it is suspicious of containing malicious code (e.g., PEMiner [30], PE-Probe [31], and Perdisci et al. [32]). These approaches that use this method for filtering, previous to dynamic unpacking, are computationally more expensive and time consuming and less effective to analyze large sets of mixed malicious and benign executables [33–35].

Artificial Intelligence and data mining algorithms have been applied as malicious detection methods and for the discovery of new malware patterns [36]. In the research effort of Babar and Khalid [29], boosted decision trees working on n-grams are found to produce better results than Naive Bayes classifiers and support vector machines (SVMs). Ye et al. [37] use automatic extraction of association rules on Windows API execution sequences to distinguish between malware and clean program files. Chandrasekaran et al. [38] used association rules, on honeytokens

of known parameters. Chouchan et al. [39] used Hidden Markov Models to detect whether a given program file is (or is not) a variant of a previous program file. Stamp et al. [40] employ profile hidden Markov Models, which have been previously used for sequence analysis in bioinformatics. Artificial Neural Networks (ANN) to detect polymorphic malware is explored in [41]. Yoo [42] employs Self-Organizing Maps to identify patterns of behavior for viruses in Windows executable files. These methods have low accuracy as a consequence, packed benign executables would likely cause false alarm, whereas malware may remain undetected.

Vulnerability pattern approach is used by Livshits et al. [43], they propose static analysis approach for finding the SQL injection attack. The main issue of this method is that it cannot detect the SQL injection attack patterns that are not known beforehand. Also, AMNESIA mechanism to prevent SQL injection at run time is proposed by Halfond et al. [44]. It uses a model based approach to detect illegal queries before it sends for execution to database. The mechanism which filters the SQL Injection in a static manner is proposed by Buehrer et al. [45]. The SQL statements by comparing the parse tree of a SQL statement before and after input and only allowing to SQL statements to execute if the parse trees match. Marco Cova et al. [46] proposed a Swaddler, which analyzes the internal state of a web application and learns the relationships between the application's critical execution points and the application's internal state.

There exists machine learning related works in the wild [47–51]. In this work we focus on the detection at the spot between application and database, detecting anomalous SQL statements (the SQL statement returns a result set of records from one or more tables), which are malicious in the sense that they include parts of injected code or differ from the set of queries usually issued within an application. Valeur et al. [52] proposed the use of an IDS based on a machine learning technique which identifies queries that do not match multiple models of typical queries at runtime, including string model and data type-independent model. It is trained by a set of typical application queries, and the quality depends on the quality of the training set. Wang et al. presented a novel method for learning SQL statements and apply machine learning techniques, such as one class classification, in order to detect malicious behavior between the database and application [53]. The approach incorporates the tree structure of SQL queries as well as input parameter and query value similarity as characteristic to distinguish malicious from benign queries. Rawat et al. use SVM for classification and prediction of SQL-Injection attack [54]. This work contains the idea that compares SQL query strings and blocks suspicious SQL-query and passes original SQL-query. Huang et al. present a new method to prevent SQLI attack based on machine learning [55]. This approach identifies SQL injection codes by HTTP parameters' attributes and the Bayesian classifier. This technique depends on the choices of patterns' attributes and the quality of the training set. They choose two values as attributes of patterns, and invent a way to generate the real-world patterns automatically. In addition Huang et al. designed a system based on machine learning for preventing SQL injection attack, which utilizes pattern classifiers to detect injection attacks and protect web applications [56]. The system captures parameters of HTTP requests, and converts them into

numeric attributes. Numeric attributes include the length of parameters and the number of keywords of parameters. Using these attributes, the system classifies the parameters by Bayesian classifier for judging whether parameters are injection patterns.

3 Methodologies Comprising the bioHAIFCS

The bioHAIFCS uses three biologically inspired Artificial Intelligence methods, namely: eSNN, MLFF, and ECF and their corresponding optimization approach with GA, in order to create a high level security framework. It acts in a smart and preemptive manner to spot the threats by making the minimum consumption of resources. These methods are presented below:

3.1 Evolving Spiking Neural Networks

The eSNN that has been developed and discussed herein is based on the *"Thorpe"* neural model [57] which intensifies the importance of the spikes taking place in an earlier moment, whereas the neural plasticity is used to monitor the learning algorithm by using one-pass learning. In order to classify real-valued data sets, each data sample is mapped into a sequence of *spikes* using the rank order population encoding (ROPE) technique [58, 59]. The topology of the developed eSNN is strictly feed-forward, organized in several layers and weight modification occurs on the connections between the neurons of the existing layers.

The details of eSNN architecture described below:

3.1.1 Rank Order Population Encoding

The ROPE method [58, 59] is an alternative to conventional rate coding scheme that uses the order of firing neuron's inputs to encode information which allows the mapping of vectors of real-valued elements into a sequence of spikes. Neurons organized into neuronal maps which share the same synaptic weights. Whenever the synaptic weight of a neuron is modified, the same modification is applied to the entire population of neurons within the map. Inhibition is also present between each neuronal map. If a neuron spikes, it inhibits all the neurons in the other maps with neighboring positions. This prevents all the neurons from learning the same pattern. When propagating new information, neuronal activity is initially reset to zero. Then, as the propagation goes on, neurons are progressively desensitized each time one of their inputs fires, thus making neuronal responses dependent upon the relative order of firing of the neuron's afferents. More precisely, let $A = \{a_1, a_2, a_3 \ldots a_{m-1}, a_m\}$ be

the ensemble of afferent neurons of neuron i and $W = \{w_{1,i}, w_{2,i}, w_{3,i} \ldots w_{m-1,i}, w_{m,i}\}$ the weights of the m corresponding connections; let mod $\in [0,1]$ be an arbitrary modulation factor. The activation level of neuron i at time t is given by Eq. (1):

$$\text{Activation}(i,t) = \sum_{j \in [1,m]} \text{mod}^{\text{order}(aj)} w_{j,i} \tag{1}$$

where $\text{order}(a_j)$ is the firing rank of neuron a_j in the ensemble A. By convention, $\text{order}(a_j) = +8$ if a neuron a_j is not fired at time t, sets the corresponding term in the above sum to zero. This kind of desensitization function could correspond to a fast shunting inhibition mechanism. Whenever a neuron reaches its threshold, it spikes and inhibits neurons at equivalent positions in the other maps so that only one neuron will respond at any particular location. Every spike also triggers a time based Hebbian-like learning rule that adjusts the synaptic weights. Let t_e be the date of arrival of the excitatory postsynaptic potential (EPSP) at synapse of weight W and t_a the date of discharge of the postsynaptic neuron.

$$\text{if } t_e < t_a \text{ then } dW = a(1\text{-}W)e^{-|\Delta o|\tau} \tag{2}$$

$$\text{else } dW = -aWe^{-|\Delta o|\tau}$$

where Δo is the difference between the date of the EPSP and the date of the neuronal discharge (expressed in terms of order of arrival instead of time), as is a constant that controls the amount of synaptic potentiation and depression [58].

ROPE technique with receptive fields allows the encoding of continuous values by using a collection of neurons with overlapping sensitivity profiles [60]. Each input variable is encoded independently by a group of one-dimensional receptive fields (Fig. 2). For a variable n, an interval $[I^n_{min,} \; I^n_{max}]$ is defined. The Gaussian receptive field of neuron i is given by its center μi:

$$\mu i = I^n_{min} + \frac{2i\text{-}3}{2} \frac{I^n_{max} - I^n_{min}}{M\text{-}2} \tag{3}$$

The width σ is given by Eq. (4):

$$\sigma = \frac{1}{\beta} \frac{I^n_{max} - I^n_{min}}{M\text{-}2} \tag{4}$$

where $1 \leq \beta \leq 2$ and the parameter β directly controls the width of each Gaussian receptive field.

Figure 1 depicts an example encoding of a single variable. For the diagram ($\beta = 2$) the input interval $[I^n_{min,} \; I^n_{max}]$ was set to $[-1.5, 1.5]$ and M=5 receptive fields were used. For an input value v=0.75 (thick straight line in left figure) the intersection points with each Gaussian is computed (triangles), which are in turn translated into spike time delays (right figure).

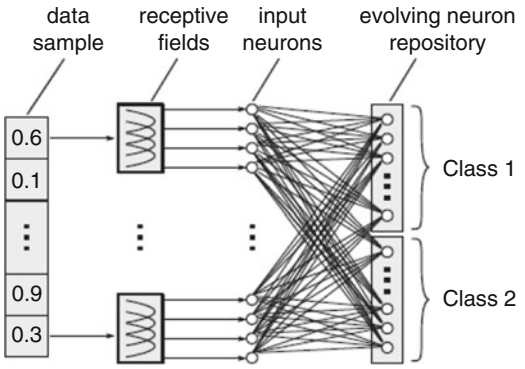


Fig. 1 The evolving spiking neural network (eSNN) architecture [23]

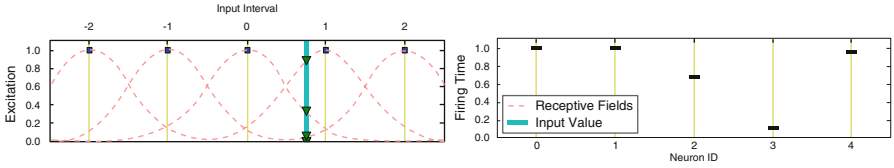


Fig. 2 Population encoding based on Gaussian receptive fields [23]

3.1.2 One-Pass Learning

The aim of the one-pass learning method is to create a repository of trained output neurons during the presentation of training samples. After presenting a certain input sample to the network, the corresponding spike train is propagated through the SANN which may result in the firing of certain output neurons. It is also possible that no output neuron is activated and in this case the network remains silent and the classification result is undetermined. If one or more output neurons have emitted a spike, the neuron with the shortest response time among all activated output neurons is determined. The label of this neuron represents the classification result for the presented input sample. The procedure is described in detail in the following Algorithm 1 [23, 60] (Fig. 2).

For each training sample i with class label $l \in L$ a new output neuron is created and fully connected to the previous layer of neurons resulting in a real-valued weight vector $w^{(i)}$ with $w_j^{(i)} \in R$ denoting the connection between the pre-synaptic neuron j and the created neuron i. In the next step, the input spikes are propagated through the network and the value of weight $w_j^{(i)}$ is computed according to the order of spike transmission through a synapse j: $w_j^{(i)}=(m_l)^{\text{order(j)}}$, $\forall j|j$ pre-synaptic neuron of i.

Parameter m_l is the modulation factor of the Thorpe neural model. Differently labeled output neurons may have different modulation factors m_l. Function order(j) represents the rank of the spike emitted by neuron j. The firing threshold $\theta^{(i)}$ of the created neuron I is defined as the fraction $c_l \in R$, $0 < c_l < 1$, of the maximal possible potential $u_{max}^{(i)}$:

Algorithm 1 Training an Evolving Spiking Neural Network (eSNN) [23]

Require: m_l, s_l, c_l **for a class label** l $\in$ L
initialize neuron repository $R_l = \{\}$
for all samples $X^{(i)}$ belonging to class l **do**
  $w_j^{(i)} \leftarrow (m_l)^{order(j)}$, $\forall$ j | j pre-synaptic neuron of i
  $u_{max}^{(i)} \leftarrow \sum_j w_j^{(i)} (m_l)^{order(j)}$
  $\theta^{(i)} \leftarrow c_l u_{max}^{(i)}$
if min(d($w^{(i)}$, $w^{(k)}$)) < s_l, $w^{(k)} \in R_l$ **then**
  $w^{(k)} \leftarrow$ merge $w^{(i)}$ and $w^{(k)}$ according to Eq. (6)
  $\theta^{(k)} \leftarrow$ merge $\theta^{(i)}$ and $\theta^{(k)}$ according to Eq. (7)
else
  $R_l \leftarrow R_l \cup \backslash\backslash \{w^{(i)}\backslash\backslash\}$
end if
end for

$$\theta^{(i)} \leftarrow c_l u_{max}^{(i)} \tag{5}$$

$$u_{max}^{(i)} \leftarrow \sum_j w_j^{(i)} (m_l)^{order(j)} \tag{6}$$

The fraction c_l is a parameter of the model and for each class label l $\in$ L a different fraction can be specified. The weight vector of the trained neuron is then compared to the weights corresponding to neurons already stored in the repository. Two neurons are considered too "similar" if the minimal *Euclidean* distance between their weight vectors is smaller than a specified similarity threshold s_l (the eSNN object uses optimal similarity threshold s=0.6). All parameters modulation factor m_l, similarity threshold s_l, PSP fraction c_l, l $\in$ L of ESNN which were included in this search space, are optimized according to the versatile quantum-inspired evolutionary algorithm (vQEA) [61]. In this case, both the firing thresholds and the weight vectors are merged according to Eqs. (7) and (8):

$$w_j^{(k)} \leftarrow \frac{w_j^{(i)} + N w_j^{(k)}}{1+N}, \forall j \mid j \text{ pre-synaptic neuron of i} \tag{7}$$

$$\theta^{(k)} \leftarrow \frac{\theta^{(i)} + N\theta^{(k)}}{1+N} \tag{8}$$

It must be clarified that integer N denotes the number of samples previously used to update neuron k. The merging is implemented as the (running) average of the connection weights, and the (running) average of the two firing thresholds. After the merging, the trained neuron i is discarded and the next sample processed. If no other neuron in the repository is similar to the trained neuron i, the neuron i is added to the repository as a new output neuron.

3.2 Multilayer Feed-Forward Neural Network

Artificial neural networks are biologically inspired classification algorithms that consist of an input layer of nodes, one or more hidden layers, and an output layer. Each node in a layer has one corresponding node in the next layer, thus creating the stacking effect [62]. Artificial neural networks are the very versatile tools and have been widely used to tackle many issues [63–67].

Feed-forward neural networks (FNN) are one of the popular structures among artificial neural networks. These efficient networks are widely used to solve complex problems by modeling complex input–output relationships [68, 69]. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In many applications the units of these networks apply a sigmoid function as an activation function.

The universal approximation theorem for neural networks states that every continuous function that maps intervals of real numbers to some output interval of real numbers can be approximated arbitrarily closely by a multi-layer perceptron with just one hidden layer. This result holds only for restricted classes of activation functions, e.g. for the sigmoidal functions.

Feed-forward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear relationships between input and output vectors. The linear output layer is most often used for function fitting (or nonlinear regression) problems.

Multi-layer networks use a variety of learning techniques, the most popular being back-propagation. Here, the output values are compared with the correct answer to compute the value of some predefined error-function. By various techniques, the error is then fed back through the network. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. After repeating this process for a sufficiently large number of training cycles, the network will usually converge to some state where the error of the calculations is small. In this case, one would say that the network has learned a certain target function. To adjust weights properly, one applies a general method for nonlinear optimization that is called gradient descent. For this, the derivative of the error function with respect to the network weights is calculated, and the weights are then changed such that the error decreases (thus going downhill on the surface of the error function).

3.3 Evolving Connectionist Systems

Evolving connectionist systems (ECOS) [70] are multi-modular, connectionist architectures that facilitate modeling of evolving processes and knowledge discovery [60]. An ECOS may consist of many evolving connectionist modules. An ECOS

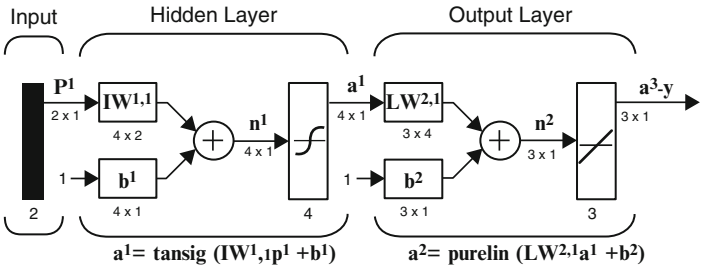


Fig. 3 Architecture of the multilayer feed-forward artificial neural network (http://www.mathworks.com/)

is a neural network that operates continuously in time and adapts its structure and functionality through a continuous interaction with the environment and with other systems according to:

- a set of parameters that are subject to change during the system operation;
- an incoming continuous flow of information with unknown distribution;
- a goal (rational) criterion (also subject to modification) that is applied to optimize the performance of the system over time.

The ECOS evolve in an open space, using constructive processes, not necessarily of fixed dimensions. Moreover, they learn in on-line incremental fast mode, possibly through one pass of data propagation. Life-long learning is a main attribute of this procedure. They operate as both individual systems and as part of an evolutionary population of such systems. They learn locally and locally partition the problem space, thus allowing for a fast adaptation and tracing processes over time. They facilitate different kinds of knowledge representation and extraction, mostly—memory based statistical and symbolic knowledge [60, 71, 72] (Fig. 3).

ECOS are connectionist structures that evolve their nodes (neurons) and connections through supervised incremental learning from input–output data pairs.

Their architecture comprises of five layers: input nodes, representing input variables; input fuzzy membership nodes, representing the membership degrees of the input values to each of the defined membership functions; rule nodes, representing cluster centers of samples in the problem space and their associated output function; output fuzzy membership nodes, representing the membership degrees to which the output values belong to defined membership functions; and output nodes, representing output variables [60, 71, 72].

ECOS learn local models from data through clustering of the data and associating a local output function for each cluster. Rule nodes evolve from the input data stream to cluster the data, and the first layer W1 connection weights of these nodes represent the coordinates of the nodes in the input space. The second layer W2 represents the local models (functions) allocated to each of the clusters.

Clusters of data are created based on similarity between data samples either in the input space or in both the input space and the output space. Samples that have a distance to an existing cluster center (rule node) N of less than a threshold Rmax are allocated to the same cluster Nc. Samples that do not fit into existing clusters form new clusters as they arrive in time. Cluster centers are continuously adjusted according to new data samples and new clusters are created incrementally. The similarity between a sample S = (x, y) and an existing rule node N = (W1, W2) can be measured in different ways, the most popular of them being the normalized Euclidean distance:

$$d(S, N) = \frac{1}{n}\left[\sum_{i=1}^{n} |x_i\text{-}W_{1N}|^2\right]^{\frac{1}{2}} \tag{9}$$

where n is the number of the input variables.

ECOS learn from data and automatically create a local output function for each cluster, the function being represented in the W2 connection weights, thus creating local models. Each model is represented as a local rule with an antecedent—the cluster area, and a consequent—the output function applied to data in this cluster.

The following is a corresponding example of such a local Rule:

- IF (data is in cluster Nc), THEN (the output is calculated with a function Fc)
- In the case of DENFIS [32], first order local fuzzy rule models are derived incrementally from data. The following rule is a characteristic example:
- IF (the value of x1 is in the area defined by a Gaussian function with a center at 0.7 and a standard deviation of 0.1) AND (the value of x2 is in the area defined by a Gaussian function with a center at 0.5 and a deviation of 0.2), THEN (the output value y is calculated with the use of the formula y= 3.7 + 0.5x1−4.2x2).

3.3.1 Evolving Classification Function

ECF, a special case of ECOS used for pattern classification, generates rule nodes in an N dimensional input space and associate them with classes. Each rule node is defined with its center, radius (influence field), and the class it belongs to. A learning mechanism is designed in such a way that the nodes can be generated.

The ECF model used here is a connectionist system for classification tasks that consists of four layers of neurons (nodes). The first layer represents the input variables; the second layer—the fuzzy membership functions; the third layer represents clusters centers (prototypes) of data in the input space; and the fourth layer represents classes [60, 70–72].

3.4 Genetic Algorithm

The genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution (http://www.mathworks.com/). The GA repeatedly modifies a population of individual solutions. At each step, the GA selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. You can apply the GA to solve a variety of optimization problems that are not well suited for standard optimization algorithms, including problems in which the objective function is discontinuous, nondifferentiable, stochastic, or highly nonlinear. Also the GA can address problems of mixed integer programming, where some components are restricted to be integer-valued.

The GA uses three main types of rules at each step to create the next generation from the current population:

- Selection rules select the individuals, called parents, that contribute to the population at the next generation.
- Crossover rules combine two parents to form children for the next generation.
- Mutation rules apply random changes to individual parents to form children.

The GA differs from a classical, derivative-based, optimization algorithm in two main ways, as follows:

- Classical Algorithm
  - Generates a single point at each iteration. The sequence of points approaches an optimal solution.
  - Selects the next point in the sequence by a deterministic computation.
- Genetic Algorithm
  - Generates a population of points at each iteration. The best point in the population approaches an optimal solution.
  - Selects the next population by computation which uses random number generators.

3.4.1 Genetic Algorithm for Offline ECF Optimization

A GA is applied to a population of solutions to a problem in order to "breed" better solutions. Solutions, in this case the parameters of the ECF network, are encoded in a binary string and each solution is given a score depending on how well it performs. Good solutions are selected more frequently for breeding, and are subjected to crossover and mutation (loosely analogous to those operations found in biological systems). After several generations, the population of solutions should converge on a "good" solution.

Given that the ECF system is a neural network that operates continuously in time and adapts its structure and functionality through a continuous interaction with the environment and with other systems according to a set of parameters P that are subject to change during the system operation; an incoming continuous flow of information with unknown distribution; a goal (rationale) criteria (also subject to modification) that is applied to optimize the performance of the system over time.

The set of parameters P of an ECOS can be regarded as a chromosome of "genes" of the evolving system and evolutionary computation can be applied for their optimization. The GA algorithm for offline ECF Optimization runs over generations of populations and standard operations are applied such as: binary encoding of the genes (parameters); roulette wheel selection criterion; multi-point crossover operation for crossover. Genes are complex structures and they cause dynamic transformation of one substance into another during the whole life of an individual, as well as the life of the human population over many generations.

Micro-array gene expression data can be used to evolve the ECF with inputs being the expression level of a certain number of selected genes and the outputs being the classes. After the ECF is trained on gene expression rules can be extracted that represent [73]. The ECF model and the GA algorithm for Offline ECF Optimization are parts from NeuCom software (http://www.kedri.aut.ac.nz/) which is a Neuro-Computing Decision Support Environment, based on the theory of ECOS [60, 70–72].

4 Description of the Proposed Hybrid Framework

Considering that the aim of the partial proposed systems is to carry out acts in a common environment, the architecture of the bioHAIFC can be simulated by a distributed multi-agent AI system. The agents are the three proposed Machine Learning systems, namely: (HESADM, ECISMD and ePSSQLI). These systems dynamically control the predefined sectors with a potential threat [74]. The synchronization of the Agents is achieved either with negotiation or with cooperation, as none of them has the full information package, there is no central control in the system, the data are distributed and the calculations are done in an asynchronous manner. The Agent communication and information exchange is done by a hybrid system of temporal programming in order to phase (in an optimal way) the potential contradiction of intensions and contradiction in the management of resources, based on priorities related to the extent of the threat and risk.

The results of the characterization of a threat are sent to the administrator of the network in a form of logs. The administrator tries to take necessary prevention actions in order to avoid the risk. Also the framework automates the potential direct termination of the TCP connection operation with the attacker for higher security and control (e.g., tcpkill host 192.168.1.2 or tcpkill host host12.blackhut.com).

The analytical description of the partial systems of the bioHAIFCS is described below:

4.1 Hybrid Evolving Spiking Anomaly Detection Model

The HESADM methodology uses eSNN classification approach and Multi-Layer Feed Forward ANN in order to classify the exact type of the intrusion or anomaly in the network with minimum computational power. The dataset which used and the general algorithm are described in detail below:

4.1.1 Data

The *KDD Cup* 1999 data set [75] was used to test the herein proposed approach. This data set was created in the LincolnLab of MIT and it is the most popular free data set used in evaluation of IDS. It contains recordings of the total network flow of a local network which was installed in the Lincoln Labs and it simulates the military network of the USA air force. The method of events' analysis includes a connection between a source IP address and a destination IP, during which a sequence of TCP packages is exchanged, by using a specific protocol and a strictly defined operation time.

The KDD Cup 1999 data includes 41 characteristics which are organized in the following four basic categories: Content Features, Traffic Features, Time-based Traffic Features, Host-based Traffic Features. Also the attacks are divided into four categories, namely: DoS, r2l, u2r, and probe.

Using the eSNN Traf_Red_Full.data In the first classification case, all (41) features were used. The data were classified as normal or abnormal. The dataset Traf_Red_Full.data has 145,738 records and the 75 % (109,303 rec.) used as train_data and the 25 % (36,435 rec.) used as test_data.

Using the SNN normalFull.data In the second classification case, the relevant normal features comprising of 11 features were used. The data were classified as normal or abnormal. The dataset normalFull.data has 145,738 records and the 75 % (109,303 rec.) were used as train_data and the 25 % (36,435 rec.) as test_data.

4.1.2 Algorithm

- **Step 1**

We choose to use the traffic oriented data, which is related to only nine features. We import the required classes that use the variable Population Encoding. This variable controls the conversion of real-valued data samples into the corresponding time spikes. The encoding is performed with 20 Gaussian receptive fields per variable (Gaussian width parameter beta=1.5). We also normalize the data to the interval [−1,1] and so we indicate the coverage of the Gaussians using i_min and i_max. For the normalization processing the following function 10 was used:

$$x_{1_{\text{norm}}} = 2* \left(\frac{x_1 - x_{\min}}{x_{\max} - x_{\min}} \right) - 1, x \in R \tag{10}$$

The data is classified into two classes namely: class 0 which contains the normal results and class 1 which comprises of the abnormal ones (DoS, r2l, u2r and probe). The eSNN object using modulation factor m=0.9, firing threshold ratio c=0.7 and similarity threshold s=0.6 in agreement with the vQEA algorithm [23, 61].

- **Step 2**

We train the eSNN with 75 % of the dataset vectors (*train_data) a*nd we test the eSNN with 25 % of the dataset vectors *(test_data).* The training process is described in Algorithm 1.

- **Step 3**

If the result of the classification is normal, the eSNN classification process is repeated but this time the relevant normal data vectors are used. These vectors are comprised of 11 features [9]. If the result is normal, then the process is terminated. If the result of the classification is abnormal, a two-layer feed-forward neural network with sigmoid function both in hidden and output layer with scaled conjugate gradient backpropagation as the learning algorithm is used to perform pattern recognition of the attack type with all features of KDD dataset (41 inputs and 5 outputs).

The outcome of the pattern recognition process is submitted in the form of an *Alert* signal to the network administrator. A Graphical display of the complete HESADM methodology can be seen in Fig. 4.

The performance metric used is the mean squared error (MSE). The MLFF ANN was developed with 41 input neurons, corresponding to the 41 input parameters of the KDD cup 1999 dataset, 33 neurons in the Hidden Layer, and 5 in the output one corresponding to the following output parameters: DoS, r2l, u2r, Probe, normal. In the hidden layer 33 neurons are used, based on the following empirical function 11 [76]:

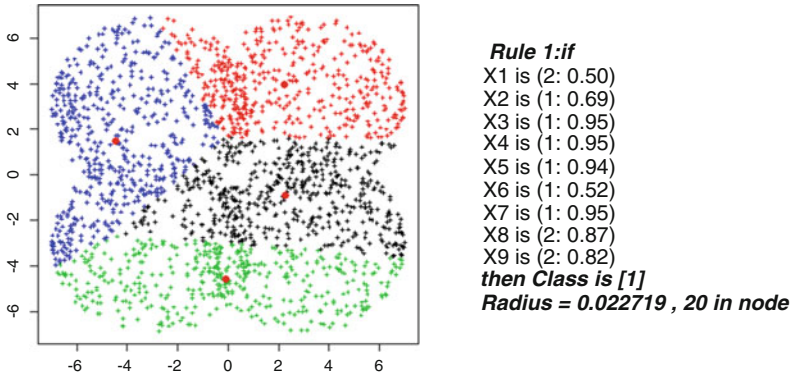


Fig. 4 Rule of the evolving connectionist system [60, 70–72]

$$\left(\frac{2}{3}*\text{Inputs}\right) + \text{Outputs} = \left(\frac{2}{3}*41\right) + 5 = 33 \tag{11}$$

The KDD cup 1999 dataset was divided randomly into 70 % (102,016 rec.) the train_data, 15 % (21,861 rec.) as test_data and the rest 15 % (21,861 records) as validation_data.

4.2 *Evolving Computational Intelligence System for Malware Detection*

The proposed herein, hybrid ECISMD methodology uses an eSNN classification approach to classify packed or unpacked executables with minimum computational power combined with the ECF method in order to detect packed malware. Finally it applies Genetic Algorithm for ECF Optimization, in order to decrease the level of false positive and false negative rates (Fig. 5).

The dataset which used and the general algorithm are described below:

4.2.1 Dataset

The full_dataset comprised of 2598 packed viruses from the Malfease Project dataset (http://malfease.oarci.net), 2231 non-packed benign executables collected from a clean installation of *Windows XP Home plus*, several common user applications and 669 packed benign executables.

The dataset was divided randomly into two parts:

- A training dataset containing 2231 patterns related to the non-packed benign executable and 2262 patterns related to the packed executables detected using unpacked software
- A testing dataset containing 1005 patterns related to the packed executables that even the best known unpacked software was not able to detect. These datasets are available at http://roberto.perdisci.googlepages.com/code [32].

The virus dataset containing 2598 malware and 669 benign executables is divided into two parts:

- A training dataset containing 1834 patterns related to the malware and 453 patterns related to the benign executables
- A test dataset containing 762 patterns related to the malware and 218 benign executables. In order to translate each executable into a pattern vector Perdisci et al. [32] use binary static analysis, to extract information such as the name of the code and data sections, the number of writable-executable sections, the code and data entropy.

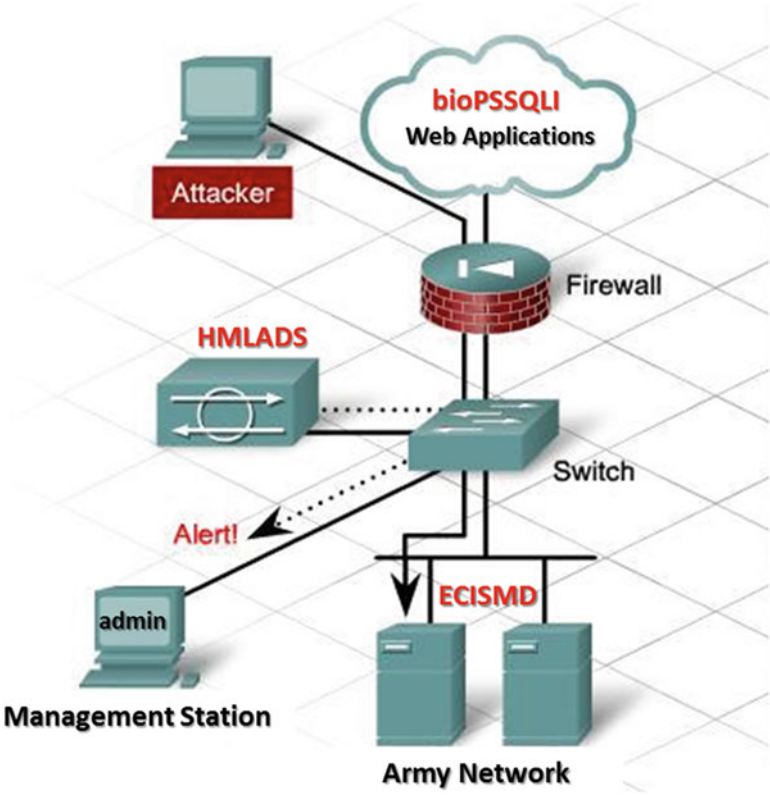


Fig. 5 Bio-inspired hybrid artificial intelligence framework for cyber security

In the first classification performed by the ECISMD, the eSNN approach was employed in order to classify packed or not packed executables.

In the second classification performed by the ECISMD, the ECF approach was employed in order to classify malware or benign executables.

4.2.2 Algorithm

- **Step 1**

The train and test *datasets* are determined and formed, related to *n* features. The required classes (packed and unpacked executables) that use the variable *Population Encoding* are imported. This variable controls the conversion of real-valued data samples into the corresponding time spikes. The encoding is performed with 20 Gaussian receptive fields per variable (Gaussian width parameter beta=1.5). The data are normalized to the interval [−1,1] and so the coverage of the Gaussians is determined by using i_min and i_max. For the normalization processing function 10 is used (Fig. 6).

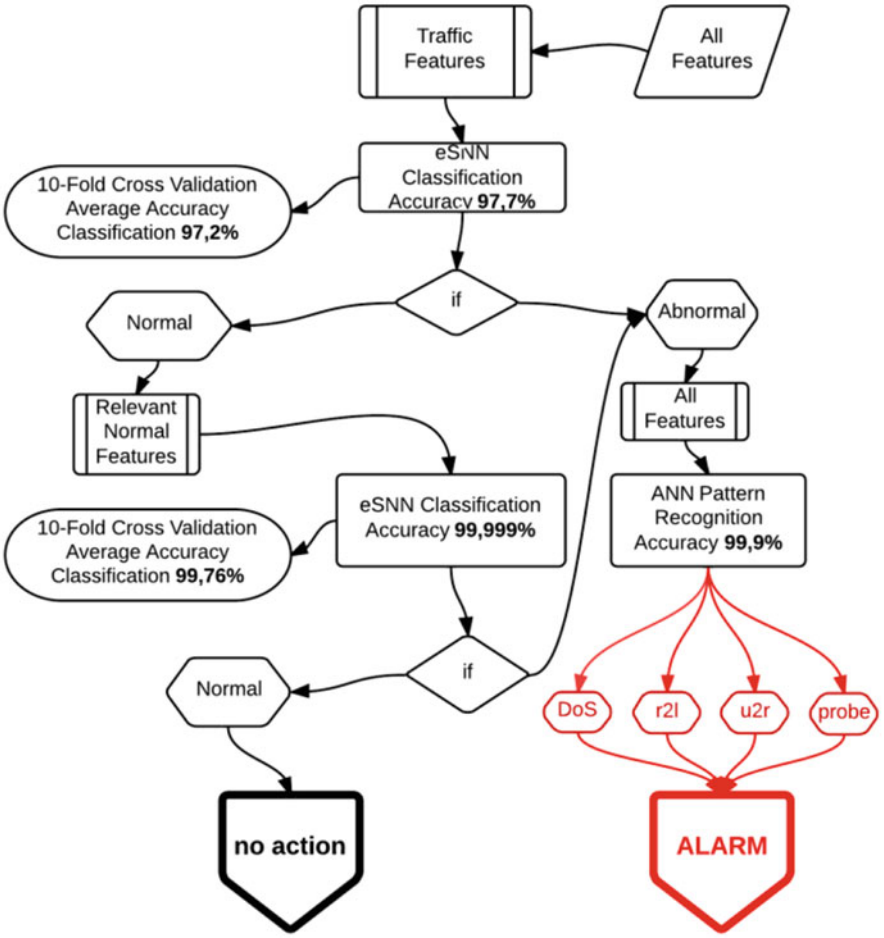


Fig. 6 The hybrid evolving spiking anomaly detection model (HESADM) methodology

The data is classified into two classes, namely: ***Class 0*** which contains the unpacked results and ***Class 1*** which comprises of the packed ones. The eSNN object using modulation factor m=0.9, firing threshold ratio c=0.7, and similarity threshold s=0.6 in agreement with the vQEA algorithm [23, 61].

- **Step 2**

The eSNN is trained with the *packed_train* dataset vectors and the testing is performed with the *packed_test* vectors. The training process is described in Algorithm 1.

- **Step 3**

If the result is unpacked, then the process is terminated and the executable file goes to the antivirus scanner. If the result of the classification is packed, the new classification process is initiated employing the ECF method. This time the malware data vectors are used. These vectors comprise of nine features and two classes malware and benign.

The learning algorithm of the ECF according to the ECOS is as follows:

- If all input vectors are fed, finish the iteration; otherwise, input a vector from the data set and calculate the distances between the vector and all rule nodes already created using Euclidean distance.
- If all distances are greater than a max-radius parameter, a new rule node is created. The position of the new rule node is the same as the current vector in the input data space and the radius of its receptive field is set to the min-radius parameter; the algorithm goes to step 1; otherwise it goes to the next step.
- If there is a rule node with a distance to the current input vector less than or equal to its radius and its class is the same as the class of the new vector, nothing will be changed; go to step 1; otherwise.
- If there is a rule node with a distance to the input vector less than or equal to its radius and its class is different from those of the input vector, its influence field should be reduced. The radius of the new field is set to the larger value from the two numbers: distance minus the min-radius; min radius. New node is created as in to represent the new data vector.
- If there is a rule node with a distance to the input vector less than or equal to the max-radius, and its class is the same as of the input vector's, enlarge the influence field by taking the distance as a new radius if only such enlarged field does not cover any other rule nodes which belong to a different class; otherwise, create a new rule node in the same way as in step 2, and go to step 1 [77].

- **Step 4**

To increase the level of integrity the Offline ECF Optimization with GA is used.

- **Step 5**

If the result of the classification is benign, the executable file goes to antivirus scanner and the process is terminated. Otherwise, the executable file is marked as malicious, it goes to the unpacker, to the antivirus scanner for verification and finally placed in quarantine and the process is terminated (Fig. 7).

4.3 Evolutionary Prevention System from SQL Injection

The proposed ePSSQLI model uses an MFFNN which has optimized with a GA. Generally, there are three methods of using a GA for training MFFNNs. Firstly, GA is utilized for finding a combination of weights and biases that provide the minimum

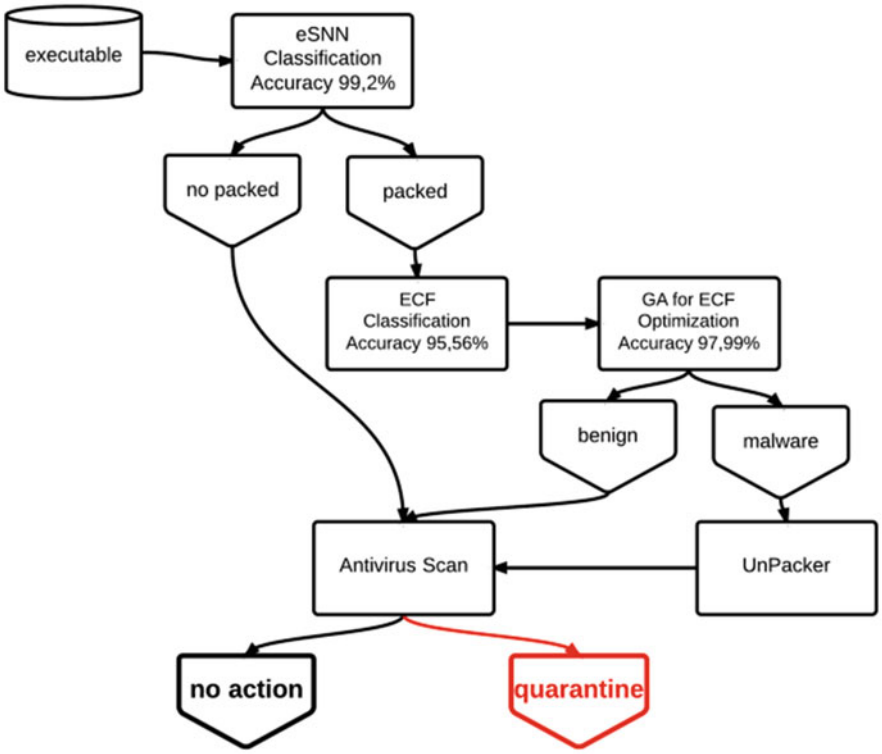


Fig. 7 Graphical display of the ECISMD algorithm

error for an MFFNN. Secondly, GA is employed to find a proper architecture for an MFFNN in a particular problem. The last method is to use a GA to tune the parameters of a gradient-based learning algorithm, such as the learning rate and momentum. In the first method, the architecture does not change during the learning process. The training algorithm is required to find proper values for all connection weights and biases in order to minimize the overall error of the MFFNN. In the second approach, the structure of the MFFNNs varies. In this case, a training algorithm determines the best structure for solving a certain problem. Changing the structure can be accomplished by manipulating the connections between neurons, the number of hidden layers, and the number of hidden nodes in each layer. In this study the GA is applied to minimize the error of MFFNN in order to classify SQL injections with high accuracy.

The dataset which used and the general algorithm are described below:

4.3.1 Dataset

The dataset used includes a list of 13,884 SQL statements that have been selected by various sources. Actually, 12,881 of them are malicious (SQL Injections) and 1003 are legit. With the help of the SQLparse module (https://github.com/andialbrecht/sqlparse) in Python, which is a non-validating SQL one, we have searched the way of syntax and use of certain SQL symbols in the construction of SQL injections commands. Also we investigated the correlation of SQL statements with the attacks of SQL injections' type.

Finally, the n-gram technique was used to search the correlation of the SQL statements sequence, with the syntax of the SQL injections commands (https://github.com/ClickSecurity/data_hacking). In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words, or base pairs according to the application. The n-grams in this case are collected from an SQL statements.

Various malicious και legit scores constitute the statistical output of the SQL statements and they were used as features. In information theory, entropy is a measure of the uncertainty associated with a random variable. The term by itself in this context usually refers to the Shannon entropy, which quantifies, in the sense of an expected value, the information contained in a message, usually in units such as bits. Equivalently, the Shannon entropy is a measure of the average information content one is missing when one does not know the value of the random variable [78].

After its adjustment, the dataset includes the following parameters:

- Length
- Entropy
- Malicious_score
- Legit_score
- Difference_score
- Class

In the pre-processing of data remove extreme values and outliers. The extreme value is a point which is far away from the average value of a parameter. The distance is measured based on a threshold which is a multiplicand of the standard deviation (Fig. 8).

We know that for a random parameter that is under normal distribution, the 95 % of all the values fall up to the value of 2*stdev whereas 99 % fall up to the value of 3*stdev. Extreme values cause significant errors in a potential model. Things become even worse when these extreme values are noise results during measurements procedure. If the number of extreme values is small, then they are removed from the data set.

The estimation of the extreme values was done under the Inter Quartile Range method [79]. This method spots extreme values and outliers based on (InterQuartile Ranges—IQR). The IQR is the difference between the third (Q3) and the first (Q1)

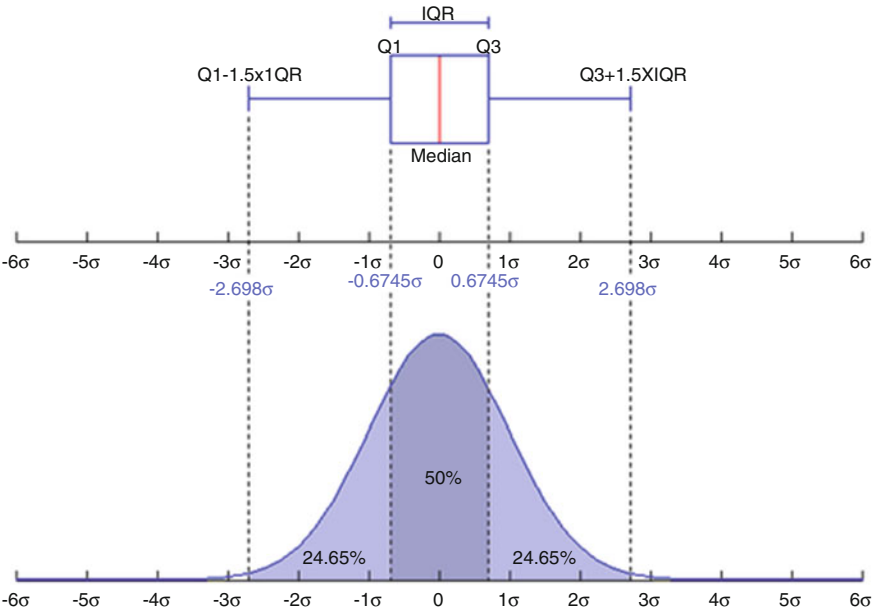


Fig. 8 Graphical display of inter quartile range method

quartile, IQR= Q3 – Q1. The quartiles divide the data into four equal parts. The IQR includes the imtermediate 50 % of the data whereas the rest 25 % is less than Q1 and the rest 25 % is higher than Q3 [2]. The calculation of the Extreme values was done as follows:

- Outliers:
  - Q3 + OF*IQR < x <= Q3 + EVF*IQR or Q1 - EVF*IQR <= x < Q1 - OF*IQR
- Extreme values:
  - x > Q3 + EVF*IQR or x < Q1 - EVF*IQR

Key: Q1=25 % quartile, Q3=75 % quartile, IQR=Interquartile Range difference between Q1 and Q3, OF=Outlier Factor, EVF=Extreme Value Factor.

With the use of the above method 12 outliers and three extreme values were removed from the data set which was reduced to 13,869 cases (12,881 malicious, 988 legit).

Also the data were Normalized so that they can have the proper input for the Learning Algorithms in the interval [−1, +1].

After a relative observation we can realize that we have created an imbalanced dataset which includes 13,869 cases from which 12,881 are malicious and 988 legit (0.0723 %). Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed

by two classes: The majority (negative) class and the minority (positive) class. The problem with class imbalances is that standard learners are often biased towards the majority class. That is because these classifiers attempt to reduce global quantities such as the error rate, not taking the data distribution into consideration. As a result examples from the overwhelming class are well classified whereas examples from the minority class tend to be misclassified.

To resolve the certain problem we use the technique synthetic minority over-sampling technique (SMOTE) in order to resample the dataset. [80]. Re-sampling provides a simple way of biasing the generalization process. It can do so by generating synthetic samples accordingly biased and controlling the amount and placement of the new samples. SMOTE is a technique which combines Informed oversampling of the minority class with random undersampling of the majority class. SMOTE is a technique which is combines Informed oversampling of the minority class with random undersampling of the majority class and produce the best results as far as re-sampli ng and modifying the probabilistic estimate techniques.

For each minority sample, SMOTE works as follows:

- Find its k-nearest minority neighbors.
- Randomly select *j* of these neighbors.
- Randomly generate synthetic samples along the lines joining the minority sample and its *j* selected neighbors (*j* depends on the amount of oversampling desired).

By applying the SMOTE approach we re-created the dataset, which includes 21,773 cases, from which 12,881 are malicious and 8892 are legit.

4.4 Algorithm

The MLFF ANN was developed with five input neurons, corresponding to the five input parameters of the dataset, five neurons in the Hidden Layer and two in the output one corresponding to the following output parameters: malicious or legit. In the hidden layer five neurons are used, based on the empirical function 11.

This adds a greater degree of integrity to the rest of security infrastructure MFF ANN, optimized with GA. The following outline summarizes how the GA works:

- The algorithm begins by creating a random initial population.
- The algorithm then creates a sequence of new populations. At each step, the algorithm uses the individuals in the current generation to create the next population.
- To create the new population, the algorithm performs the following steps:
  - Scores each member of the current population by computing its fitness value.

- Scales the raw fitness scores to convert them into a more usable range of values.
- Selects members, called parents, based on their fitness.
- Some of the individuals in the current population that have lower fitness are chosen as *elite*. These elite individuals are passed to the next population.
- Produces children from the parents. Children are produced either by making random changes to a single parent—*mutation*—or by combining the vector entries of a pair of parents—*crossover*.
- Replaces the current population with the children to form the next generation.
- The algorithm stops when one of the stopping criteria is met.

5 Results

Each subsystem was tested based on multiple scenarios and different datasets were used for each case of threat. The results obtained are very encouraging as the accuracy is as high as 99 %, resulting in a reduction of the false alarms to the minimum. This fact, combined with the flexibility of the proposed system and with its generalization ability and the spotting of zero-day threats, makes its use suitable for critical applications like the one of military networks protection. The results of each case are presented below:

5.1 Hybrid Evolving Spiking Anomaly Detection Model

5.1.1 eSNN Approach

- In the first classification using the eSNN Traf_Red_Full.data the data classified as normal or abnormal. The results are shown below:
  - Classification Accuracy: 97.7 %
  - No. of evolved neurons: Class 0: 794 neurons, Class 1: 809 neurons
  - The average accuracy after applying tenfold Classification in the Traf_Red_Full.data was as high as 97.2 %.
- In the second classification case using the SNN normalFull.data, the relevant normal features comprising of 11 features were used. The data were classified as normal or abnormal. The results are shown below:
  - Classification Accuracy: 99.99 %
  - No. of evolved neurons: Class 0: 646 neurons, Class 1: 136 neurons
  - The average accuracy after applying tenfold Classification in the normalFull.data was as high as 99.76 %.

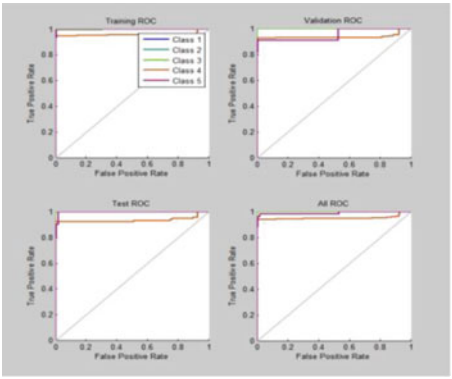


Fig. 9 ROC analysis

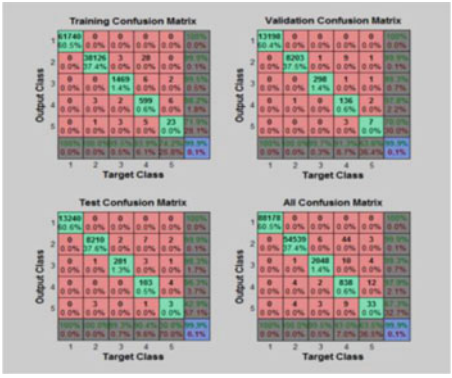


Fig. 10 Confusion matrix

5.1.2 MLFF ANN Approach

The classification accuracy is as high as 99.9% and all the performance metrics support the high level of convergence of the model.

In Fig. 9 the colored lines in each axis represent the ROC curves. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) as the threshold is varied. A perfect test would show points in the upper-left corner, with 100 % sensitivity and 100 % specificity. For this problem, the network performs very well.

Figure 10 shows the confusion matrices for training, testing, and validation, and the three kinds of data combined. The network outputs are very accurate, by the high numbers of correct responses in the green squares and the low numbers of incorrect responses in the red squares. The lower right blue squares illustrate the overall accuracies.

5.2 ECISMD Results

Table 1 reports the average accuracy which computed over tenfold cross-validation obtained with RBF ANN, Naïve Bayes, multi layer perceptron (MLP), Support Vector Machine (SVM), k-Nearest-Neighbors (k-NN), and eSNN. The best results on the testing dataset were obtained by using the eSNN classifier, to classify packed or not packed executables.

Table 2 reports the results obtained with six classifiers and optimized ECF network (RBF Network, Naïve Bayes, MLP, Lib SVM, k-NN, ECF, and optimized ECF). The best results on the testing dataset were obtained by using the optimized ECF which classifies virus or benign executables (Table 3).

5.3 ePSSQLI Results

5.3.1 MFF ANN

The classification accuracy of the MFF ANN that uses tenfold Cross Validation before the optimization is equal to 97.7 %. The rest of the measurements and the confusion matrix are presented below (Table 4):

Table 1 Comparison of various approaches for the packed dataset

Packed dataset		
Classifier	Train accuracy (%)	Test accuracy (%)
RBFNetwork	98.3085	98.0859
NaiveBayes	98.3975	97.1144
MLP	99.5326	96.2189
LibSVM	99.4436	89.8507
k-NN	99.4436	96.6169
eSNN	99.8	99.2

Table 2 Comparison of various approaches for the virus dataset

Virus dataset		
Classifier	Train accuracy (%)	Test accuracy (%)
RBFNetwork	94.4031	93.0612
NaiveBayes	94.0533	92.3469
MLP	97.7551	97.289
LibSVM	94.6218	94.2857
k-NN	98.1198	96.8367
ECF	99.05	95.561
Optimized ECF	99.87	97.992

Table 3 Metrics of the MFF ANN

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.986	0.034	0.976	0.986	0.981	0.986	Malicious
0.966	0.014	0.980	0.966	0.973	0.986	Legit

Table 4 Confusion matrix of the MFF ANN

Malicious	Legit
12,702	179
306	8586

Table 5 Metrics of the MFF ANN with GA

TP rate	FP rate	Precision	Recall	F-measure	ROC area	Class
0.997	0.003	0.998	0.997	0.997	0.998	Malicious
0.997	0.003	0.996	0.997	0.996	0.998	Legit

Table 6 Confusion matrix of the MFF ANN with GA

Malicious	Legit
12,845	36
31	8861

5.3.2 MFF ANN Optimized with GA

The initial parameters of GA are as below (Table 5):

- Selection: *Roulette wheel*
- Crossover: *Single point (probability = 1)*
- Mutation: *Uniform (probability = 0.01)*
- Population size: *200*
- Maximum number of generations: *250*

The classification accuracy of the MFF ANN that uses tenfold Cross Validation after its optimization with GA is 99.6 %. The rest of the measurements and the confusion matrix are presented below (Table 6):

The good performance and reliability of the proposed scheme that uses MFF ANN with GA is shown in Table 7 below. Table 7 presents the results of the categorization with the same dataset and by employing tenfold Cross Validation and other Machine Learning approaches.

6 Discussion: Conclusions

This paper proposes the use of a Bio-Inspired Hybrid Artificial Intelligence Framework for Cyber Security, which is based on the combination of three timely methods of Artificial Intelligence.

Table 7 Comparison of various approaches for the SQLI dataset

SQLI dataset	
Classifier	Accuracy (%)
MFF ANN with GA	99.6
RBFNetwork	97.3
fNaiveBayes	95.6
BayesNet	98.7
SVM	98.5
k-NN	98.3
Random forest	99.1

The function of the subsystems aims in the time spotting of the cyber-attacks which are untraceable with the classical passive protection approaches.

More specifically, this paper proposes the HESADM system, which spots potential anomalies of a network and the attacks that might bypass the firewall and the IDS. The second subsystem is ECISMD which scans the packed executable files and then spots malicious code untraceable by antivirus. The third one is ePSSQLI which spots in time the SQL Injections attacks. The result of each categorization is sent to the administrator of the system so that he/she can impose proper actions. An automatic disconnection from the attacker is also included.

The combination of the subsystems under the proposed framework takes place based on a temporal scheduling which succeeds the optimal distribution of the resources and the maximum availability and performance of the system. The use of the proposed systems can be done regardless of the framework.

The testing has resulted in an accuracy level of 99 %. Also a comparative analysis has revealed that the proposed algorithm outperforms the existing ones.

As a future direction, aiming to improve the efficiency of biologically realistic ANN for pattern recognition, it would be important to evaluate the eSNN model with ROC analysis and to perform feature minimization in order to achieve minimum processing time. Other coding schemes could be explored and compared on the same security task. Also, the ECISMD could be improved towards a better online learning with self-modified parameter values. Finally, the MFF ANN with GA which used in the ePSSQLI system could be compared with other optimization schemes like particle swarm optimization.

References

1. Garcıa Teodoro, P., Dıaz-Verdejo, J., Macia-Fernandez, G., Vazquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Elsevier Comput. Security **28**, 18–28 (2009)
2. Demertzis, K., Iliadis, L.: A hybrid network anomaly and intrusion detection approach based on evolving spiking neural network classification. In: E-Democracy, Security, Privacy and Trust in a Digital World. Communications in Computer and Information Science, vol. 441, pp. 11–23. (2014). doi:10.1007/978-3-319-11710-2_2

3. Yan, W., Zhang, Z., Ansari, N.: Revealing packed malware. IEEE Secur. Priv. **6**(5), 65–69 (2007)
4. Cesare, S., Xiang, Y.: Software Similarity and Classification. Springer, New York (2012)
5. Demertzis, K., Iliadis, L.: Evolving computational intelligence system for malware detection. In: Advanced Information Systems Engineering Workshops. Lecture Notes in Business Information Processing, vol. 178, pp. 322–334. (2014). doi:10.1007/978-3-319-07869-4_30
6. Open Web Application Security Project (OWASP): (2014) https://www.owasp.org
7. Dorothy, D.E.: An intrusion-detection model. IEEE Trans. Softw. Eng. **13**, 222–232 (1987). doi:10.1109/TSE.1987.232894
8. Puketza, N., Zhang, K., Chung, M., Mukherjee, B., Olsson, R.A.: A methodology for testing intrusion detection system. IEEE Trans. Softw. Eng. **22**, 719–729 (1996). doi:10.1109/32.544350
9. *B*harti, K., Jain, S., Shukla, S.: Fuzzy K-mean clustering via random forest for intrusiion detection system. Int. J. Comput. Sci. Eng. **02**(06), 2197–2200 (2010)
10. Mehdi B., Mohammad B.: An overview to software architecture in intrusion detection system. Int. J. Soft Comput. Softw. Eng. (2012). doi:10.7321/jscse.v1.n1.1
11. Muna, M., Jawhar, T., Monica, M.: Design network intrusion system using hybrid fuzzy neural network. Int. J. Comput. Sci. Secur. **4**(3), 285–294 (2009)
12. Jakir, H., Rahman, A., Sayeed, S., Samsuddin, K., Rokhani, F.: A modified hybrid fuzzy clustering algorithm for data partitions. Aust. J. Basic Appl. Sci. **5**, 674–681 (2011)
13. Suguna, J., Selvi, A.M.: Ensemble fuzzy clustering for mixed numeric and categorical data. Int. J. Comput. Appl. **42**, 19–23 (2012). doi:10.5120/5673-7705
14. Vladimir, V.: The Nature of Statistical Learning Theory, 2nd edn., p. 188. Springer, New York (1995). ISBN-10: 0387945598
15. John, G.H.: Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, (UAI' 95), pp. 338–345. Morgan Kaufmann Publishers Inc., San Francisco (1995)
16. Sang-Jun, H., Sung-Bae, C.: Evolutionary neural networks for anomaly detection based on the behavior of a program. IEEE Trans. Syst. Man Cybern. **36**, 559–570 (2005) doi:10.1109/TSMCB.2005.860136
17. Mehdi, M., Mohammad, Z.: A neural network based system for intrusion detection and classification of attacks. In: IEEE International Conference on Advances in Intelligent Systems - Theory and Applications (2004)
18. Zhou, T.-J.: The research of intrusion detection based on genetic neural network. In: Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, pp. 276–281, 30–31 Aug 2008. IEEE Xplore Press, Hong Kong (2008). doi:10.1109/ICWAPR.2008.4635789
19. Novikov, D., Yampolskiy, R.V., Reznik, L.: Anomaly detection based intrusion detection. In: Proceedings of the Third International Conference on Information Technology: New Generations, pp. 420–425, 10–12 April 2006. IEEE Xplore Press, Las Vegas (2006) doi:10.1109/ITNG.2006.33
20. Dahlia, A., Zainaddin, A., Mohd Hanapi, Z.: Hybrid of fuzzy clustering neural network over nsl dataset for intrusion detection system. J. Comput. Sci. **9**(3), 391–403 (2013). ISSN: 1549-3636 2013. doi:10.3844/jcssp.2013391 403 [Science Publications]
21. Tartakovskya, A.G., Rozovskii, B.L., Rudolf, B., Blazek, R.B., Kim, H.J.: A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. IEEE Trans. Signal Process. **54**(9) (2006). doi:10.1109/TSP.2006.879308
22. Mukhopadhyay, I.: Implementation of Kalman filter in intrusion detection system. In: Proceeding of ISCI Technologies, Vientiane (2008)
23. Simei Gomes, W., Lubica, B., Kasabov Nikola, K.: Adaptive learning procedure for a network of spiking neurons and visual pattern recognition. In: Advanced Concepts for Intelligent Vision Systems. Springer, New York (2006)
24. Babar, K., Khalid, F.: Generic unpacking techniques., Computer, Control and Communication, 2nd International Conference on IC4 IEEE (2009), DOI:10.1109/IC4.2009.4909168 (2009)

25. Royal, P., Halpin, M., Dagon, D., Edmonds, R.: Polyunpack: automating the hidden-code extraction of unpack-executing malware. In: ACSAC (2006)
26. Kang, M., Poosankam, P., Yin, H.: Renovo: a hidden code extractor for packed executables. In: 2007 ACM Workshop on Recurring Malcode (2007)
27. Martignoni, L., Christodorescu, M., Jha, S.: Omniunpack: fast, generic, and safe unpacking of malware. In: Proceedings of the ACSAC, pp. 431/441 (2007)
28. Yegneswaran, V., Saidi, H., Porras, P., Sharif, M.: Eureka: a framework for enabling static analysis on malware. Technical Report SRI-CSL-08-01 (2008)
29. Danielescu, A.: Anti-debugging and anti-emulation techniques. Code-Breakers J. **5**(1), 27–30 (2008)
30. Farooq, M.: PE-Miner: mining structural information to detect malicious executables in realtime. In: 12th Symposium on Recent Advances in ID, pp. 121–141. Springer, New York (2009)
31. Shaq, M., Tabish, S., Farooq, M.: PE-probe: leveraging packer detection and structural information to detect malicious portable executables. In: Proceedings of the Virus Bulletin Conference (2009)
32. Perdisci, R., Lanzi, A., Lee, W.: McBoost: boosting scalability in malware collection and analysis using statistical classiffication of executables. In: Proceedings of the 2008 Annual Computer Security Applications Conference, pp. 301/310 (2008). ISSN: 1063–9527
33. Kolter, J.Z., Maloof, M.A.: Learning to detect and classify malicious executables in the wild. J. ML Res. **7**, 2721–2744 (2006)
34. Ugarte-Pedrero, X., Santos, I., Bringas, P.G., Gastesi , M., Esparza, J.M.: Semi-supervised Learning for Packed Executable Detection, Network and System Security (NSS), 5th International Conference on, (2011). DOI: 10.1109/ICNSS.2011.6060027
35. Ugarte-Pedrero, X., Santos, I., Laorden, C., Sanz, B., Bringas, G.P.: Collective classification for packed executable identification. In: ACM CEAS (2011)
36. Gavrilut, D., Cimpoes, M., Anton, D., Ciortuz, L.: Malware detection using machine learning. In: Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 735–741 (2009). ISBN: 978-83-60810-22-4
37. Ye, Y., Wang, D., Li, T., Ye, D.: Imds: Intelligent Malware Detection System. ACM, New York (2007)
38. Chandrasekaran, M., Vidyaraman, V., Upadhyaya S.J.: Spycon: emulating user activities to detect evasive spyware. Performance, Computing, and Communications Conference, 2007. In: IPCCC 2007. IEEE International Conference on (2007). DOI:10.1109/PCCC.2007.358933
39. Chouchane, M.R., Walenstein, A., Lakhotia, A.: Using Markov Chains to filter machine-morphed variants of malicious programs. In: 3rd International Conference on Malicious and Unwanted Software, 2008, MALWARE 2008, pp. 77–84 (2008)
40. Stamp, M., Attaluri, S., McGhee, S.: Profile hidden marko v models and metamorphic virus detection. J. Comput. Virol. **5**(2):151-169 (2009). DOI: 10.1007/s11416-008-0105-1
41. Santamarta, R.: Generic detection and classification of polymorphic malware using neural pattern recognition, white paper, ReverseMode. http://www.reversemode.com/ (2006)
42. Yoo, I.: Visualizing windows executable viruses using self-organizing maps. In: VizSEC/DMSEC '04: ACM Workshop (2004)
43. Livshits, V.B., Lam, M.S.: Finding Security vulnerability in Java applications with static analysis. In: Proceedings of the 14th USS, August 2005
44. Halfond, W.G.J., Orso, A., Manolios, P.: WASP: protecting web applications using positive tainting and syntax-aware evaluation. IEEE Trans. Softw. Eng. **34**, 181–191 (2008)
45. Buehrer, G.T., Weide, B.W., Sivilotti, Using Parse tree validation to prevent SQL injection attacks. In: Proceeding of the 5th International Workshop on Software Engineering and Middleware (SEM '056), pp. 106–113, September 2005
46. Cova, M., Balzarotti, D., Felmetsger, V., Vigna, G.: Swaddler: an approach for the anamoly based character distribution models in the detection of SQL injection attacks. In: Recent Advances in Intrusion Detection System, pp. 63–86. Springerlink, New York (2007)

47. Gerstenberger, R.: Anomaliebasierte Angriffserkennung im FTP-Protokoll. Master's Thesis, University of Potsdam, Germany (2008)
48. Dùssel, P., Gehl, C., Laskov, P., Rieck, K.: Incorporation of application layer protocol syntax into anomaly detection. In: Sekar, R., Pujari, A.K. (eds.) ICISS 2008. LNCS, vol. 5352, pp. 188–202. Springer, Heidelberg (2008)
49. Bockermann, C., Apel, M., Meier, M.: Learning sql. for database intrusion detection using context-sensitive modelling. In: Detection of Intrusions and Malware, and Vulnerability Assessment, vol. 5587/2009, pp. 196–205. Springer Berlin/Heidelberg (2009)
50. Dewhurst, R.: Damn Vulnerable Web Application (DVWA). http://www.dvwa.co.uk/ (2012)
51. Bernardo Damele, A.G., Stampar, M.: Sqlmap: automatic SQL injection and database takeover tool. http://sqlmap.sourceforge.net/ (2012)
52. Valeur, F., Mutz, D., Vigna, G.: A Learning-based approach to the detection of SQL attacks. In: Proceedings of the Conference on Detection of Intrusions and Malware and Vulnerability Assessment, Vienna, pp. 123–140 (2005)
53. Wang, Y., Li, Z.: SQL injection detection with composite kernel in support vector machine. Int. J. Secur. Appl. **6**(2), 191 (2012)
54. Romi Rawat, R., Kumar Shrivastav, S.: SQL injection attack detection using SVM. Int. J. Comput. Appl. **42**(13), 0975–8887 (2012)
55. Huang, Z., Hong Cheon, E.: An approach to prevention of SQL injection attack based on machine learning. In: Proceedings of the First Yellow Sea International Conference on Ubiquitous Computing, Weihai (2011)
56. Hong Cheon, E., Huang, Z., Sik Lee, Y.: Preventing SQL injection attack based on machine learning. Int. J. Adv. Comput. Technol. **5**(9), (2013). doi:10.4156/ijact.vol5.issue9.115
57. Thorpe, S.J., Arnaud, D., van Rullen, R.: Spike-based strategies for rapid processing. Neural Netw. **14**(6–7), 715–725 (2001)
58. Delorme A., Perrinet L., Thorpe S.J., Networks of integrate-and-fire neurons using rank order coding b: spike timing dependant plasticity and emergence of orientation selectivity. Neurocomputing **38–40**(1–4), 539–545 (2000)
59. Thorpe, S.J., Gautrais, J.: Rank order coding. In: CNS '97: Proceding of the 6th Annual Conference on Computational Neuroscience: Trends in Research, pp. 113–118. Plenum Press, New York (1998)
60. Nikola, K.: Evolving Connectionist Systems: The Knowledge Engineering Approach. Springer, New York (2006)
61. Schliebs, S., Defoin-Platel, M., Kasabov, N.: Integrated feature and parameter optimization for an evolving spiking neural network. In: 15th International Conference, ICONIP 2008. Lecture Notes in Computer Science, vol. 5506, pp. 1229–1236, 25–28 Nov 2008. Springer, New York (2009)
62. Shrivastava, S., Singh, M.P.: Performance evaluation of feed-forward neural network with soft computing techniques for hand written English alphabets. Appl. Soft Comput. **11**(1), 1156–1182 (2011)
63. Shao, Y.E., Hsu, B.-S.: Determining the contributors for a multivariate SPC chart signal using artificial neural networks and support vector machine. J. ICIC **5**(12(B)), 4899–4906 (2009)
64. Chou, P.-H., Hsu, C.-H., Wu, C.-F., Li, P.-H., Wu, M.-J.: Application of back-propagation neural network for e-commerce customers patterning. ICIC Express Lett. **3**(3(B)), 775–785 (2009)
65. He, C., Li, H., Wang, B., Yu, W., Liang, X.: Prediction of compressive yield load for metal hollow sphere with crack based on artificial neural network. ICIC Express Lett. **3**(4(B)), 1263–1268 (2009)
66. Wu, J.K., Kang, J., Chen, M.H., Chen, G.T.: Fuzzy neural network model based on particle swarm optimization for short-term load forecasting. In: Proceedings of CSU-EPSA **19**(1), 63–67 (2007)
67. Li, D.K., Zhang, H.X., Li, S.A.: Development cost estimation of aircraft frame based on BP neural networks. FCCC **31**(9), 27–29 (2006)

68. Karimi, B., Menhaj, M.B., Saboori, I.: Multilayer feed forward neural networks for controlling decentralized large-scale non-affine nonlinear systems with guaranteed stability. Int. J. Innov. Comput. Inf. Control **6**(11), 4825–4841 (2010)
69. ZareNezhad, B., Aminian, A.: A multi-layer feed forward neural network model for accurate prediction of fue gas sulfuric acid dew points in process industries. Appl. Therm. Eng. **30**(6–7), 692–696 (2010)
70. Huang, L., Song, Q., Kasabov, N.: Evolving connectionist system based role allocation for robotic soccer. Playing, Intelligent Control, 2005. Proceedings of the IEEE International Symposium on (2005). Mediterrean Conference on Control and Automation (2005). DOI:10.1109/.2005.1466988
71. Kasabov, N.: Evolving fuzzy neural networks for on-line supervised/ unsupervised, knowledge–based learning. IEEE Trans. Cybern. **31**(6), 902–918 (2001)
72. Song, Q., Kasabov, N.: Weighted data normalization and feature selection. In: Proceedings 8th Intelligence Information Systems Conference (2003)
73. Kasabov, N., Song Q.: GA-parameter optimization of evolving connectionist systems for classification and a case study from bioinformatics. In: 9th Conference on Neural Information ICONIP '02, IEEE ICONIP. 1198128 (2002)
74. Vlassis, N.: A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence. Morgan and Claypool Publishers, San Rafael (2008). ISBN: 978-1-59829-526-9
75. Stolfo Salvatore, J., Wei, F., Lee, W., Andreas, P., Chan, P.K.: Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection: results from the JAM project. In: Proceedings of DARPA Information Survivability Conference and Exposition, DISCEX '00 (2000)
76. Jeff, H.: Introduction to Neural Networks with Java, 1st edn. (2008). ISBN: 097732060X
77. Goh, L., Song, Q., Kasabov, N.: A novel feature selection method to improve classification of gene expression data. In: 2nd Asia-Pacific IT Conference, vol. 29 (2004)
78. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
79. Zwillinger, D., Kokoska, S.: CRC Standard Probability and Statistics Tables and Formulae, CRC Press Print (1999). ISBN: 978-1-58488-059-2, eBook ISBN: 978-1-4200-5026-4
80. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: J. Artif. Intell. Res., **16**(1), 321–357 (2002)

Integral Estimates for the Composition of Green's and Bounded Operators

Shusen Ding and Yuming Xing

Abstract We develop integral estimates for the composition of Green's operator and a bounded operator applied to differential forms. These results give the higher integrability of this composite operator.

Keywords: Green's operator • Differential forms • Norm inequalities • A-harmonic equations

1 Introduction

Let G be Green's operator and F be a general bounded operator defined on the space of smooth differential forms. The purpose of this chapter is to establish some integral estimates for the composition $G \circ F$ of Green's operator G and the bounded operator F. Many different versions of L^p-norm inequalities and estimates for operators and their compositions have been developed during the recent years, see [1–8]. These results provide effective tools for some areas of mathematics, including partial differential equations, harmonic analysis, and operator theory. The estimates obtained in this chapter will give the higher integrability of the composite operator $G \circ F$. Specifically, we will estimate the L^s norm of $G \circ F(u)$ in terms of L^p norm of u, where u is a smooth differential form and the integral exponent s of $G \circ F(u)$ could be much larger than p, the integral exponent of the smooth differential form u.

In this chapter, we keep using the same notations appearing in [1]. Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$ be a domain with $|\Omega| < \infty$, B and σB be the balls with the same center and $\text{diam}(\sigma B) = \sigma \text{diam}(B)$. We do not distinguish the balls from cubes in this chapter. Differential forms are extensions of differentiable functions in $\mathbb{R}^n$. A function $u(x_1, x_2, \cdots, x_n)$ is called a 0-form. A k-form $u(x)$ is generated by $\{dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}\}$, $k = 1, 2, \cdots, n$, that is,

S. Ding (✉)
Department of Mathematics, Seattle University, Seattle, WA 98122, USA
e-mail: sding@seattleu.edu

Y. Xing
Department of Mathematics, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China
e-mail: xyuming@hit.edu.cn

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_8

$$u(x) = \sum_I \omega_I(x)dx_I = \sum \omega_{i_1 i_2 \cdots i_k}(x)dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k},$$

where $I = (i_1, i_2, \cdots, i_k)$, $1 \le i_1 < i_2 < \cdots < i_k \le n$. If all coefficients functions $\omega_{i_1 i_2 \cdots i_k}(x)$ are differentiable, then $u(x)$ is called a differential k-form. Much progress has been made for differential forms satisfying some versions of harmonic equations, see [9–12] for example. Let $\wedge^l = \wedge^l(\mathbb{R}^n)$ be the set of all l-forms in $\mathbb{R}^n$, $D'(\Omega, \wedge^l)$ be the space of all differential l-forms in Ω, and $L^p(\Omega, \wedge^l)$ be the l-forms $u(x) = \sum_I u_I(x)dx_I$ in Ω satisfying $\int_\Omega |u_I|^p < \infty$ for all ordered l-tuples I, $l = 1, 2, \cdots, n$. We denote the exterior derivative by d and the Hodge star operator by $\star$. The Hodge codifferential operator $d^\star$ is given by $d^\star = (-1)^{nl+1} \star d \star$, $l = 1, 2, \cdots, n$. For $u \in D'(\Omega, \wedge^l)$, the vector-valued differential form

$$\nabla u = \left(\frac{\partial u}{\partial x_1}, \cdots, \frac{\partial u}{\partial x_n} \right)$$

consists of differential forms $\frac{\partial u}{\partial x_i} \in D'(\Omega, \wedge^l)$, where the partial differentiation is applied to the coefficients of u. Let $|E|$ be the n-dimensional Lebesgue measure of a set $E \subseteq \mathbb{R}^n$. For a function u, the average of u over B is defined by $u_B = \frac{1}{|B|} \int_B u dx$. All integrals involved in this paper are the Lebesgue integrals. We use $C^\infty(\Omega, \wedge^l)$ to denote the space of smooth l-forms on Ω and the Green's operator G be defined on $C^\infty(\Omega, \wedge^l)$ by assigning $G(u)$ to be a solution of the Poisson's equation

$$\Delta G(u) = u - H(u),$$

where H is the harmonic projection operator, see [1] and [6] for more results about Green's operator. In this chapter, we always assume that F is any operator bounded from $L^p(\Omega, \wedge)$ to itself for some p, $1 < p < \infty$, i.e.,

$$\|F(u)\|_{p,\Omega} \le C\|u\|_{p,\Omega}$$

for $u \in \wedge$, where C is a constant independent of u. For example, F can be the potential operator P, or the homotopy operator T. For any subset $E \subset \mathbb{R}^n$ and $p > 1$, we use $W^{1,p}(E, \wedge^l)$ to denote the Sobolev space of l-forms which equals $L^p(E, \wedge^l) \cap L_1^p(E, \wedge^l)$ with norm

$$\|u\|_{W^{1,p}(E)} = \|u\|_{W^{1,p}(E,\wedge^l)} = \operatorname{diam}(E)^{-1}\|u\|_{p,E} + \|\nabla u\|_{p,E}. \tag{1}$$

From [13], we know that for each differential form u, we have the decomposition

$$u = d(Tu) + T(du). \tag{2}$$

and

$$\|\nabla(Tu)\|_{p,\Omega} \le C|\Omega|\|u\|_{p,\Omega}, \quad \text{and} \quad \|Tu\|_{p,\Omega} \le C|\Omega|\operatorname{diam}(\Omega)\|u\|_{p,\Omega}, \tag{3}$$

where T is the homotopy operator.

The nonlinear differential equation for differential forms

$$d^{\star}A(x, du) = B(x, du) \tag{4}$$

is called non-homogeneous A-harmonic equation, where $A : \Omega \times \wedge^l(\mathbb{R}^n) \to \wedge^l(\mathbb{R}^n)$ and $B : \Omega \times \wedge^l(\mathbb{R}^n) \to \wedge^{l-1}(\mathbb{R}^n)$ satisfy the conditions:

$$|A(x, \xi)| \leq a|\xi|^{p-1},\ A(x, \xi) \cdot \xi \geq |\xi|^p \ \text{ and } \ |B(x, \xi)| \leq b|\xi|^{p-1}$$

for almost every $x \in \Omega$ and all $\xi \in \wedge^l(\mathbb{R}^n)$. Here $a, b > 0$ are constants and $1 < p < \infty$ is a fixed exponent associated with (4). A solution to (4) is an element u of the Sobolev space $W^{1,p}_{\text{loc}}(\Omega, \wedge^{l-1})$ such that

$$\int_{\Omega} A(x, du) \cdot d\varphi + B(x, du) \cdot \varphi = 0$$

for all $\varphi \in W^{1,p}_{\text{loc}}(\Omega, \wedge^{l-1})$ with compact support. If u is a function (0-form) in $\mathbb{R}^n$, the Eq. (4) reduces to

$$\text{div}A(x, \nabla u) = B(x, \nabla u).$$

If the operator $B = 0$, the Eq. (4) becomes

$$d^{\star}A(x, du) = 0$$

which is called the (homogeneous) A-harmonic equation. See [1] for more recent progress made in the study of the A-harmonic equation.

2 Local Estimates

In this section, we prove the local norm inequalities for the composite operator $G \circ F$. We will need the following lemmas.

Lemma 1 ([13]). *Let $u \in D'(Q, \wedge^l)$ and $du \in L^p(Q, \wedge^{l+1})$. Then, $u - u_Q$ is in $L^{np/(n-p)}(Q, \wedge^l)$ and*

$$\left(\int_Q |u - u_Q|^{np/(n-p)} dx\right)^{(n-p)/np} \leq C_p(n) \left(\int_Q |du|^p dx\right)^{1/p} \tag{5}$$

for Q a cube or a ball in $\mathbb{R}^n$, $l = 0, 1, \cdots, n-1$, and $1 < p < n$.

Lemma 2 ([6]). *Let u be a smooth differential form defined in Ω and $1 < s < \infty$. Then, there exists a positive constant $C = C(s)$, independent of u, such that*

$$\|dd^*G(u)\|_{s,B} + \|d^*dG(u)\|_{s,B} + \|dG(u)\|_{s,B} + \|d^*G(u)\|_{s,B} + \|G(u)\|_{s,B} \leq C(s)\|u\|_{s,B} \tag{6}$$

for any ball $B \subset \Omega$.

Lemma 3 ([12]). *Let u be a solution of the A-harmonic equation (4) in a domain Ω and $0 < s, t < \infty$. Then, there exists a constant C, independent of u, such that*

$$\|u\|_{s,B} \leq C|B|^{(t-s)/st}\|u\|_{t,\sigma B} \tag{7}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Combining the second inequality in (3) and Lemma 2, we have the following Lemma 4 immediately.

Lemma 4. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, F be a bounded operator and G be Green's operator. Then, for any constant $p > 1$, there exists a constant C, independent of u, such that*

$$\|G(F(u))\|_{p,B} \leq C\|u\|_{p,B} \tag{8}$$

for all balls $B \subset \Omega$.

We first prove the following local norm inequality, which gives the higher integrability of $G \circ F$ applied to differential form u.

Theorem 1. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $1 < p < n$, and F be a bounded operator and G be Green's operator. Then, for any constant s with $0 < s < np/(n-p)$, there exists a constant C, independent of u, such that*

$$\|G(F(u)) - (G(F(u)))_B\|_{s,B} \leq C|B|^{1/s+1/n-1/p}\|u\|_{p,\sigma B} \tag{9}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof. Let $B \subset \Omega$ be any ball. For any constant $p > 1$ and any differential form u, applying Lemma 1 to $G(F(u))$, Lemma 2 and noticing that F is bounded, we obtain

$$\begin{aligned}\left(\int_B |G(F(u)) - (G(F(u)))_B|^{np/(n-p)}dx\right)^{(n-p)/np} &\leq C_1\left(\int_B |dG(F(u))|^p dx\right)^{1/p} \\ &= C_2\left(\int_B |F(u)|^p dx\right)^{1/p} \\ &\leq C_3\left(\int_B |u|^p dx\right)^{1/p}. \end{aligned} \tag{10}$$

Using the monotonic property of the L^p-norms, for any s with $0 < s < np/(n-p)$, we obtain

$$\begin{aligned}&\left(\frac{1}{|B|}\int_B |G(F(u)) - (G(F(u)))_B|^s dx\right)^{1/s} \\ &\quad\leq \left(\frac{1}{|B|}\int_B |G(F(u)) - (G(F(u)))_B|^{np/(n-p)}dx\right)^{(n-p)/np}. \end{aligned} \tag{11}$$

Combining (10) and (11) yields

$$\left(\int_B |G(F(u)) - (G(F(u)))_B|^s dx\right)^{1/s} \le C_9 |B|^{1/s+1/n-1/p} \left(\int_B |u|^p dx\right)^{1/p}.$$

We have completed the proof of Theorem 1.

We should note that the above inequality (9) can be written as the following version

$$\left(\frac{1}{|B|}\int_B |G(F(u)) - (G(F(u)))_B|^s dx\right)^{1/s} \le C|B|^{1/n} \left(\frac{1}{|B|}\int_B |u|^p dx\right)^{1/p}. \tag{12}$$

It should also be noticed that in both (9) and (12), the integral exponent s on the left-hand side could be much larger than the integral exponent p on the right-hand side since $np/(n-p) \to \infty$ as $p \to n-$, which gives the higher integrability of the composite operator $G(F(u)) - (G(F(u)))_B$ for the case $1 < p < n$. Next, we prove the similar higher integrability of $G(F(u)) - (G(F(u)))_B$ for the case $p \ge n$.

Theorem 2. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $p \ge n$, and F be a bounded operator and G be Green's operator. Then, for any $s > 0$, there exists a constant C, independent of u, such that*

$$\left(\frac{1}{|B|}\int_B |G(F(u)) - (G(F(u)))_B|^s dx\right)^{1/s} \le C|B|^{1/n} \left(\frac{1}{|B|}\int_{\sigma B} |u|^p dx\right)^{1/p} \tag{13}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Note that (12) can also be written as the norm version

$$\|G(F(u)) - (G(F(u)))_B\|_{s,B} \le C|B|^{1/s+1/n-1/p} \|u\|_{p,\sigma B} \tag{14}$$

Proof. Choose $k = \max\{1, s/p\}$ and $q = knp/(n+kp)$. Since $n - p \le 0$ now, then

$$q - p = \frac{p(k(n-p)-n)}{n+kp} < 0, \tag{15}$$

that is, $q < p$. Also, $1 < q = knp/(n+kp) < n$. Applying Lemma 1 to $G(F(u))$ and noticing the monotonic property of the L^p space, we have

$$\begin{aligned}\left(\int_B |G(F(u)) - (G(F(u)))_B|^{nq/(n-q)} dx\right)^{(n-q)/nq} &\le C_2 \left(\int_B |dG(F(u))|^q dx\right)^{1/q} \\ &= C_3 \left(\int_B |(F(u))|^q dx\right)^{1/q}\end{aligned}$$

$$\leq C_4 \left(\int_B |u|^q dx \right)^{1/q}$$
$$\leq C_5 |B|^{1/q-1/p} \left(\int_B |u|^p dx \right)^{1/p}. \tag{16}$$

Note that $nq/(n-q) = kp > s$, using the monotonic property of the L^p space again, and (16),

$$\begin{aligned}
&\left(\int_B |G(F(u)) - (G(F(u)))_B|^s dx \right)^{1/s} \\
&\leq |B|^{1/s-1/kp} \left(\int_B |G(F(u)) - (G(F(u)))_B|^{nq/(n-q)} dx \right)^{(n-q)/nq} \\
&\leq C_6 B|^{1/q-1/p+1/s-1/kp} \left(\int_B |u|^p dx \right)^{1/p} \\
&\leq C_6 B|^{1/n+1/s-1/p} \left(\int_B |u|^p dx \right)^{1/p},
\end{aligned} \tag{17}$$

that is,

$$\left(\frac{1}{|B|} \int_B |G(F(u)) - (G(F(u)))_B|^s dx \right)^{1/s} \leq C_6 |B|^{1/n} \left(\frac{1}{|B|} \int_B |u|^p dx \right)^{1/p}.$$

The proof of Theorem 2 has been completed.

Let φ be a strictly increasing convex function on $[0, \infty)$ with $\varphi(0) = 0$, and u be a differential form in a bounded domain $D \subset \mathbb{R}^n$ such that $\varphi(k(|u| + |u_D|)) \in L^1(D; \mu)$ for any real number $k > 0$ and $\mu(\{x \in D : |u - u_D| > 0\}) > 0$, where μ is a Radon measure defined by $d\mu = w(x)dx$ for a weight $w(x)$. It has been proved that for any positive constant a, it follows that

$$\int_D \varphi(\frac{1}{2}|u - u_D|) d\mu \leq C_1 \int_D \varphi(a|u|) d\mu \leq C_2 \int_D \varphi(2a|u - u_D|) d\mu, \tag{18}$$

where C_1 and C_2 are some positive constants. Choosing $\varphi(t) = t^p$, $p > 1$, $w(x) = 1$ and D to be a ball B in (18), we know that the norms $\|u\|_{p,B}$ and $\|u - u_B\|_{p,B}$ are comparable, that is,

$$\|u - u_B\|_{p,B} \leq C_1 \|u\|_{p,B} \leq C_2 \|u - u_B\|_{p,B} \tag{19}$$

for any ball B with $|\{x \in B : |u - u_D| > 0\}| > 0$. We prove the local higher integrability of $G \circ F$ in the following theorem.

Theorem 3. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $1 < p < n$, and F be a bounded operator and G be Green's operator. If $u \in L^p_{\text{loc}}(\Omega, \wedge^l)$, then $G(F(u)) \in L^s_{\text{loc}}(\Omega, \wedge^l)$ for any $0 < s < np/(n-p)$. Specifically, there exists a constant C, independent of u, such that*

$$\|G(F(u))\|_{s,B} \leq C|B|^{1/s+1/n-1/p}\|u\|_{p,\sigma B} \tag{20}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof. We may assume that the measure

$$|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| > 0.$$

Otherwise, if $|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| = 0$, then $G(F(u)) = G(F(u)))_B$ almost everywhere in B. Therefore, $G(F(u))$ is a closed form. Thus, $G(F(u))$ is a solution of the A-harmonic equation (4). By Lemma 3, for any differential form u and any $m, k > 0$, there exists a constant C_1, independent of u, such that

$$\|u\|_{m,B} \leq C_1|B|^{(k-m)/km}\|u\|_{k,\sigma B} \tag{21}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$. Particularly, for $G(F(u))$ and s, p appearing in our theorem, we have

$$\|G(F(u))\|_{s,B} \leq C_1|B|^{(p-s)/sp}\|G(F(u))\|_{p,\sigma B}. \tag{22}$$

From (22) and Lemma 4, we have

$$\begin{aligned}
\|G(F(u))\|_{s,B} &\leq C_1|B|^{(p-s)/sp}\|G(F(u))\|_{p,\sigma B} \\
&\leq C_2|B|^{(p-s)/sp}|B|^{1+1/n}\|u\|_{p,\sigma B} \\
&= C_2|B|^{1/s+1/n-1/p}|B|\|u\|_{p,\sigma B} \\
&= C_2|B|^{1/s+1/n-1/p}|\Omega|\|u\|_{p,\sigma B} \\
&= C_3|B|^{1/s+1/n-1/p}\|u\|_{p,\sigma B},
\end{aligned} \tag{23}$$

that is, inequality (20) holds. Next, we may assume that

$$|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| > 0.$$

Hence, we can use (18). Choosing $\varphi(t) = t^{np/(n-p)}$ in (18), we find that for any differential form v

$$\left(\int_B |v|^{np/(n-p)}dx\right)^{(n-p)/np} \leq C_4\left(\int_B |v - v_B|^{np/(n-p)}dx\right)^{(n-p)/np}. \tag{24}$$

Replacing v by $G(F(u))$ in (24) it follows that

$$\left(\int_B |G(F(u))|^{np/(n-p)}dx\right)^{(n-p)/np} \leq C_5\left(\int_B |G(F(u)) - (G(F(u)))_B|^{np/(n-p)}dx\right)^{(n-p)/np}. \tag{25}$$

By the monotonic property of the L^p-space, for any s with $0 < s < np/(n-p)$, we obtain

$$\left(\frac{1}{|B|}\int_B |G(F(u))|^s dx\right)^{1/s} \leq \left(\frac{1}{|B|}\int_B |G(F(u))|^{np/(n-p)}dx\right)^{(n-p)/np}. \tag{26}$$

Combining (26), (25) and using (10), we find that

$$\begin{aligned}&\left(\frac{1}{|B|}\int_B |G(F(u))|^s dx\right)^{1/s}\\ &\leq \left(\frac{1}{|B|}\int_B |G(F(u))|^{np/(n-p)}dx\right)^{(n-p)/np}\\ &\leq C_6\left(\frac{1}{|B|}\int_B |G(F(u)) - (G(F(u)))_B|^{np/(n-p)}dx\right)^{(n-p)/np}\\ &\leq C_7|B|^{1/n}\left(\frac{1}{|B|}\int_B |u|^p dx\right)^{1/p},\end{aligned} \tag{27}$$

that is,

$$\left(\int_B |G(F(u))|^s dx\right)^{1/s} \leq C|B|^{1/s+1/n-1/p}\left(\int_B |u|^p dx\right)^{1/p}. \tag{28}$$

The above inequality (28) shows that if $u \in L^p_{\mathrm{loc}}(\Omega, \wedge^l)$, then $G(F(u)) \in L^s_{\mathrm{loc}}(\Omega, \wedge^l)$. We have completed the proof of Theorem 3.

It should be noticed that the inequality (20) can be written as the following version

$$\left(\frac{1}{|B|}\int_B |G(F(u))|^s dx\right)^{1/s} \leq C|B|^{1/n}\left(\frac{1}{|B|}\int_B |u|^p dx\right)^{1/p}. \tag{29}$$

As we noticed earlier, in both (20) and (29), the integral exponent s on the left-hand side could be much larger than the integral exponent p on the right-hand side because of $np/(n-p) \to \infty$ as $p \to n-$, which gives the higher integrability of the composite operator $G \circ F$ for the case $1 < p < n$. In the following theorem, we prove the higher integrability of $G \circ F$ for the case $p \geq n$.

Theorem 4. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $p \geq n$, and F be a bounded operator and G be Green's operator. If $u \in L^p_{\text{loc}}(\Omega, \wedge^l)$, then $G(F(u)) \in L^s_{\text{loc}}(\Omega, \wedge^l)$ for any $s > 0$. Moreover, there exists a constant C, independent of u, such that*

$$\left(\frac{1}{|B|}\int_B |G(F(u))|^s dx\right)^{1/s} \leq C|B|^{1/n}\left(\frac{1}{|B|}\int_{\sigma B} |u|^p dx\right)^{1/p} \tag{30}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Notice that (30) can also be written as the norm version

$$\|G(F(u))\|_{s,B} \leq C|B|^{1/s+1/n-1/p}\|u\|_{p,\sigma B} \tag{31}$$

Proof. First, assume that the measure $|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| = 0$. By the same method developed in the proof of Theorem 3, we can show inequality (30) holds for any ball B with $\sigma B \subset \Omega$ for some $\sigma > 1$. Next, we assume that $|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| > 0$. Set $k = \max\{1, s/p\}$ and $q = knp/(n + kp)$. Since $n - p \leq 0$ now, then

$$q - p = \frac{p(k(n-p) - n)}{n + kp} < 0, \tag{32}$$

that is, $q < p$. Also, $1 < q = knp/(n + kp) < n$. Applying Lemma 1 to $G(F(u))$ and from the monotonic property of the L^p norm, we find that

$$\begin{aligned}
&\left(\int_B |G(F(u)) - (G(F(u)))_B|^{nq/(n-q)} dx\right)^{(n-q)/nq} \\
&\leq C_2\left(\int_B |dG(F(u))|^q dx\right)^{1/q} \\
&= C_3\left(\int_B |(F(u))_B|^q dx\right)^{1/q} \\
&\leq C_4\left(\int_B |F(u)|^q dx\right)^{1/q} \\
&\leq C_5\left(\int_B |u|^q dx\right)^{1/q} \\
&\leq C_5|B|^{1/q-1/p}\left(\int_B |u|^p dx\right)^{1/p}.
\end{aligned} \tag{33}$$

Since the measure $|\{x \in B : |G(F(u)) - (G(F(u)))_B| > 0\}| > 0$ now, then we can use (18). Choosing $\varphi(t) = t^{nq/(n-q)}$ in (18), we find that for any differential form ω

$$\left(\int_B |\omega|^{nq/(n-q)} dx\right)^{(n-q)/nq} \leq C_6\left(\int_B |\omega - \omega_B|^{nq/(n-q)} dx\right)^{(n-q)/nq}. \tag{34}$$

Replacing ω by $G(F(u))$ in (34), we have

$$\left(\int_B |G(F(u))|^{nq/(n-q)}dx\right)^{(n-q)/nq}$$
$$\leq C_7\left(\int_B |G(F(u)) - (G(F(u)))_B|^{nq/(n-q)}dx\right)^{(n-q)/nq}. \tag{35}$$

Note that $nq/(n-q) = kp > s$, using the monotonic property of the L^p norm again, (35) and (33),

$$\begin{aligned}
&\left(\int_B |G(F(u))|^s dx\right)^{1/s}\\
&\leq |B|^{1/s-1/kp}\left(\int_B |G(F(u))|^{nq/(n-q)}dx\right)^{(n-q)/nq}\\
&\leq C_7|B|^{1/s-1/kp}\left(\int_B |G(F(u)) - (G(F(u)))_B|^{nq/(n-q)}dx\right)^{(n-q)/nq}\\
&\leq C_7 B|^{1/q-1/p+1/s-1/kp}\left(\int_B |u|^p dx\right)^{1/p}\\
&\leq C_7 B|^{1/n+1/s-1/p}\left(\int_B |u|^p dx\right)^{1/p},
\end{aligned} \tag{36}$$

that is,

$$\left(\frac{1}{|B|}\int_B |G(F(u))|^s dx\right)^{1/s} \leq C_5|B|^{1/n}\left(\frac{1}{|B|}\int_B |u|^p dx\right)^{1/p}.$$

The proof of Theorem 4 has been completed.

Now, we prove the higher order imbedding theorems of the composite operator $G\circ F$ in the following theorem.

Theorem 5. *Let $u \in D'(\Omega, \wedge^l)$ be a solution of A-harmonic equation (4), $l = 1, 2, \cdots, n$, and F be a bounded operator and G be Green's operator. Then, or any constant $s > 0$, there exists a constant C, independent of u, such that*

$$\|G(F(u)) - (G(F(u)))_B\|_{W^{1,s}(B)} \leq C|B|^{1+1/s-1/p}\|u\|_{p,\sigma B} \tag{37}$$

for all balls B with $\sigma B \subset \Omega$ for some $\sigma > 1$.

Proof. For any differential form v, the decomposition

$$v = Tdv + dTv \tag{38}$$

holds. Applying (38) to differential form $G(F(u))$

$$G(F(u)) = dT(G(F(u))) + Td(G(F(u))). \tag{39}$$

Noticing the fact that $dT(G(F(u))) = (G(F(u)))_B$ for any differential form u, and using (39), (3), Lemma 2 and Lemma 3, we obtain

$$\begin{aligned}
&\|G(F(u)) - (G(F(u)))_B\|_{W^{1,s}(B)} \\
&\quad = \|Td(G(F(u)))\|_{W^{1,s}(B)} \\
&\quad = (\mathrm{diam}(B))^{-1}\|Td(G(F(u)))\|s, B + \|\nabla Td(G(F(u)))\|s, B \\
&\quad \le (\mathrm{diam}(B))^{-1}C_1|B|\mathrm{diam}(B)\|dG(F(u))\|s, B + C_2|B|\|dG(F(u))\|s, B \\
&\quad \le C_3|B|\|F(u)\|s, B \\
&\quad = C_4|B|\|u\|s, B \\
&\quad \le C_5|B|^{1+1/s-1/p}\|u\|_{p,B}.
\end{aligned} \tag{40}$$

We have completed the proof of Theorem 5.

3 Global Estimates

In this section, we prove the global higher integrability theorems for the composition $G \circ F$.

Lemma 5. *Each domain Ω has a modified Whitney cover of cubes $\mathcal{V} = \{Q_i\}$ such that*

$$\cup_i Q_i = \Omega, \ \sum_{Q_i \in \mathcal{V}} \chi_{\sqrt{\frac{5}{4}}Q_i} \le N\chi_\Omega \tag{41}$$

and some $N > 1$, and if $Q_i \cap Q_j \neq \emptyset$, then there exists a cube R (this cube need not be a member of $\mathcal{V}$) in $Q_i \cap Q_j$ such that $Q_i \cup Q_j \subset NR$. Moreover, if Ω is δ-John, then there is a distinguished cube $Q_0 \in \mathcal{V}$ which can be connected with every cube $Q \in \mathcal{V}$ by a chain of cubes $Q_0, Q_1, \cdots, Q_k = Q$ from $\mathcal{V}$ and such that $Q \subset \rho Q_i$, $i = 0, 1, 2, \cdots, k$, for some $\rho = \rho(n, \delta)$.

Theorem 6. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $1 < p < n$, and F be a bounded operator and G be Green's operator. If $u \in L^p(\Omega, \wedge^l)$, then $G(F(u)) \in L^s(\Omega, \wedge^l)$ for any $0 < s < np/(n-p)$. Moreover, there exist constants C_1 and C_2, independent of u, such that*

$$\|G(F(u))\|_{s,\Omega} \le C_1|\Omega|^{1/s+1/n-1/p}\|u\|_{p,\Omega}, \tag{42}$$

and

$$\|G(F(u)) - (G(F(u)))_{\Omega}\|_{s,\Omega} \leq C|\Omega|^{1/s+1/n-1/p}\|u\|_{p,\Omega} \tag{43}$$

for any bounded and convex domain $\Omega \subset \mathbb{R}^n$.

Proof. First, we prove that inequality (42) holds. Using the Cover Lemma and Theorem 3, and noticing $1/s + 1/n - 1/p > 0$ since $0 < s < np/(n-p)$, we have

$$\begin{aligned}
\|G(F(u))\|_{s,\Omega} &\leq \sum_{B\in\mathcal{V}} \|G(F(u))\|_{s,B} \\
&\leq \sum_{B\in\mathcal{V}} \left(C_1|B|^{1/s+1/n-1/p}\|u\|_{p,\sigma B}\right) \\
&\leq \sum_{B\in\mathcal{V}} \left(C_1|\Omega|^{1/s+1/n-1/p}\|u\|_{p,\sigma B}\right) \\
&\leq C_2|\Omega|^{1/s+1/n-1/p}N\|u\|_{p,\Omega} \\
&\leq C_3|\Omega|^{1/s+1/n-1/p}\|u\|_{p,\Omega},
\end{aligned} \tag{44}$$

that is, inequality (42) holds. Next, we show that inequality (43) also holds. It is well known that for any differential form ω and any bounded and convex domain D, we have

$$\|\omega_D\|_{s,D} \leq C_4\|\omega\|_{s,D}. \tag{45}$$

Replacing ω by $G(F(u))$ and D by Ω in (45), we obtain

$$\|(G(F(u)))_{\Omega}\|_{s,\Omega} \leq C_4\|G(F(u))\|_{s,\Omega}. \tag{46}$$

Hence, using (46) and (44), we have

$$\begin{aligned}
\|G(F(u)) - (G(F(u)))_{\Omega}\|_{s,\Omega} &\leq \|G(F(u))\|_{s,\Omega} + \|(G(F(u)))_{\Omega}\|_{s,\Omega} \\
&\leq \|G(F(u))\|_{s,\Omega} + C_5\|G(F(u))\|_{s,\Omega} \\
&\leq (1 + C_5)\|G(F(u))\|_{s,\Omega} \\
&\leq C_6|\Omega|^{1/s+1/n-1/p}\|u\|_{p,\Omega},
\end{aligned}$$

which indicates that (43) holds. The proof of Theorem 6 has been completed.

We have proved the global higher integrability of $G \circ F$ for the case $1 < p < n$. Using Theorem 4 and the same method as we did in the proof of Theorem 6, we can prove the global higher integrability of $G \circ F$ for the case $p \geq n$ in the following theorem.

Theorem 7. *Let $u \in D'(\Omega, \wedge^l)$, $l = 1, 2, \cdots, n$, $p \geq n$, and F be a bounded operator and G be Green's operator. If $u \in L^p(\Omega, \wedge^l)$, then $G(F(u)) \in L^s(\Omega, \wedge^l)$ for any $s > 0$. Moreover, there exist constants C_1 and C_2, independent of u, such that*

$$\|G(F(u))\|_{s,\Omega} \leq C_1 |\Omega|^{1/s+1/n-1/p} \|u\|_{p,\Omega}, \tag{47}$$

and

$$\|G(F(u)) - (G(F(u)))_\Omega\|_{s,\Omega} \leq C|\Omega|^{1/s+1/n-1/p} \|u\|_{p,\Omega} \tag{48}$$

for any bounded and convex domain $\Omega \subset \mathbb{R}^n$.

Remark. (1) We only extend some local results to the global cases. We can also extend the other local results to the global versions. Considering the length of the chapter, we only include a few global versions here. (2) We think that the global results can be proved in more general domains, such as the L^p-averaging domains [14] and $L^\varphi(\mu)$-averaging domains [15]. (3) Our norm inequalities can be extended into the weighted cases by routine procedure.

References

1. Agarwal, R.P., Ding, S., Nolder, C.A.: Inequalities for Differential Forms. Springer, New York (2009)
2. Xing, Y., Wu, C.: Global weighted inequalities for operators and harmonic forms on manifolds. J. Math. Anal. Appl. **294**, 294–309 (2004)
3. Shi, G., Xing, Y., Sun, B.: Poincaré-type inequalities for the composite operator in L^A-averaging domains. Abstr. Appl. Anal. **2014**, Article ID 675464 (2014)
4. Bi, H., Ding, S.: Some strong (p, q)-type inequalities for the homotopy operator. Comput. Math. Appl. **62**, 1780–1789 (2011)
5. Ding, S., Liu, B.: Global estimates for singular integrals of the composite operator. Ill. J. Math. **53**, 1173–1185 (2009)
6. Scott, C.: L^p-theory of differential forms on manifolds. Trans. Am. Math. Soc. **347**, 2075–2096 (1995)
7. Fang, R.: Poincaré inequalities for composition operators with norm. Abstr. Appl. Anal. **2014**, Article ID 818201 (2014)
8. Li, X., Wang, Y., Xing, Y.: Norm comparison estimates for the composite operator. J. Funct. Spaces **2014**, Article ID 943986 (2014)
9. Ding, S.: Two-weight Caccioppoli inequalities for solutions of nonhomogeneous A-harmonic equations on Riemannian manifolds. Proc. Am. Math. Soc. **132**, 2367–2375 (2004)
10. Wang, Y., Wu, C.: Sobolev imbedding theorems and Poincaré inequalities for Green's operator on solutions of the nonhomogeneous A-harmonic equation. Comput. Math. Appl. **47**, 1545–1554 (2004)
11. Xing, Y.: Weighted integral inequalities for solutions of the A-harmonic equation. J. Math. Anal. Appl. **279**, 350–363 (2003)
12. Nolder, C.A.: Hardy-Littlewood theorems for A-harmonic tensors. Ill. J. Math. **43**, 613–631 (1999)

13. Iwaniec, T., Lutoborski, A.: Integral estimates for null Lagrangians. Arch. Ration. Mech. Anal. **125**, 25–79 (1993)
14. Staples, S.G.: L^p-averaging domains and the Poincaré inequality. Ann. Acad. Sci. Fenn. Ser. A I Math. **14**, 103–127 (1989)
15. Ding, S.: $L^\varphi(\mu)$-averaging domains and the quasihyperbolic metric. Comput. Math. Appl. **47**, 1611–1618 (2004)

A Survey of Reverse Inequalities for f-Divergence Measure in Information Theory

S.S. Dragomir

Abstract In this paper we survey some discrete inequalities for the f-divergence measure in Information Theory by the use of recent reverses of the celebrated Jensen's inequality. Applications in connection with Hölder's inequality and for particular measures such as Kullback–Leibler divergence measure, Hellinger discrimination, χ^2-distance and variation distance are provided as well.

Keywords: Convex functions • Jensen's inequality • Reverse of Jensen's inequality • Reverse of Hölder's inequality • f-Divergence measure • Kullback–Leibler divergence measure • Hellinger discrimination • χ^2-Distance • Variation distance • Grüss' type inequality

1 Introduction

Given a convex function $f : \mathbb{R}_+ \to \mathbb{R}_+$, the *$f$-divergence functional,* or *f-divergence measure*

$$I_f(p,q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) \tag{1}$$

was introduced by Csiszár in [13, 14] as a generalized measure of information, a "distance function" on the set of probability distributions $\mathbb{P}^n$.

The restriction to discrete distributions is only for convenience, similar results hold for more general distributions.

The definition (1) can be extended for nonconvex function, however in this case the positivity property of $I_f(p,q)$ is not always assured.

S.S. Dragomir (✉)
School of Engineering and Science, Victoria University, PO Box 14428, Melbourne City, VIC 8001, Australia

School of Computational and Applied Mathematics, University of the Witwatersrand, Private Bag 3, Johannesburg 2050, South Africa
e-mail: Sever.Dragomir@vu.edu.au

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_9

As in Csiszár [14], we interpret the following, otherwise undefined expressions, as indicated:

$$f(0) = \lim_{t\to 0+} f(t), \quad 0f\left(\frac{0}{0}\right) = 0,$$

$$0f\left(\frac{a}{0}\right) = \lim_{\varepsilon\to 0+} f\left(\frac{a}{\varepsilon}\right) = a \lim_{t\to\infty} \frac{f(t)}{t}, \quad a > 0.$$

The immediately following results were essentially given by Csiszár and Körner [15].

Theorem 1 (Csiszár and Körner [15]). *If $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, then $I_f(p,q)$ is jointly convex in p and q.*

The following lower bound for the f-divergence functional also holds.

Theorem 2 (Csiszár and Körner [15]). *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be convex, then for every $p, q \in \mathbb{R}_+^n$, we have the inequality:*

$$I_f(p,q) \geq \sum_{i=1}^{n} q_i f\left(\frac{\sum_{i=1}^n p_i}{\sum_{i=1}^n q_i}\right). \tag{2}$$

If f is strictly convex, equality holds in (2) iff

$$\frac{p_1}{q_1} = \frac{p_2}{q_2} = \cdots = \frac{p_n}{q_n}. \tag{3}$$

Corollary 1. *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be convex and normalized, i.e.,*

$$f(1) = 0, \tag{4}$$

then, for any $p, q \in \mathbb{R}_+^n$ with

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i, \tag{5}$$

we have the inequality,

$$I_f(p,q) \geq 0. \tag{6}$$

If f is strictly convex, equality holds in (6) iff $p_i = q_i$ for all $i \in \{1,\ldots,n\}$.

In particular, if p, q are probability vectors, then (5) is assured. Corollary 1 then shows that, for strictly convex and normalized $f : \mathbb{R}_+ \to \mathbb{R}$,

$$I_f(p,q) \geq 0 \quad \text{for all } p, q \in \mathbb{P}^n. \tag{7}$$

The equality holds in (7) iff $p = q$.

These are "distance properties", however, I_f is not a metric since it violates the triangle inequality, and is asymmetric, i.e., for general $p, q \in \mathbb{R}_+^n$, $I_f(p,q) \neq I_f(q,p)$.

2 Some Examples

In the examples below we obtain, for suitable choices of the kernel f, some of the best known distance functions I_f used in mathematical statistics, information theory and signal processing, see [1–12, 16, 32–52, 52–60] and [65–92].

Example 1 (Kullback–Leibler). For

$$f(t) := t \log t, \ t > 0 \tag{8}$$

the f-divergence is

$$I_f(p,q) = KL(p,q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right), \tag{9}$$

called the *Kullback–Leibler distance* [63, 64].

Example 2 (Hellinger). Let

$$f(t) = \frac{1}{2}\left(1 - \sqrt{t}\right)^2, \ t > 0. \tag{10}$$

Then I_f gives the *Hellinger distance* [70]

$$I_f(p,q) = h^2(p,q) = \frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{p_i} - \sqrt{q_i}\right)^2, \tag{11}$$

which is symmetric.

Example 3 (Renyi). For $\alpha > 1$, let

$$f(t) = t^{\alpha}, \ t > 0. \tag{12}$$

Then

$$I_f(p,q) = D_{\alpha}(p,q) = \sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha}, \tag{13}$$

which is the *α-order entropy* [80].

Example 4 (χ^2-Distance). Let

$$f(t) = (t-1)^2, \ t > 0. \tag{14}$$

Then

$$I_f(p,q) = D_{\chi^2}(p,q) = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i} \tag{15}$$

$$= \sum_{i=1}^{n} \frac{p_i^2}{q_i} - 2P_n + Q_n$$

$$\left(= \sum_{i=1}^{n} \frac{p_i^2 - q_i^2}{q_i} \quad \text{if } P_n = Q_n \right)$$

is the χ^2-*distance between* p and q, where $P_n = \sum_{i=1}^{n} p_i$ and $Q_n = \sum_{i=1}^{n} q_i$.

Finally, we have

Example 5 (Variation Distance). Let $f(t) = |t-1|, \ t > 0$. The corresponding f-divergence, called the *variation distance*, is symmetric,

$$V(p,q) = \sum_{i=1}^{n} |p_i - q_i|.$$

For other examples of divergence measures, see the paper [61] by J.N. Kapur, where further references are given (see also [62]).

For other examples of divergence measures and further references, see [61] and [85].

In this paper we survey some discrete inequalities for the f-divergence measure in Information Theory by the use of recent reverses of the celebrated Jensen's inequality. Applications in connection with Hölder's inequality and for particular measures such as Kullback–Leibler divergence measure, Hellinger discrimination, χ^2-distance and variation distance are provided as well.

3 A Reverse Inequality Due to Dragomir and Ionescu

If $x_i, y_i \in \mathbb{R}$ and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i = 1$, then we may consider the *Čebyšev functional*

$$T_w(x,y) := \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i. \tag{16}$$

The following result is known in the literature as the *Grüss inequality*

$$|T_w(x,y)| \leq \frac{1}{4}(\Gamma-\gamma)(\Delta-\delta), \tag{17}$$

provided

$$-\infty < \gamma \leq x_i \leq \Gamma < \infty, \quad -\infty < \delta \leq y_i \leq \Delta < \infty \tag{18}$$

for $i = 1, \ldots, n$.

The constant $\frac{1}{4}$ is sharp in the sense that it cannot be replaced by a smaller constant.

If we assume that $-\infty < \gamma \leq x_i \leq \Gamma < \infty$ for $i = 1, \ldots, n$, then by the Grüss inequality for $y_i = x_i$ and by the Schwarz's discrete inequality, we have

$$\sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right| \leq \left[\sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{j=1}^{n} w_j x_j \right)^2 \right]^{\frac{1}{2}} \leq \frac{1}{2}(\Gamma-\gamma). \tag{19}$$

In order to provide a reverse of the celebrated Jensen's inequality for convex functions, S.S. Dragomir obtained in 2002 [28] the following result:

Theorem 3. *Let $f : [m,M] \to \mathbb{R}$ be a differentiable convex function on (m,M). If $x_i \in [m,M]$ and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i = 1$, then one has the counterpart of Jensen's weighted discrete inequality:*

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i \right) \\ &\leq \sum_{i=1}^{n} w_i f'(x_i) x_i - \sum_{i=1}^{n} w_i f'(x_i) \sum_{i=1}^{n} w_i x_i \\ &\leq \frac{1}{2}\left[f'(M) - f'(m) \right] \sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right|. \end{aligned} \tag{20}$$

Remark 1. We notice that the inequality between the first and the second term in (20) was proved in 1994 by Dragomir and Ionescu, see [49].

On making use of (19), we can state the following string of reverse inequalities

$$0 \leq \sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i \right) \tag{21}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} w_i f'(x_i)\, x_i - \sum_{i=1}^{n} w_i f'(x_i) \sum_{i=1}^{n} w_i x_i \\
&\leq \frac{1}{2}\left[f'(M) - f'(m)\right] \sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right| \\
&\leq \frac{1}{2}\left[f'(M) - f'(m)\right] \left[\sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{j=1}^{n} w_j x_j \right)^2 \right]^{\frac{1}{2}} \\
&\leq \frac{1}{4}\left[f'(M) - f'(m)\right](M-m),
\end{aligned}
$$

provided that $f : [m, M] \subset \mathbb{R} \to \mathbb{R}$ is a differentiable convex function on (m, M), $x_i \in [m, M]$ and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i = 1$.

Remark 2. We notice that the inequality between the first, second, and last term from (21) was proved in the general case of positive linear functionals in 2001 by S.S. Dragomir in [24].

For various Jensen's type inequalities, see [17–51].

4 Further Reverse Inequalities

The following reverse of the Jensen's inequality holds:

Theorem 4 (Dragomir [43]). *Let $f : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$, $\mathring{I}$ is the interior of I. If $x_i \in [m, M]$ and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i = 1$, then*

$$
\begin{aligned}
0 &\leq \sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i \right) \qquad (22) \\
&\leq \frac{\left(M - \sum_{i=1}^{n} w_i x_i\right)\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M-m} \Psi_f\left(\sum_{i=1}^{n} w_i x_i; m, M \right) \\
&\leq \frac{\left(M - \sum_{i=1}^{n} w_i x_i\right)\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M-m} \sup_{t \in (m,M)} \Psi_f(t; m, M) \\
&\leq \left(M - \sum_{i=1}^{n} w_i x_i \right)\left(\sum_{i=1}^{n} w_i x_i - m \right) \frac{f'_-(M) - f'_+(m)}{M-m} \\
&\leq \frac{1}{4}(M-m)\left[f'_-(M) - f'_+(m)\right],
\end{aligned}
$$

where $\Psi_f(\cdot; m, M) : (m, M) \to \mathbb{R}$ *is defined by*

$$\Psi_f(t; m, M) = \frac{f(M) - f(t)}{M - t} - \frac{f(t) - f(m)}{t - m}.$$

We also have the inequality

$$\begin{aligned} 0 &\le \sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i\right) \\ &\le \frac{\left(M - \sum_{i=1}^{n} w_i x_i\right)\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M - m} \Psi_f\left(\sum_{i=1}^{n} w_i x_i; m, M\right) \\ &\le \frac{1}{4}(M - m)\,\Psi_f\left(\sum_{i=1}^{n} w_i x_i; m, M\right) \\ &\le \frac{1}{4}(M - m) \sup_{t \in (m,M)} \Psi_f(t; m, M) \\ &\le \frac{1}{4}(M - m)\left[f'_-(M) - f'_+(m)\right], \end{aligned} \tag{23}$$

provided that $\sum_{i=1}^{n} w_i x_i \in (m, M)$.

Proof. By the convexity of f we have that

$$\begin{aligned} &\sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i\right) \\ &= \sum_{i=1}^{n} w_i f\left[\frac{m(M - x_i) + M(x_i - m)}{M - m}\right] \\ &\quad - f\left(\sum_{i=1}^{n} w_i \left[\frac{m(M - x_i) + M(x_i - m)}{M - m}\right]\right) \\ &\le \sum_{i=1}^{n} w_i \frac{(M - x_i) f(m) + (x_i - m) f(M)}{M - m} \\ &\quad - f\left(\frac{m\left(M - \sum_{i=1}^{n} w_i x_i\right) + M\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M - m}\right) \end{aligned} \tag{24}$$

$$= \frac{\left(M - \sum_{i=1}^{n} w_i x_i\right) f(m) + \left(\sum_{i=1}^{n} w_i x_i - m\right) f(M)}{M - m}$$
$$- f\left(\frac{m\left(M - \sum_{i=1}^{n} w_i x_i\right) + M\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M - m}\right) := B.$$

By denoting

$$\Delta_f(t; m, M) := \frac{(t-m) f(M) + (M-t) f(m)}{M - m} - f(t), \quad t \in [m, M]$$

we have

$$\begin{aligned} \Delta_f(t; m, M) &= \frac{(t-m) f(M) + (M-t) f(m) - (M-m) f(t)}{M-m} \\ &= \frac{(t-m) f(M) + (M-t) f(m) - (M-t+t-m) f(t)}{M-m} \\ &= \frac{(t-m)[f(M) - f(t)] - (M-t)[f(t) - f(m)]}{M-m} \\ &= \frac{(M-t)(t-m)}{M-m} \Psi_f(t; m, M) \end{aligned} \tag{25}$$

for any $t \in (m, M)$.

Therefore we have the equality

$$B = \frac{\left(M - \sum_{i=1}^{n} w_i x_i\right)\left(\sum_{i=1}^{n} w_i x_i - m\right)}{M - m} \Psi_f\left(\sum_{i=1}^{n} w_i x_i; m, M\right) \tag{26}$$

provided that $\sum_{i=1}^{n} w_i x_i \in (m, M)$.

For $\sum_{i=1}^{n} w_i x_i = m$ or $\sum_{i=1}^{n} w_i x_i = M$ the inequality (22) is obvious. If $\sum_{i=1}^{n} w_i x_i \in (m, M)$, then

$$\begin{aligned} \Psi_f\left(\sum_{i=1}^{n} w_i x_i; m, M\right) &\leq \sup_{t \in (m,M)} \Psi_f(t; m, M) \\ &= \sup_{t \in (m,M)} \left[\frac{f(M) - f(t)}{M - t} - \frac{f(t) - f(m)}{t - m}\right] \\ &\leq \sup_{t \in (m,M)} \left[\frac{f(M) - f(t)}{M - t}\right] + \sup_{t \in (m,M)} \left[-\frac{f(t) - f(m)}{t - m}\right] \end{aligned}$$

$$= \sup_{t\in(m,M)}\left[\frac{f(M)-f(t)}{M-t}\right] - \inf_{t\in(m,M)}\left[\frac{f(t)-f(m)}{t-m}\right]$$
$$= f'_-(M) - f'_+(m)$$

which by (24) and (26) produces the desired result (22).

Since, obviously

$$\frac{\left(M-\sum_{i=1}^n w_i x_i\right)\left(\sum_{i=1}^n w_i x_i - m\right)}{M-m} \le \frac{1}{4}(M-m),$$

then by (24) and (26) we deduce the second inequality (23). The last part is clear.

Corollary 2. *Let $f: I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \overset{\circ}{I}$. If $x_i \in [m, M]$, then we have the inequalities*

$$\begin{aligned} 0 &\le \frac{1}{n}\sum_{i=1}^n f(x_i) - f\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \\ &\le \frac{\left(M-\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i - m\right)}{M-m}\Psi_f\left(\frac{1}{n}\sum_{i=1}^n x_i; m, M\right) \\ &\le \frac{\left(M-\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i - m\right)}{M-m}\sup_{t\in(m,M)}\Psi_f(t; m, M) \\ &\le \left(M-\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i - m\right)\frac{f'_-(M)-f'_+(m)}{M-m} \\ &\le \frac{1}{4}(M-m)\left[f'_-(M)-f'_+(m)\right], \end{aligned} \tag{27}$$

and

$$\begin{aligned} 0 &\le \frac{1}{n}\sum_{i=1}^n f(x_i) - f\left(\frac{1}{n}\sum_{i=1}^n x_i\right) \\ &\quad \frac{\left(M-\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n x_i - m\right)}{M-m}\Psi_f\left(\frac{1}{n}\sum_{i=1}^n x_i; m, M\right) \\ &\le \frac{1}{4}(M-m)\Psi_f\left(\frac{1}{n}\sum_{i=1}^n x_i; m, M\right) \end{aligned} \tag{28}$$

$$\leq \frac{1}{4}(M-m)\sup_{t\in(m,M)}\Psi_f(t;m,M)$$

$$\leq \frac{1}{4}(M-m)\left[f'_-(M)-f'_+(m)\right],$$

where $\frac{1}{n}\sum_{i=1}^{n}x_i \in (m,M)$.

Remark 3. Define the weighted arithmetic mean of the positive n-tuple $x = (x_1,\ldots,x_n)$ with the nonnegative weights $w = (w_1,\ldots,w_n)$ by

$$A_n(w,x) := \frac{1}{W_n}\sum_{i=1}^{n} w_i x_i$$

where $W_n := \sum_{i=1}^{n} w_i > 0$ and the weighted geometric mean of the same n-tuple, by

$$G_n(w,x) := \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{1/W_n}.$$

It is well known that the following arithmetic mean–geometric mean inequality holds true

$$A_n(w,x) \geq G_n(w,x).$$

Applying the inequality between the first and third term in (27) for the convex function $f(t) = -\ln t, t > 0$ we have

$$1 \leq \frac{A_n(w,x)}{G_n(w,x)} \leq \exp\left[\frac{1}{Mm}(M-A_n(w,x))(A_n(w,x)-m)\right] \tag{29}$$

$$\leq \exp\left[\frac{1}{4}\frac{(M-m)^2}{mM}\right],$$

provided that $0 < m \leq x_i \leq M < \infty$ for $i \in \{1,\ldots,n\}$.

Also, if we apply the inequality (28) for the same function f, we get that

$$1 \leq \frac{A_n(w,x)}{G_n(w,x)} \tag{30}$$

$$\leq \left[\left(\frac{M}{A_n(w,x)}\right)^{M-A_n(w,x)}\left(\frac{m}{A_n(w,x)}\right)^{A_n(w,x)-m}\right]^{-\frac{1}{4}(M-m)}$$

$$\leq \exp\left[\frac{1}{4}\frac{(M-m)^2}{mM}\right].$$

The following result also holds:

Theorem 5 (Dragomir [43]). *With the assumptions of Theorem 4, we have the inequalities*

$$\begin{aligned} 0 &\leq \sum_{i=1}^{n} w_i f(x_i) - f\left(\sum_{i=1}^{n} w_i x_i\right) \\ &\leq 2\max\left\{\frac{M-\sum_{i=1}^{n} w_i x_i}{M-m}, \frac{\sum_{i=1}^{n} w_i x_i - m}{M-m}\right\} \\ &\quad \times \left[\frac{f(m)+f(M)}{2} - f\left(\frac{m+M}{2}\right)\right] \\ &\leq \frac{1}{2}\max\left\{M-\sum_{i=1}^{n} w_i x_i, \sum_{i=1}^{n} w_i x_i - m\right\}\left[f'_{-}(M) - f'_{+}(m)\right]. \end{aligned} \tag{31}$$

Proof. First of all, we recall the following result obtained by the author in [36] that provides a refinement and a reverse for the weighted Jensen's discrete inequality:

$$\begin{aligned} n \min_{i\in\{1,\ldots,n\}}\{p_i\}\left[\frac{1}{n}\sum_{i=1}^{n} f(x_i) - f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right] \\ \leq \frac{1}{P_n}\sum_{i=1}^{n} p_i f(x_i) - f\left(\frac{1}{P_n}\sum_{i=1}^{n} p_i x_i\right) \\ \leq n \max_{i\in\{1,\ldots,n\}}\{p_i\}\left[\frac{1}{n}\sum_{i=1}^{n} f(x_i) - f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right], \end{aligned} \tag{32}$$

where $f: C \to \mathbb{R}$ is a convex function defined on the convex subset C of the linear space X, $\{x_i\}_{i\in\{1,\ldots,n\}} \subset C$ are vectors and $\{p_i\}_{i\in\{1,\ldots,n\}}$ are nonnegative numbers with $P_n := \sum_{i=1}^{n} p_i > 0$.

For $n = 2$ we deduce from (32) that

$$\begin{aligned} 2\min\{t, 1-t\}\left[\frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right)\right] \\ \leq tf(x) + (1-t)f(y) - f(tx + (1-t)y) \\ \leq 2\max\{t, 1-t\}\left[\frac{f(x)+f(y)}{2} - f\left(\frac{x+y}{2}\right)\right] \end{aligned} \tag{33}$$

for any $x, y \in C$ and $t \in [0,1]$.

If we use the second inequality in (33) for the convex function $f: I \to \mathbb{R}$ and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$, we have for $t = \frac{M-\sum_{i=1}^{n} w_i x_i}{M-m}$ that

$$\begin{aligned}&\frac{\left(M-\sum_{i=1}^{n} w_i x_i\right) f(m)+\left(\sum_{i=1}^{n} w_i x_i-m\right) f(M)}{M-m}\\&-f\left(\frac{m\left(M-\sum_{i=1}^{n} w_i x_i\right)+M\left(\sum_{i=1}^{n} w_i x_i-m\right)}{M-m}\right)\\&\leq 2\max\left\{\frac{M-\sum_{i=1}^{n} w_i x_i}{M-m}, \frac{\sum_{i=1}^{n} w_i x_i-m}{M-m}\right\}\\&\times\left[\frac{f(m)+f(M)}{2}-f\left(\frac{m+M}{2}\right)\right].\end{aligned} \tag{34}$$

Utilizing the inequality (24) and (34) we deduce the first inequality in (31).

Since

$$\begin{aligned}&\frac{\frac{f(m)+f(M)}{2}-f\left(\frac{m+M}{2}\right)}{M-m}\\&=\frac{1}{4}\left[\frac{f(M)-f\left(\frac{m+M}{2}\right)}{M-\frac{m+M}{2}}-\frac{f\left(\frac{m+M}{2}\right)-f(m)}{\frac{m+M}{2}-m}\right]\end{aligned}$$

and, by the gradient inequality, we have that

$$\frac{f(M)-f\left(\frac{m+M}{2}\right)}{M-\frac{m+M}{2}} \leq f'_{-}(M)$$

and

$$\frac{f\left(\frac{m+M}{2}\right)-f(m)}{\frac{m+M}{2}-m} \geq f'_{+}(m),$$

then we get

$$\frac{\frac{f(m)+f(M)}{2}-f\left(\frac{m+M}{2}\right)}{M-m} \leq \frac{1}{4}\left[f'_{-}(M)-f'_{+}(m)\right]. \tag{35}$$

On making use of (34) and (35) we deduce the last part of (31).

Corollary 3. *With the assumptions in Corollary 2, we have the inequalities*

$$\begin{aligned}&0\leq\frac{1}{n}\sum_{i=1}^{n} f(x_i)-f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\\&\leq 2\max\left\{\frac{M-\frac{1}{n}\sum_{i=1}^{n} x_i}{M-m}, \frac{\frac{1}{n}\sum_{i=1}^{n} x_i-m}{M-m}\right\}\end{aligned} \tag{36}$$

$$\times \left[\frac{f(m)+f(M)}{2}-f\left(\frac{m+M}{2}\right)\right] \tag{37}$$
$$\leq \frac{1}{2}\max\left\{M-\frac{1}{n}\sum_{i=1}^{n}x_i, \frac{1}{n}\sum_{i=1}^{n}x_i-m\right\}\left[f'_-(M)-f'_+(m)\right].$$

Remark 4. Since, obviously,

$$\frac{M-\sum_{i=1}^{n}w_ix_i}{M-m}, \frac{\sum_{i=1}^{n}w_ix_i-m}{M-m}\leq 1,$$

then we obtain from the first inequality in (31) the simpler, however coarser inequality, namely

$$0\leq \sum_{i=1}^{n}w_if(x_i)-f\left(\sum_{i=1}^{n}w_ix_i\right) \tag{38}$$
$$\leq 2\left[\frac{f(m)+f(M)}{2}-f\left(\frac{m+M}{2}\right)\right].$$

This inequality was obtained in 2008 by S. Simic in [84].

Remark 5. With the assumptions in Remark 3 we have the following reverse of the arithmetic mean–geometric mean inequality

$$1\leq \frac{A_n(w,x)}{G_n(w,x)}\leq \left(\frac{A(m,M)}{G(m,M)}\right)^{2\max\left\{\frac{M-A_n(w,x)}{M-m}, \frac{A_n(w,x)-m}{M-m}\right\}}, \tag{39}$$

where $A(m,M)$ is the arithmetic mean while $G(m,M)$ is the geometric mean of the positive numbers m and M.

5 Applications for the Hölder Inequality

If $x_i, y_i \geq 0$ for $i\in\{1,\ldots,n\}$, then the *Hölder inequality* holds true

$$\sum_{i=1}^{n}x_iy_i\leq \left(\sum_{i=1}^{n}x_i^p\right)^{1/p}\left(\sum_{i=1}^{n}y_i^q\right)^{1/q},$$

where $p>1, \frac{1}{p}+\frac{1}{q}=1$.

Assume that $p>1$. If $z_i\in\mathbb{R}$ for $i\in\{1,\ldots,n\}$, satisfies the bounds

$$0<m\leq z_i\leq M<\infty \text{ for } i\in\{1,\ldots,n\}$$

and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^n w_i > 0$, then from (22) we have

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n w_i z_i^p}{W_n} - \left(\frac{\sum_{i=1}^n w_i z_i}{W_n} \right)^p \\ &\leq \frac{\left(M - \frac{\sum_{i=1}^n w_i z_i}{W_n} \right) \left(\frac{\sum_{i=1}^n w_i z_i}{W_n} - m \right)}{M - m} B_p (m, M) \\ &\leq p \frac{M^{p-1} - m^{p-1}}{M - m} \left(M - \frac{\sum_{i=1}^n w_i z_i}{W_n} \right) \left(\frac{\sum_{i=1}^n w_i z_i}{W_n} - m \right) \\ &\leq \frac{1}{4} p (M - m) \left(M^{p-1} - m^{p-1} \right), \end{aligned} \tag{40}$$

where $\Psi_p (\cdot; m, M) : (m, M) \to \mathbb{R}$ is defined by

$$\Psi_p (t; m, M) = \frac{M^p - t^p}{M - t} - \frac{t^p - m^p}{t - m}$$

while

$$B_p (m, M) := \sup_{t \in (m, M)} \Psi_p (t; m, M). \tag{41}$$

From (23) we also have the inequality

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n w_i z_i^p}{W_n} - \left(\frac{\sum_{i=1}^n w_i z_i}{W_n} \right)^p \\ &\leq \frac{1}{4} (M - m) \Psi_p \left(\frac{\sum_{i=1}^n w_i z_i}{W_n}; m, M \right) \leq \frac{1}{4} p (M - m) \left(M^{p-1} - m^{p-1} \right). \end{aligned} \tag{42}$$

Proposition 1 (Dragomir [43]). *If $x_i \geq 0$, $y_i > 0$ for $i \in \{1, \ldots, n\}$, $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants $\gamma, \Gamma > 0$ and such that*

$$\gamma \leq \frac{x_i}{y_i^{q-1}} \leq \Gamma \text{ for } i \in \{1, \ldots, n\},$$

then we have

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n x_i^p}{\sum_{i=1}^n y_i^q} - \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} \right)^p \\ &\leq \frac{B_p (\gamma, \Gamma)}{\Gamma - \gamma} \left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} \right) \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma \right) \end{aligned} \tag{43}$$

$$\leq p\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\left(\Gamma-\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^q}\right)\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^q}-\gamma\right)$$

$$\leq \frac{1}{4}p\left(\Gamma-\gamma\right)\left(\Gamma^{p-1}-\gamma^{p-1}\right),$$

and

$$0 \leq \frac{\sum_{i=1}^{n}x_i^p}{\sum_{i=1}^{n}y_i^q}-\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^q}\right)^p \tag{44}$$

$$\leq \frac{1}{4}\left(\Gamma-\gamma\right)\Psi_p\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^q};\gamma,\Gamma\right)\leq \frac{1}{4}p\left(\Gamma-\gamma\right)\left(\Gamma^{p-1}-\gamma^{p-1}\right),$$

where $B_p(\cdot,\cdot)$ and $\Psi_p(\cdot;\cdot,\cdot)$ are defined above.

Proof. The inequalities (43) and (44) follow from (40) and (42) by choosing

$$z_i=\frac{x_i}{y_i^{q-1}} \text{ and } w_i=y_i^q.$$

The details are omitted.

Remark 6. We observe that for $p=q=2$ we have $\Psi_2(t;\gamma,\Gamma)=\Gamma-\gamma=B_2(\gamma,\Gamma)$ and then from the first inequality in (43) we get the following reverse of the Cauchy–Bunyakovsky–Schwarz inequality:

$$\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}x_iy_i\right)^2 \tag{45}$$

$$\leq\left(\Gamma-\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^2}\right)\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^2}-\gamma\right)\left(\sum_{i=1}^{n}y_i^2\right)^2$$

provided that $x_i\geq 0$, $y_i>0$ for $i\in\{1,\ldots,n\}$ and there exists the constants $\gamma,\Gamma>0$ such that

$$\gamma\leq\frac{x_i}{y_i}\leq\Gamma \text{ for } i\in\{1,\ldots,n\}.$$

Corollary 4 (Dragomir [43]). *With the assumptions of Proposition 1 we have the following additive reverses of the Hölder inequality*

$$0\leq\left(\sum_{i=1}^{n}x_i^p\right)^{1/p}\left(\sum_{i=1}^{n}y_i^q\right)^{1/q}-\sum_{i=1}^{n}x_iy_i \tag{46}$$

$$\leq \left[\frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\right]^{1/p}\left(\Gamma-\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)^{1/p}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}-\gamma\right)^{1/p}$$
$$\times\sum_{i=1}^n y_i^q$$
$$\leq p^{1/p}\left(\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\right)^{1/p}\left(\Gamma-\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)^{1/p}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}-\gamma\right)^{1/p}$$
$$\times\sum_{i=1}^n y_i^q$$
$$\leq \frac{1}{4^{1/p}}p^{1/p}(\Gamma-\gamma)^{1/p}\left(\Gamma^{p-1}-\gamma^{p-1}\right)^{1/p}\sum_{i=1}^n y_i^q$$

and

$$0\leq\left(\sum_{i=1}^n x_i^p\right)^{1/p}\left(\sum_{i=1}^n y_i^q\right)^{1/q}-\sum_{i=1}^n x_i y_i \tag{47}$$
$$\leq \frac{1}{4^{1/p}}(\Gamma-\gamma)^{1/p}\Psi_p^{1/p}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q};m,M\right)\sum_{i=1}^n y_i^q$$
$$\leq \frac{1}{4^{1/p}}p^{1/p}(\Gamma-\gamma)^{1/p}\left(\Gamma^{p-1}-\gamma^{p-1}\right)^{1/p}\sum_{i=1}^n y_i^q$$

where $p>1$ *and* $\frac{1}{p}+\frac{1}{q}=1$.

Proof. By multiplying in (43) with $\left(\sum_{i=1}^n y_i^q\right)^p$ we have

$$\sum_{i=1}^n x_i^p\left(\sum_{i=1}^n y_i^q\right)^{p-1}-\left(\sum_{i=1}^n x_i y_i\right)^p$$
$$\leq \frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\left(\Gamma-\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}-\gamma\right)\left(\sum_{i=1}^n y_i^q\right)^p$$
$$\leq p\frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma}\left(\Gamma-\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}-\gamma\right)\left(\sum_{i=1}^n y_i^q\right)^p$$
$$\leq \frac{1}{4}p(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right)\left(\sum_{i=1}^n y_i^q\right)^p,$$

which is equivalent to

$$\begin{aligned}
&\sum_{i=1}^{n} x_i^p \left(\sum_{i=1}^{n} y_i^q\right)^{p-1} \\
&\leq \left(\sum_{i=1}^{n} x_i y_i\right)^p + \frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma\right) \\
&\quad \times \left(\sum_{i=1}^{n} y_i^q\right)^p \\
&\leq \left(\sum_{i=1}^{n} x_i y_i\right)^p + p\left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma\right) \\
&\quad \times \left(\sum_{i=1}^{n} y_i^q\right)^p \frac{\Gamma^{p-1}-\gamma^{p-1}}{\Gamma-\gamma} \\
&\leq \left(\sum_{i=1}^{n} x_i y_i\right)^p + \frac{1}{4} p(\Gamma-\gamma)\left(\Gamma^{p-1}-\gamma^{p-1}\right)\left(\sum_{i=1}^{n} y_i^q\right)^p.
\end{aligned} \tag{48}$$

Taking the power $1/p$ with $p > 1$ and employing the following elementary inequality that state that for $p > 1$ and $\alpha, \beta > 0$,

$$(\alpha+\beta)^{1/p} \leq \alpha^{1/p} + \beta^{1/p}$$

we have from the first part of (48) that

$$\begin{aligned}
&\left(\sum_{i=1}^{n} x_i y_i\right)^{1/p} \left(\sum_{i=1}^{n} y_i^q\right)^{1-\frac{1}{p}} \\
&\leq \sum_{i=1}^{n} x_i y_i + \left[\frac{B_p(\gamma,\Gamma)}{\Gamma-\gamma}\right]^{1/p}\left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)^{1/p}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma\right)^{1/p} \\
&\quad \times \sum_{i=1}^{n} y_i^q
\end{aligned} \tag{49}$$

and since $1-\frac{1}{p} = \frac{1}{q}$ we get from (49) the first inequality in (46). The rest is obvious.

The inequality (47) can be proved in a similar manner, however the details are omitted.

If $z_i \in \mathbb{R}$ for $i \in \{1, \ldots, n\}$, satisfies the bounds

$$0 < m \leq z_i \leq M < \infty \text{ for } i \in \{1, \ldots, n\}$$

and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^n w_i > 0$, then from (31) we also have the inequality

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n w_i z_i^p}{W_n} - \left(\frac{\sum_{i=1}^n w_i z_i}{W_n} \right)^p \\ &\leq 2 \left[\frac{m^p + M^p}{2} - \left(\frac{m+M}{2} \right)^p \right] \\ &\quad \times \max \left\{ \frac{M - \frac{\sum_{i=1}^n w_i z_i}{W_n}}{M-m}, \frac{\frac{\sum_{i=1}^n w_i z_i}{W_n} - m}{M-m} \right\} \\ &\leq \frac{1}{2} p \left(M^{p-1} - m^{p-1} \right) \max \left\{ M - \frac{\sum_{i=1}^n w_i z_i}{W_n}, \frac{\sum_{i=1}^n w_i z_i}{W_n} - m \right\}. \end{aligned} \tag{50}$$

From the inequality (50) we can state:

Proposition 2 (Dragomir [43]). *With the assumptions of Proposition 1 we have*

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n x_i^p}{\sum_{i=1}^n y_i^q} - \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} \right)^p \\ &\leq 2 \cdot \frac{\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2} \right)^p}{\Gamma - \gamma} \max \left\{ \Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}, \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma \right\} \\ &\leq \frac{1}{2} p \left(\Gamma^{p-1} - \gamma^{p-1} \right) \max \left\{ \Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}, \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma \right\}. \end{aligned} \tag{51}$$

Finally, the following additive reverse of the Hölder inequality can be stated as well:

Corollary 5 (Dragomir [43]). *With the assumptions of Proposition 1 we have*

$$\begin{aligned} 0 &\leq \left(\sum_{i=1}^n x_i^p \right)^{1/p} \left(\sum_{i=1}^n y_i^q \right)^{1/q} - \sum_{i=1}^n x_i y_i \\ &\leq 2^{1/p} \cdot \left(\frac{\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2} \right)^p}{\Gamma - \gamma} \right)^{1/p} \\ &\quad \times \max \left\{ \left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} \right)^{1/p}, \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma \right)^{1/p} \right\} \sum_{i=1}^n y_i^q \end{aligned} \tag{52}$$

$$\leq \frac{1}{2^{1/p}} p^{1/p} \max\left\{\left(\Gamma - \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q}\right)^{1/p}, \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^q} - \gamma\right)^{1/p}\right\}$$

$$\times \left(\Gamma^{p-1} - \gamma^{p-1}\right)^{1/p} \sum_{i=1}^n y_i^q.$$

Remark 7. As a simpler, however coarser inequality we have the following result:

$$0 \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} \left(\sum_{i=1}^n y_i^q\right)^{1/q} - \sum_{i=1}^n x_i y_i \tag{53}$$

$$\leq 2^{1/p} \cdot \left[\frac{\gamma^p + \Gamma^p}{2} - \left(\frac{\gamma + \Gamma}{2}\right)^p\right]^{1/p} \sum_{i=1}^n y_i^q,$$

where x_i and y_i are as above.

6 Applications for f-Divergence

Consider the f-divergence

$$I_f(p,q) := \sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right) \tag{54}$$

defined on the set of probability distributions $p, q \in \mathbb{P}^n$, where f is convex on $(0,\infty)$. It is assumed that $f(u)$ is zero and strictly convex at $u = 1$. By appropriately defining this convex function, various divergences are derived.

The following result holds:

Proposition 3 (Dragomir [43]). *Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathbb{P}^n$ and there exists the constants $0 < r < 1 < R < \infty$ such that*

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1,\ldots,n\}. \tag{55}$$

Then we have the inequalities

$$0 \leq I_f(p,q) \leq \frac{(R-1)(1-r)}{R-r} \sup_{t\in(r,R)} \Psi_f(t;r,R) \tag{56}$$

$$\leq (R-1)(1-r)\frac{f'_-(R) - f'_+(r)}{R-r}$$

$$\leq \frac{1}{4}(R-r)\left[f'_-(R) - f'_+(r)\right],$$

and $\Psi_f(\cdot; r, R) : (r, R) \to \mathbb{R}$ *is defined by*

$$\Psi_f(t; r, R) = \frac{f(R) - f(t)}{R - t} - \frac{f(t) - f(r)}{t - r}.$$

We also have the inequality

$$I_f(p, q) \leq \frac{1}{4}(R - r)\frac{f(R)(1 - r) + f(r)(R - 1)}{(R - 1)(1 - r)} \tag{57}$$
$$\leq \frac{1}{4}(R - r)\left[f'_-(R) - f'_+(r)\right].$$

The proof follows by Theorem 4 by choosing $w_i = q_i$, $x_i = \frac{p_i}{q_i}$, $m = r$ and $M = R$ and performing the required calculations. The details are omitted.

Utilising the same approach and Theorem 5 we can also state that:

Proposition 4 (Dragomir [43]). *With the assumptions of Proposition 3 we have*

$$0 \leq I_f(p, q) \leq 2\max\left\{\frac{R - 1}{R - r}, \frac{1 - r}{R - r}\right\} \tag{58}$$
$$\times \left[\frac{f(r) + f(R)}{2} - f\left(\frac{r + R}{2}\right)\right]$$
$$\leq \frac{1}{2}\max\{R - 1, 1 - r\}\left[f'_-(R) - f'_+(r)\right].$$

The above results can be utilized to obtain various inequalities for the divergence measures in Information Theory that are particular instances of f-divergence.

Consider the Kullback–Leibler divergence

$$KL(p, q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right), \; p, q \in \mathbb{P}^n.$$

For the convex function $f : (0, \infty) \to \mathbb{R}, f(t) = -\ln t$ we have

$$I_f(p, q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) = -\sum_{i=1}^{n} q_i \ln\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i}\right) = KL(q, p)$$

If $p, q \in \mathbb{P}^n$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1, \ldots, n\}, \tag{59}$$

then we get from the second inequality in (56) that

$$0 \leq KL(q,p) \leq \frac{(R-1)(1-r)}{rR}, \tag{60}$$

from the first inequality in (57) that

$$0 \leq KL(q,p) \leq \frac{1}{4}(R-r)\ln\left[R^{-\frac{1}{R-1}} r^{-\frac{1}{1-r}}\right]$$

and from the first inequality in (58) that

$$0 \leq KL(q,p) \leq 2\max\left\{\frac{R-1}{R-r}, \frac{1-r}{R-r}\right\} \ln\left(\frac{A(r,R)}{G(r,R)}\right) \tag{61}$$

where $A(r,R)$ is the arithmetic mean and $G(r,R)$ is the geometric mean of the positive numbers r and R.

For the convex function $f:(0,\infty)\to\mathbb{R}, f(t)=t\ln t$ we have

$$I_f(p,q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) = KL(p,q).$$

If $p,q \in \mathbb{P}^n$ such that there exists the constants $0<r<1<R<\infty$ with the property (59), then we get from the second inequality in (56) that

$$0 \leq KL(p,q) \leq \frac{(R-1)(1-r)}{L(r,R)}, \tag{62}$$

where $L(r,R)$ is the Logarithmic mean of r, R, namely

$$L(r,R) = \frac{R-r}{\ln R - \ln r}.$$

From the first inequality in (57) we also have:

$$0 \leq KL(p,q) \leq \frac{1}{4}(R-r)\frac{R-r+\ln\left(R^{1-r}r^{R-1}\right)}{(R-1)(1-r)}. \tag{63}$$

Finally, by the first inequality in (58) we have

$$0 \leq KL(p,q) \leq 2\max\left\{\frac{R-1}{R-r}, \frac{1-r}{R-r}\right\} \ln\left[\frac{G\left(r^r,R^R\right)}{[A(r,R)]^{A(r,R)}}\right]. \tag{64}$$

7 More Reverse Inequalities

For the *Lebesgue measurable* function $g : [\alpha, \beta] \to \mathbb{R}$ we introduce the *Lebesgue p-norms* defined as

$$\|g\|_{[\alpha,\beta],p} := \left(\int_\alpha^\beta |g(t)|^p \, dt \right)^{1/p} \text{ if } g \in L_p[\alpha, \beta],$$

for $p \geq 1$ and

$$\|g\|_{[\alpha,\beta],\infty} := ess \sup_{t \in [\alpha,\beta]} |g(t)| \text{ if } g \in L_\infty[\alpha, \beta],$$

for $p = \infty$.

The following result also holds:

Theorem 6 (Dragomir [44]). *Let $\Phi : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$, $\mathring{I}$ is the interior of I. If $x_i \in I$ and $w_i \geq 0$ for $i \in \{1, \ldots, n\}$ with $\sum_{i=1}^n w_i = 1$, denote $\bar{x}_w := \sum_{i=1}^n w_i x_i \in I$, then we have the inequality*

$$0 \leq \sum_{i=1}^n w_i \Phi(x_i) - \Phi(\bar{x}_w) \tag{65}$$
$$\leq \frac{(M - \bar{x}_w) \int_m^{\bar{x}_w} |\Phi'(t)| \, dt + (\bar{x}_w - m) \int_{\bar{x}_w}^M |\Phi'(t)| \, dt}{M - m} := \Theta_\Phi(\bar{x}_w; m, M),$$

where $\Theta_\Phi(\bar{x}_w; m, M)$ satisfies the bounds

$$\Theta_\Phi(\bar{x}_w; m, M) \tag{66}$$
$$\leq \begin{cases} \left[\frac{1}{2} + \frac{\left| \bar{x}_w - \frac{m+M}{2} \right|}{M-m} \right] \int_m^M |\Phi'(t)| \, dt \\ \left[\frac{1}{2} \int_m^M |\Phi'(t)| \, dt + \frac{1}{2} \left| \int_{\bar{x}_w}^M |\Phi'(t)| \, dt - \int_m^{\bar{x}_w} |\Phi'(t)| \, dt \right| \right], \end{cases}$$

$$\Theta_\Phi(\bar{x}_w; m, M) \tag{67}$$
$$\leq \frac{(\bar{x}_w - m)(M - \bar{x}_w)}{M - m} \left[\|\Phi'\|_{[\bar{x}_w, M],\infty} + \|\Phi'\|_{[m,\bar{x}_w],\infty} \right]$$
$$\leq \frac{1}{2}(M - m) \frac{\|\Phi'\|_{[\bar{w}_p, M],\infty} + \|\Phi'\|_{[m,\bar{w}_p],\infty}}{2} \leq \frac{1}{2}(M - m) \|\Phi'\|_{[m,M],\infty}$$

and

$$\begin{aligned}
&\Theta_{\Phi}\left(\bar{x}_w; m, M\right) \qquad (68)\\
&\leq \frac{1}{M-m}\Big[\left(\bar{x}_w - m\right)\left(M - \bar{x}_w\right)^{1/q}\left\|\Phi'\right\|_{[\bar{x}_w, M], p}\\
&\quad + \left(M - \bar{x}_w\right)\left(\bar{x}_w - m\right)^{1/q}\left\|\Phi'\right\|_{[m, \bar{x}_w], p}\Big]\\
&\leq \frac{1}{M-m}\left[\left(\bar{x}_w - m\right)^q\left(M - \bar{x}_w\right) + \left(M - \bar{x}_w\right)^q\left(\bar{x}_w - m\right)\right]^{1/q}\left\|\Phi'\right\|_{[m, M], p}
\end{aligned}$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

Proof. By the convexity of Φ we have that

$$\begin{aligned}
&\sum_{i=1}^{n} w_i \Phi\left(x_i\right) - \Phi\left(\bar{x}_w\right) \qquad (69)\\
&= \sum_{i=1}^{n} w_i \Phi\left[\frac{m\left(M - x_i\right) + M\left(x_i - m\right)}{M-m}\right] - \Phi\left(\bar{x}_w\right)\\
&\leq \sum_{i=1}^{n} w_i \frac{\left(M - x_i\right)\Phi\left(m\right) + \left(x_i - m\right)\Phi\left(M\right)}{M-m} - \Phi\left(\bar{x}_w\right)\\
&= \frac{\left(M - \bar{x}_w\right)\Phi\left(m\right) + \left(\bar{x}_w - m\right)\Phi\left(M\right)}{M-m} - \Phi\left(\bar{x}_w\right) = B.
\end{aligned}$$

By denoting

$$\Lambda_{\Phi}\left(t; m, M\right) := \frac{\left(t - m\right)\Phi\left(M\right) + \left(M - t\right)\Phi\left(m\right)}{M-m} - \Phi\left(t\right), \quad t \in [m, M]$$

we have

$$\begin{aligned}
\Lambda_{\Phi}\left(t; m, M\right) &= \frac{\left(t - m\right)\Phi\left(M\right) + \left(M - t\right)\Phi\left(m\right)}{M-m} - \Phi\left(t\right) \qquad (70)\\
&= \frac{\left(t - m\right)\Phi\left(M\right) + \left(M - t\right)\Phi\left(m\right) - \left(M - m\right)\Phi\left(t\right)}{M-m}\\
&= \frac{\left(t - m\right)\Phi\left(M\right) + \left(M - t\right)\Phi\left(m\right) - \left(M - t + t - m\right)\Phi\left(t\right)}{M-m}\\
&= \frac{\left(t - m\right)\left[\Phi\left(M\right) - \Phi\left(t\right)\right] - \left(M - t\right)\left[\Phi\left(t\right) - \Phi\left(m\right)\right]}{M-m}
\end{aligned}$$

for any $t \in [m, M]$. Also

$$B = \Lambda_\Phi (\bar{x}_w; m, M).$$

Taking the modulus on (70) and, noticing that, by the convexity of Φ we have

$$\Lambda_\Phi (t; m, M) = \frac{(t-m)\Phi(M) + (M-t)\Phi(m)}{M-m} - \Phi\left(\frac{(t-m)M + (M-t)m}{M-m}\right) \geq 0$$

for any $t \in [m, M]$, then we have

$$\begin{aligned} \Lambda_\Phi (t; m, M) &\leq \frac{(t-m)|\Phi(M) - \Phi(t)| + (M-t)|\Phi(t) - \Phi(m)|}{M-m} \\ &= \frac{(t-m)\left|\int_t^M \Phi'(s)\,ds\right| + (M-t)\left|\int_m^t \Phi'(s)\,ds\right|}{M-m} \\ &\leq \frac{(t-m)\int_t^M |\Phi'(s)|\,ds + (M-t)\int_m^t |\Phi'(s)|\,ds}{M-m} \end{aligned} \tag{71}$$

for any $t \in [m, M]$.

Finally, if we write the inequality (71) for $t = \bar{x}_w \in [m, M]$ and utilize the inequality (69), we deduce the desired result (65).

Now, we observe that

$$\begin{aligned} &\frac{(t-m)\int_t^M |\Phi'(s)|\,ds + (M-t)\int_m^t |\Phi'(s)|\,ds}{M-m} \\ &\leq \begin{cases} \max\{t-m, M-t\}\int_m^M |\Phi'(t)|\,dt \\ \max\left\{\int_t^M |\Phi'(s)|\,ds, \int_m^t |\Phi'(s)|\,ds\right\}(M-m) \end{cases} \\ &= \begin{cases} \left[\frac{1}{2}(M-m) + \left|t - \frac{m+M}{2}\right|\right]\int_m^M |\Phi'(t)|\,dt \\ \left[\frac{1}{2}\int_m^M |\Phi'(s)|\,ds + \frac{1}{2}\left|\int_t^M |\Phi'(s)|\,ds - \int_m^t |\Phi'(s)|\,ds\right|\right](M-m) \end{cases} \end{aligned} \tag{72}$$

for any $t \in [m, M]$. This proves the inequality (66).

By the Hölder's inequality we have

$$\int_t^M \left|\Phi'(s)\right| ds \leq \begin{cases} (M-t)\,\|\Phi'\|_{[t,M],\infty} \\ (M-t)^{1/q}\,\|\Phi'\|_{[t,M],p} \text{ if } p > 1, \frac{1}{p} + \frac{1}{q} = 1 \end{cases}$$

and

$$\int_m^t \left|\Phi'(s)\right| ds \leq \begin{cases} (t-m)\left\|\Phi'\right\|_{[m,t],\infty} \\ (t-m)^{1/q}\left\|\Phi'\right\|_{[m,t],p} \text{ if } p>1, \frac{1}{p}+\frac{1}{q}=1, \end{cases}$$

which give that

$$\begin{aligned} &\frac{(t-m)\int_t^M |\Phi'(s)|\,ds + (M-t)\int_m^t |\Phi'(s)|\,ds}{M-m} \\ &\leq \frac{(t-m)(M-t)\left\|\Phi'\right\|_{[t,M],\infty} + (M-t)(t-m)\left\|\Phi'\right\|_{[m,t],\infty}}{M-m} \\ &= \frac{(t-m)(M-t)}{M-m}\left[\left\|\Phi'\right\|_{[t,M],\infty} + \left\|\Phi'\right\|_{[m,t],\infty}\right] \\ &\leq \frac{1}{2}(M-m)\frac{\left\|\Phi'\right\|_{[t,M],\infty} + \left\|\Phi'\right\|_{[m,t],\infty}}{2} \\ &\leq \frac{1}{2}(M-m)\max\left\{\left\|\Phi'\right\|_{[t,M],\infty}, \left\|\Phi'\right\|_{[m,t],\infty}\right\} = \frac{1}{2}(M-m)\left\|\Phi'\right\|_{[m,M],\infty} \end{aligned} \tag{73}$$

and

$$\begin{aligned} &\frac{(t-m)\int_t^M |\Phi'(s)|\,ds + (M-t)\int_m^t |\Phi'(s)|\,ds}{M-m} \\ &\leq \frac{(t-m)(M-t)^{1/q}\left\|\Phi'\right\|_{[t,M],p} + (M-t)(t-m)^{1/q}\left\|\Phi'\right\|_{[m,t],p}}{M-m} \\ &\leq \frac{1}{M-m}\left[\left((t-m)(M-t)^{1/q}\right)^q + \left((M-t)(t-m)^{1/q}\right)^q\right]^{1/q} \\ &\quad\times\left[\left\|\Phi'\right\|_{[t,M],p}^p + \left\|\Phi'\right\|_{[m,t],p}^p\right]^{1/p} \\ &= \frac{1}{M-m}\left[(t-m)^q(M-t) + (M-t)^q(t-m)\right]^{1/q}\left\|\Phi'\right\|_{[m,M],p} \end{aligned} \tag{74}$$

for any $t \in [m,M]$.

These prove the desired inequalities (67) and (68).

Remark 8. Define the weighted arithmetic mean of the positive n-tuple $x = (x_1,\ldots,x_n)$ with the nonnegative weights $w = (w_1,\ldots,w_n)$ by

$$A_n(w,x) := \frac{1}{W_n}\sum_{i=1}^n w_i x_i$$

where $W_n := \sum_{i=1}^n w_i > 0$ and the weighted geometric mean of the same n-tuple, by

$$G_n(w,x) := \left(\prod_{i=1}^n x_i^{w_i}\right)^{1/W_n}.$$

It is well known that the following arithmetic mean–geometric mean inequality holds true

$$A_n(w,x) \geq G_n(w,x).$$

On applying the inequality (65) for the convex function $\Phi(t) = -\ln t$, we have the following reverse of the arithmetic mean–geometric mean inequality

$$1 \leq \frac{A_n(w,x)}{G_n(w,x)} \leq \left(\frac{A_n(w,x)}{m}\right)^{M-A_n(w,x)} \left(\frac{M}{A_n(w,x)}\right)^{A_n(w,x)-m}. \tag{75}$$

8 Applications for the Hölder Inequality

If $x_i, y_i \geq 0$ for $i \in \{1,\ldots,n\}$, then the *Hölder inequality* holds true

$$\sum_{i=1}^n x_i y_i \leq \left(\sum_{i=1}^n x_i^p\right)^{1/p} \left(\sum_{i=1}^n y_i^q\right)^{1/q},$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

Assume that $p > 1$. If $z_i \in \mathbb{R}$ for $i \in \{1,\ldots,n\}$, satisfies the bounds

$$0 < m \leq z_i \leq M < \infty \text{ for } i \in \{1,\ldots,n\}$$

and $w_i \geq 0$ $(i = 1,\ldots,n)$ with $W_n := \sum_{i=1}^n w_i > 0$, then from Theorem 6 we have amongst other the following inequality

$$\begin{aligned} 0 &\leq \frac{\sum_{i=1}^n w_i z_i^p}{W_n} - \left(\frac{\sum_{i=1}^n w_i z_i}{W_n}\right)^p \\ &\leq (M^p - m^p)\left[\frac{1}{2} + \frac{1}{M-m}\left|\frac{\sum_{i=1}^n w_i z_i}{W_n} - \frac{m+M}{2}\right|\right]. \end{aligned} \tag{76}$$

From this inequality we can state that:

Proposition 5 (Dragomir [44]). *If $x_i \geq 0, y_i > 0$ for $i \in \{1,\ldots,n\}$, $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants $\gamma, \Gamma > 0$ and such that*

$$\gamma \leq \frac{x_i}{y_i^{q-1}} \leq \Gamma \text{ for } i \in \{1,\ldots,n\},$$

then we have

$$0 \le \frac{\sum_{i=1}^{n} x_i^p}{\sum_{i=1}^{n} y_i^q} - \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}\right)^p \tag{77}$$
$$\le (\Gamma^p - \gamma^p)\left[\frac{1}{2} + \frac{1}{\Gamma - \gamma}\left|\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q} - \frac{\gamma + \Gamma}{2}\right|\right].$$

Finally, the following additive reverse of the Hölder inequality can be stated as well:

Corollary 6 (Dragomir [44]). *With the assumptions of Proposition 5 we have*

$$0 \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p}\left(\sum_{i=1}^{n} y_i^q\right)^{1/q} - \sum_{i=1}^{n} x_i y_i \tag{78}$$
$$\le (\Gamma^p - \gamma^p)^{1/p}\left[\frac{1}{2} + \frac{1}{\Gamma - \gamma}\left|\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q} - \frac{\gamma + \Gamma}{2}\right|\right]^{1/p}\sum_{i=1}^{n} y_i^q.$$

Remark 9. We observe that for $p = q = 2$ we have from the first inequality in (77) the following reverse of the Cauchy–Bunyakovsky–Schwarz inequality

$$\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_i y_i\right)^2 \tag{79}$$
$$\le \left(\Gamma^2 - \gamma^2\right)\left[\frac{1}{2} + \frac{1}{\Gamma - \gamma}\left|\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2} - \frac{\gamma + \Gamma}{2}\right|\right]\left(\sum_{i=1}^{n} y_i^2\right)^2$$

provided that $x_i \ge 0$, $y_i > 0$ for $i \in \{1, \ldots, n\}$ and there exists the constants $\gamma, \Gamma > 0$ such that

$$\gamma \le \frac{x_i}{y_i} \le \Gamma \text{ for } i \in \{1, \ldots, n\}.$$

9 Applications for f-Divergence

Consider the f-divergence

$$I_f(p, q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) \tag{80}$$

defined on the set of probability distributions $p, q \in \mathbb{P}^n$, where f is convex on $(0, \infty)$. It is assumed that $f(u)$ is zero and strictly convex at $u = 1$. By appropriately defining this convex function, various divergences are derived.

Proposition 6 (Dragomir [44]). *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathbb{P}^n$ and there exists the constants $0 < r < 1 < R < \infty$ such that*

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1, \ldots, n\}. \tag{81}$$

Then we have the inequalities

$$0 \leq I_f(p, q) \leq B_f(r, R) \tag{82}$$

where

$$B_f(r, R) := \frac{(R-1)\int_r^1 |f'(t)|\, dt + (1-r)\int_1^R |f'(t)|\, dt}{R-r}. \tag{83}$$

Moreover, we have the following bounds for $B_f(r, R)$

$$B_f(r, R) \tag{84}$$
$$\leq \begin{cases} \left[\frac{1}{2} + \frac{\left|1 - \frac{r+R}{2}\right|}{R-r}\right] \int_r^R |f'(t)|\, dt \\ \left[\frac{1}{2}\int_r^R |f'(t)|\, dt + \frac{1}{2}\left|\int_1^R |f'(t)|\, dt - \int_r^1 |f'(t)|\, dt\right|\right], \end{cases}$$

and

$$B_f(r, R) \tag{85}$$
$$\leq \frac{(1-r)(R-1)}{R-r}\left[\|f'\|_{[1,R],\infty} + \|f'\|_{[r,1],\infty}\right]$$
$$\leq \frac{1}{2}(R-r)\frac{\|f'\|_{[1,R],\infty} + \|f'\|_{[r,1],\infty}}{2} \leq \frac{1}{2}(R-r)\|f'\|_{[r,R],\infty}$$

and

$$B_f(r, R) \tag{86}$$
$$\leq \frac{1}{R-r}\left[(1-r)(R-1)^{1/q}\|f'\|_{[1,R],p} + (R-1)(1-r)^{1/q}\|f'\|_{[r,1],p}\right]$$
$$\leq \frac{1}{R-r}\left[(1-r)^q(R-1) + (R-1)^q(1-r)\right]^{1/q}\|f'\|_{[r,R],p}$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

The proof follows by Theorem 6 by choosing $w_i = q_i$, $x_i = \frac{p_i}{q_i}$, $m = r$ and $M = R$ and performing the required calculations. The details are omitted.

The above results can be utilized to obtain various inequalities for the divergence measures in information theory that are particular instances of f-divergence.

Consider the Kullback–Leibler divergence

$$KL(p,q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right), \ p,q \in \mathbb{P}^n.$$

For the convex function $f : (0,\infty) \to \mathbb{R}, f(t) = -\ln t$ we have

$$I_f(p,q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) = -\sum_{i=1}^{n} q_i \ln\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i}\right) = KL(q,p)$$

If $p,q \in \mathbb{P}^n$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1,\ldots,n\}, \tag{87}$$

then we get from the inequality (83)

$$0 \leq KL(q,p) \leq \ln\left(\frac{R^{1-r}}{r^{R-1}}\right)^{\frac{1}{R-r}}. \tag{88}$$

For $\alpha > 1$, let

$$f(t) = t^{\alpha}, \ t > 0.$$

Then

$$I_f(p,q) = D_{\alpha}(p,q) = \sum_{i=1}^{n} p_i^{\alpha} q_i^{1-\alpha},$$

which is the α-order entropy.

If $p,q \in \mathbb{P}^n$ such that (87) holds true, then by (83) we have

$$0 \leq D_{\alpha}(p,q) \leq \frac{(R-1)(1-r^{\alpha}) + (1-r)(R^{\alpha}-1)}{R-r}.$$

10 A Refinement and Another Reverse

For a real function $g : [m, M] \to \mathbb{R}$ and two distinct points $\alpha, \beta \in [m, M]$ we recall that the *divided difference* of g in these points is defined by

$$[\alpha, \beta; g] := \frac{g(\beta) - g(\alpha)}{\beta - \alpha}.$$

Theorem 7 (Dragomir [41]). *Let $f : I \to \mathbb{R}$ be a continuous convex function on the interval of real numbers I and $m, M \in \mathbb{R}$, $m < M$ with $[m, M] \subset \mathring{I}$, $\mathring{I}$ the interior of I. Let $\bar{\mathbf{a}} = (a_1, \ldots, a_n)$, $\bar{\mathbf{p}} = (p_1, \ldots, p_n)$ be n-tuples of real numbers with $p_i \geq 0$ $(i \in \{1, \ldots, n\})$ and $\sum_{i=1}^{n} p_i = 1$. If $m \leq a_i \leq M$, $i \in \{1, \ldots, n\}$, with $\sum_{i=1}^{n} p_i a_i \neq m, M$, then*

$$\begin{aligned}
&\left| \sum_{i=1}^{n} p_i \right| f(a_i) - f\left(\sum_{j=1}^{n} p_j a_j \right) \Bigg| \, sgn\left(a_i - \sum_{j=1}^{n} p_j a_j \right) \Bigg| \\
&\leq \sum_{i=1}^{n} p_i f(a_i) - f\left(\sum_{i=1}^{n} p_i a_i \right) \\
&\leq \frac{1}{2} \left(\left[\sum_{i=1}^{n} p_i a_i, M; f \right] - \left[m, \sum_{i=1}^{n} p_i a_i; f \right] \right) \sum_{i=1}^{n} p_i \left| a_i - \sum_{j=1}^{n} p_j a_j \right| \\
&\leq \frac{1}{2} \left(\left[\sum_{i=1}^{n} p_i a_i, M; f \right] - \left[m, \sum_{i=1}^{n} p_i a_i; f \right] \right) \left[\sum_{i=1}^{n} p_i a_i^2 - \left(\sum_{j=1}^{n} p_j a_j \right)^2 \right]^{1/2}.
\end{aligned} \tag{89}$$

If the lateral derivatives $f'_+(m)$ and $f'_-(M)$ are finite, then we also have the inequalities

$$\begin{aligned}
0 &\leq \sum_{i=1}^{n} p_i f(a_i) - f\left(\sum_{i=1}^{n} p_i a_i \right) \\
&\leq \frac{1}{2} \left(\left[\sum_{i=1}^{n} p_i a_i, M; f \right] - \left[m, \sum_{i=1}^{n} p_i a_i; f \right] \right) \sum_{i=1}^{n} p_i \left| a_i - \sum_{j=1}^{n} p_j a_j \right| \\
&\leq \frac{1}{2} \left[f'_-(M) - f'_+(m) \right] \sum_{i=1}^{n} p_i \left| a_i - \sum_{j=1}^{n} p_j a_j \right|
\end{aligned} \tag{90}$$

$$\leq \frac{1}{2}\left[f'_{-}(M)-f'_{+}(m)\right]\left[\sum_{i=1}^{n}p_i a_i^2-\left(\sum_{j=1}^{n}p_j a_j\right)^2\right]^{1/2}.$$

Proof. We recall that if $f: I\to\mathbb{R}$ is a continuous convex function on the interval of real numbers I and $\alpha\in I$ then the *divided difference function* $f_\alpha: I\setminus\{\alpha\}\to\mathbb{R}$,

$$f_\alpha(t):=[\alpha,t;f]:=\frac{f(t)-f(\alpha)}{t-\alpha}$$

is monotonic nondecreasing on $I\setminus\{\alpha\}$.

For $\bar{a}_p:=\sum_{j=1}^{n}p_j a_j\in(m,M)$, we consider now the sequence

$$f_{\bar{a}_p}(i):=\frac{f(a_i)-f\left(\bar{a}_p\right)}{a_i-\bar{a}_p}.$$

We will show that $f_{\bar{a}_p}(i)$ and $h_i:=a_i-\bar{a}_p,\in\{1,\ldots,n\}$ are synchronous.

Let $i,j\in\{1,\ldots,n\}$ with $a_i,a_j\neq\bar{a}_p$. Assume that $a_i\geq a_j$, then by the monotonicity of f_α we have

$$f_{\bar{a}_p}(i)=\frac{f(a_i)-f\left(\bar{a}_p\right)}{a_i-\bar{a}_p}\tag{91}$$

$$\geq\frac{f\left(a_j\right)-f\left(\bar{a}_p\right)}{a_j-\bar{a}_p}=f_{\bar{a}_p}(j)$$

and

$$h_i\geq h_j\tag{92}$$

which shows that

$$\left[f_{\bar{a}_p}(i)-f_{\bar{a}_p}(j)\right]\left(h_i-h_j\right)\geq 0.\tag{93}$$

If $a_i<a_j$, then the inequalities (91) and (92) reverse but the inequality (93) still holds true.

Utilising the continuity property of the modulus we have

$$\begin{aligned}\left|\left[\left|f_{\bar{a}_p}(i)\right|-\left|f_{\bar{a}_p}(j)\right|\right]\left(h_i-h_j\right)\right|&\leq\left|\left[f_{\bar{a}_p}(i)-f_{\bar{a}_p}(j)\right]\left(h_i-h_j\right)\right|\\&=\left[f_{\bar{a}_p}(i)-f_{\bar{a}_p}(j)\right]\left(h_i-h_j\right)\end{aligned}$$

for any $i,j\in\{1,\ldots,n\}$.

Multiplying with $p_i, p_j \geq 0$ and summing over i and j from 1 to n we have

$$\left| \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left[\left| f_{\bar{a}_p}(i) \right| - \left| f_{\bar{a}_p}(j) \right| \right] (h_i - h_j) \right| \tag{94}$$
$$\leq \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left[f_{\bar{a}_p}(i) - f_{\bar{a}_p}(j) \right] (h_i - h_j).$$

A simple calculation shows that

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left[\left| f_{\bar{a}_p}(i) \right| - \left| f_{\bar{a}_p}(j) \right| \right] (h_i - h_j) \tag{95}$$
$$= \sum_{i=1}^{n} p_i \left| f_{\bar{a}_p}(i) \right| h_i - \sum_{i=1}^{n} p_i \left| f_{\bar{a}_p}(i) \right| \sum_{i=1}^{n} p_i h_i$$
$$= \sum_{i=1}^{n} p_i \left| \frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right| (a_i - \bar{a}_p)$$
$$- \sum_{i=1}^{n} p_i \left| \frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right| \sum_{i=1}^{n} p_i (a_i - \bar{a}_p)$$
$$= \sum_{i=1}^{n} p_i \left| \frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right| (a_i - \bar{a}_p)$$
$$= \sum_{i=1}^{n} p_i \left| f(a_i) - f(\bar{a}_p) \right| sgn(a_i - \bar{a}_p)$$

and

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \left[f_{\bar{a}_p}(i) - f_{\bar{a}_p}(j) \right] (h_i - h_j) \tag{96}$$
$$= \sum_{i=1}^{n} p_i f_{\bar{a}_p}(i) h_i - \sum_{i=1}^{n} p_i f_{\bar{a}_p}(i) \sum_{i=1}^{n} p_i h_i$$
$$= \sum_{i=1}^{n} p_i \left(\frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right) (a_i - \bar{a}_p)$$
$$- \sum_{i=1}^{n} p_i \left(\frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right) \sum_{i=1}^{n} p_i (a_i - \bar{a}_p)$$

$$= \sum_{i=1}^{n} p_i \left(\frac{f(a_i) - f(\bar{a}_p)}{a_i - \bar{a}_p} \right) (a_i - \bar{a}_p)$$
$$= \sum_{i=1}^{n} p_i f(a_i) - f\left(\sum_{i=1}^{n} p_i a_i \right).$$

On making use of the identities (95) and (96) we obtain from (94) the first inequality in (89).

Now, since $\bar{a}_p := \sum_{j=1}^{n} p_j a_j \in (m, M)$ then we have by the monotonicity of $f_{\bar{a}_p}(i)$ that

$$\left[m, \bar{a}_p; f\right] = \frac{f(\bar{a}_p) - f(m)}{\bar{a}_p - m} \leq f_{\bar{a}_p}(i) \tag{97}$$
$$\leq \frac{f(M) - f(\bar{a}_p)}{M - \bar{a}_p} = \left[\bar{a}_p, M; f\right]$$

for any $i \in \{1, \ldots, n\}$.

Applying now the *Grüss' type inequality* obtained by Cerone and Dragomir in [9]

$$\left| \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i \right| \leq \frac{1}{2} (\Gamma - \gamma) \sum_{i=1}^{n} w_i \left| x_i - \sum_{j=1}^{n} w_j x_j \right|$$

provided

$$-\infty < \delta \leq y_i \leq \Delta < \infty \tag{98}$$

for $i = 1, \ldots, n$, we have that

$$\sum_{i=1}^{n} p_i f(a_i) - f\left(\sum_{i=1}^{n} p_i a_i \right)$$
$$\leq \frac{1}{2} \left(\left[\bar{a}_p, M; f\right] - \left[m, \bar{a}_p; f\right] \right) \sum_{i=1}^{n} p_i \left| a_i - \sum_{j=1}^{n} p_j a_j \right|,$$

which proves the second inequality in (89).

The last bound in (89) is obvious by Cauchy–Bunyakovsky–Schwarz discrete inequality.

If the lateral derivatives $f'_+(m)$ and $f'_-(M)$ are finite, then by the convexity of f we have the *gradient inequalities*

$$\frac{f(M)-f\left(\bar{a}_p\right)}{M-\bar{a}_p} \leq f'_-(M)$$

and

$$\frac{f\left(\bar{a}_p\right)-f(m)}{\bar{a}_p-m} \geq f'_+(m),$$

where $\bar{a}_p \in (m,M)$. These imply that

$$\left[\bar{a}_p, M; f\right] - \left[m, \bar{a}_p; f\right] \leq f'_-(M) - f'_+(m)$$

and the proof of the third inequality in (90) is concluded.

The rest is obvious.

Remark 10. Define the weighted arithmetic mean of the positive n-tuple $x = (x_1, \ldots, x_n)$ with the nonnegative weights $w = (w_1, \ldots, w_n)$ by

$$A_n(w,x) := \frac{1}{W_n}\sum_{i=1}^{n} w_i x_i$$

where $W_n := \sum_{i=1}^{n} w_i > 0$ and the weighted geometric mean of the same n-tuple, by

$$G_n(w,x) := \left(\prod_{i=1}^{n} x_i^{w_i}\right)^{1/W_n}.$$

It is well known that the following arithmetic mean–geometric mean inequality holds

$$A_n(w,x) \geq G_n(w,x).$$

Applying the inequality (90) for the convex function $f(t) = -\ln t, t > 0$ we have the following reverse of the arithmetic mean–geometric mean inequality

$$\begin{aligned} 1 &\leq \frac{A_n(w,x)}{G_n(w,x)} \\ &\leq \left[\frac{\left(\frac{A_n(w,x)}{m}\right)^{A_n(w,x)-m}}{\left(\frac{M}{A_n(w,x)}\right)^{M-A_n(w,x)}}\right]^{\frac{1}{2}A_n(w,|x-A_n(w,x)|)} \\ &\leq \exp\left[\frac{1}{2}\frac{M-m}{mM}A_n\left(w,|x-A_n(w,x)|\right)\right], \end{aligned} \tag{99}$$

provided that $0 < m \leq x_i \leq M < \infty$ for $i \in \{1, \ldots, n\}$.

11 Applications for the Hölder Inequality

If $x_i, y_i \geq 0$ for $i \in \{1, \ldots, n\}$, then the *Hölder inequality* holds true

$$\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} x_i^p \right)^{1/p} \left(\sum_{i=1}^{n} y_i^q \right)^{1/q},$$

where $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$.

Assume that $p > 1$. If $z_i \in \mathbb{R}$ for $i \in \{1, \ldots, n\}$, satisfies the bounds

$$0 < m \leq z_i \leq M < \infty \text{ for } i \in \{1, \ldots, n\}$$

and $w_i \geq 0$ $(i = 1, \ldots, n)$ with $W_n := \sum_{i=1}^{n} w_i > 0$, then from Theorem 7 we have amongst other the following inequality

$$\begin{aligned} &\left| \frac{1}{W_n} \sum_{i=1}^{n} \left| z_i^p - \left(\frac{\sum_{i=1}^{n} w_i z_i}{W_n} \right)^p \right| w_i sgn \left[z_i - \frac{\sum_{i=1}^{n} w_i z_i}{W_n} \right] d\mu \right| \qquad (100) \\ &\leq \frac{\sum_{i=1}^{n} w_i z_i^p}{W_n} - \left(\frac{\sum_{i=1}^{n} w_i z_i}{W_n} \right)^p \\ &\leq \frac{1}{2} \left(\left[\frac{\sum_{i=1}^{n} w_i z_i}{W_n}, M; (\cdot)^p \right] - \left[m, \frac{\sum_{i=1}^{n} w_i z_i}{W_n}; (\cdot)^p \right] \right) \tilde{D}_w(z) \\ &\leq \frac{1}{2} \left(\left[\frac{\sum_{i=1}^{n} w_i z_i}{W_n}, M; (\cdot)^p \right] - \left[m, \frac{\sum_{i=1}^{n} w_i z_i}{W_n}; (\cdot)^p \right] \right) \tilde{D}_{w,2}(z) \\ &\leq \frac{1}{4} \left(\left[\frac{\sum_{i=1}^{n} w_i z_i}{W_n}, M; (\cdot)^p \right] - \left[m, \frac{\sum_{i=1}^{n} w_i z_i}{W_n}; (\cdot)^p \right] \right) (M - m), \end{aligned}$$

where $\frac{\sum_{i=1}^{n} w_i z_i}{W_n} \in (m, M)$ and

$$\tilde{D}_w(z) := \frac{1}{W_n} \sum_{i=1}^{n} w_i \left| z_i - \frac{\sum_{j=1}^{n} w_j z_j}{W_n} \right|$$

while

$$\tilde{D}_{w,2}(z) = \left[\frac{\sum_{i=1}^{n} w_i z_i^2}{W_n} - \left(\frac{\sum_{i=1}^{n} w_i z_i}{W_n} \right)^2 \right]^{\frac{1}{2}}.$$

The following result related to the Hölder inequality holds:

Proposition 7 (Dragomir [41]). *If $x_i \geq 0$, $y_i > 0$ for $i \in \{1,\dots,n\}$, $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$ and there exists the constants $\gamma, \Gamma > 0$ and such that*

$$\gamma \leq \frac{x_i}{y_i^{q-1}} \leq \Gamma \text{ for } i \in \{1,\dots,n\},$$

then we have

$$\begin{aligned}
&\left|\sum_{i=1}^{n}\left|\frac{x_i^p}{y_i^q} - \left(\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} y_j^q}\right)^p\right| y_i^q sgn\left[\frac{x_i}{y_i^{q-1}} - \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} y_j^q}\right]\right| \qquad (101)\\
&\leq \frac{\sum_{i=1}^{n} x_i^p}{\sum_{i=1}^{n} y_i^q} - \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}\right)^p\\
&\leq \frac{1}{2}\left(\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}, \Gamma; (\cdot)^p\right] - \left[\gamma, \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}; (\cdot)^p\right]\right) \tilde{D}_{y^q}\left(\frac{x}{y^{q-1}}\right)\\
&\leq \frac{1}{2}\left(\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}, \Gamma; (\cdot)^p\right] - \left[\gamma, \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}; (\cdot)^p\right]\right) \tilde{D}_{y^q,2}\left(\frac{x}{y^{q-1}}\right)\\
&\leq \frac{1}{4}\left(\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}, \Gamma; (\cdot)^p\right] - \left[\gamma, \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^q}; (\cdot)^p\right]\right)(\Gamma - \gamma),
\end{aligned}$$

where

$$\tilde{D}_{y^q}\left(\frac{x}{y^{q-1}}\right) = \frac{1}{\sum_{i=1}^{n} y_i^q}\sum_{i=1}^{n} y_i^q\left|\frac{x_i}{y_i^{q-1}} - \frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} y_j^q}\right|$$

and

$$\tilde{D}_{y^q,2}\left(\frac{x}{y^{q-1}}\right) = \left[\frac{1}{\sum_{i=1}^{n} y_i^q}\sum_{i=1}^{n}\frac{x_i^2}{y_i^{q-2}} - \left(\frac{\sum_{j=1}^{n} x_j y_j}{\sum_{j=1}^{n} y_j^q}\right)^2\right]^{\frac{1}{2}}.$$

Proof. The inequalities (102) follow from (100) by choosing

$$z_i = \frac{x_i}{y_i^{q-1}} \text{ and } w_i = y_j^q.$$

The details are omitted.

Remark 11. We observe that for $p = q = 2$ we have from the first inequality in (101) the following reverse of the Cauchy–Bunyakovsky–Schwarz inequality

$$\left|\sum_{i=1}^{n}\left|\frac{x_i^2}{y_i^2}-\left(\frac{\sum_{j=1}^{n}x_jy_j}{\sum_{j=1}^{n}y_j^2}\right)^2\right|y_i^2 sgn\left(\frac{x_i}{y_i}-\frac{\sum_{j=1}^{n}x_jy_j}{\sum_{j=1}^{n}y_j^2}\right)\right| \tag{102}$$

$$\leq \frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}y_i^2}-\left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}y_i^2}\right)^2$$

$$\leq \frac{1}{2}(\Gamma-\gamma)\frac{1}{\sum_{i=1}^{n}y_i^2}\sum_{i=1}^{n}y_i^2\left|\frac{x_i}{y_i}-\frac{\sum_{j=1}^{n}x_jy_j}{\sum_{j=1}^{n}y_j^2}\right|$$

$$\leq \frac{1}{2}(\Gamma-\gamma)\left[\frac{1}{\sum_{i=1}^{n}y_i^2}\sum_{i=1}^{n}x_i^2-\left(\frac{\sum_{j=1}^{n}x_jy_j}{\sum_{j=1}^{n}y_j^2}\right)^2\right]^{\frac{1}{2}}$$

$$\leq \frac{1}{4}(\Gamma-\gamma)^2,$$

provided that there exists the constants $\gamma, \Gamma > 0$ such that

$$\gamma \leq \frac{x_i}{y_i} \leq \Gamma \text{ for } i \in \{1,\ldots,n\}.$$

12 Applications for f-Divergence

Consider the *f-divergence*

$$I_f(p,q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) \tag{103}$$

defined on the set of probability distributions $p, q \in \mathbb{P}^n$, where f is convex on $(0,\infty)$. It is assumed that $f(u)$ is zero and strictly convex at $u = 1$.

Proposition 8 (Dragomir [41]). *Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with the property that $f(1) = 0$. Assume that $p, q \in \mathbb{P}^n$ and there exists the constants $0 < r < 1 < R < \infty$ such that*

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1,\ldots,n\}. \tag{104}$$

Then we have

$$0 \leq I_f(p,q) \leq \frac{1}{2}([1,R;f]-[r,1;f])D_v(p,q) \tag{105}$$

$$\leq \frac{1}{2}\left[f'_-(R)-f'_+(r)\right]D_v(p,q)$$

$$\leq \frac{1}{2}\left[f'_{-}(R)-f'_{+}(r)\right]\left[D_{\chi^2}(p,q)\right]^{1/2}$$
$$\leq \frac{1}{4}(R-r)\left[f'_{-}(R)-f'_{+}(r)\right],$$

where $D_v(p,q)=\sum_{i=1}^{n}|p_i-q_i|$ *and* $D_{\chi^2}(p,q)=\sum_{i=1}^{n}\frac{p_i^2}{q_i}-1$.

Proof. From (90) we have

$$0 \leq \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right)-f(1)$$
$$\leq \frac{1}{2}([1,R;f]-[r,1;f])\sum_{i=1}^{n} q_i\left|\frac{p_i}{q_i}-1\right|$$
$$\leq \frac{1}{2}\left[f'_{-}(R)-f'_{+}(r)\right]\sum_{i=1}^{n} q_i\left|\frac{p_i}{q_i}-1\right|$$
$$\leq \frac{1}{2}\left[f'_{-}(R)-f'_{+}(r)\right]\left(\sum_{i=1}^{n}\frac{p_i^2}{q_i}-1\right)^{1/2} \leq \frac{1}{4}(R-r)\left[f'_{-}(R)-f'_{+}(r)\right]$$

i.e., the desired result (105).

Remark 12. The inequality

$$I_f(p,q) \leq \frac{1}{4}(R-r)\left[f'_{-}(R)-f'_{+}(r)\right] \tag{106}$$

was obtained for the discrete divergence measures in 2000 by S.S. Dragomir, see [32].

Proposition 9 (Dragomir [41]). *With the assumptions in Proposition 8 we have*

$$\left|I_{|f|(sgn(\cdot)-1)}(p,q)\right| \leq I_f(p,q) \tag{107}$$
$$\leq \frac{1}{2}([1,R;f]-[r,1;f])\,D_v(p,q)$$
$$\leq \frac{1}{2}([1,R;f]-[r,1;f])\left[D_{\chi^2}(p,q)\right]^{1/2}$$
$$\leq \frac{1}{4}([1,R;f]-[r,1;f])(R-r),$$

where $I_{|f|(sgn(\cdot)-1)}(p,q)$ *is the generalized f-divergence for the non-necessarily convex function* $|f|(sgn(\cdot)-1)$ *and is defined by*

$$I_{|f|(sgn(\cdot)-1)}(p,q) := \sum_{i=1}^{n} q_i \left| f\left(\frac{p_i}{q_i}\right) \right| sgn\left(\frac{p_i}{q_i} - 1\right). \tag{108}$$

Proof. From the inequality (89) we have

$$\begin{aligned}
&\left| \sum_{i=1}^{n} q_i \left| f\left(\frac{p_i}{q_i}\right) - f(1) \right| sgn\left(\frac{p_i}{q_i} - 1\right) \right| \\
&\leq \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) - f(1) \\
&\leq \frac{1}{2}\left([1,R;f] - [r,1;f]\right) \sum_{i=1}^{n} q_i \left| \frac{p_i}{q_i} - 1 \right| \\
&\leq \frac{1}{2}\left([1,R;f] - [r,1;f]\right) \left(\sum_{i=1}^{n} \frac{p_i^2}{q_i} - 1 \right)^{1/2} \\
&\leq \frac{1}{4}\left([1,R;f] - [r,1;f]\right)(R-r)
\end{aligned}$$

from where we get the desired result (107).

The above results can be utilized to obtain various inequalities for the divergence measures in information theory that are particular instances of f-divergence.

Consider the *Kullback–Leibler divergence*

$$KL(p,q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right),\ p,q \in \mathbb{P}^n.$$

For the convex function $f : (0,\infty) \to \mathbb{R}, f(t) = -\ln t$ we have

$$I_f(p,q) := \sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right) = -\sum_{i=1}^{n} q_i \ln\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{n} q_i \ln\left(\frac{q_i}{p_i}\right) = KL(q,p).$$

If $p,q \in \mathbb{P}^n$ such that there exists the constants $0 < r < 1 < R < \infty$ with

$$r \leq \frac{p_i}{q_i} \leq R \text{ for } i \in \{1,\ldots,n\}, \tag{109}$$

then we get from the first inequality in (105) that

$$0 \leq KL(q,p) \leq \frac{1}{2} D_v(p,q) \ln\left(\frac{1}{R^{R-1} r^{1-r}}\right).$$

For the convex function $f:(0,\infty)\to\mathbb{R}, f(t)=t\ln t$ we have

$$I_f(p,q):=\sum_{i=1}^{n} q_i f\left(\frac{p_i}{q_i}\right)=KL(p,q).$$

If $p,q\in\mathbb{P}^n$ are such that there exists the constants $0<r<1<R<\infty$ with the property (109), then we get from the first inequality in (105) that

$$0\le KL(p,q)\le \frac{1}{2}D_v(p,q)\ln\left(R^{\frac{R}{R-1}}r^{\frac{r}{1-r}}\right).$$

References

1. Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. R. Stat. Soc. Sec. B **28**, 131–142 (1966)
2. Barnett, N.S., Cerone, P., Dragomir, S.S., Sofo, A.: Approximating Csiszár f-divergence by the use of Taylor's formula with integral remainder. Math. Ineq. Appl. **5**(3), 417–434 (2002)
3. Barnett, N.S., Cerone, P., Dragomri, S.S., Sofo, A.: Approximating two mappings associated to Csiszár f-divergence via Taylor's expansion. Pan Am. Math. J. **12**(4), 105–117 (2002)
4. Barnett, N.S., Cerone, P., Dragomir, S.S.: Some new inequalities for Hermite-Hadamard difference in information theory. In: Cho, Y.J., Kim, J.K., Choi, Y.K. (eds.) Stochastic Analysis and Applications, pp. 7–20. Nova Science Publishers, New York (2003)
5. Beth Bassat, M.: f-Entropies, probability of error and feature selection. Inform. Control **39**, 227–242 (1978)
6. Beran, R.: Minimum Hellinger distance estimates for parametric models. Ann. Stat. **5**, 445–463 (1977)
7. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943)
8. Burbea, I., Rao, C.R.: On the convexity of some divergence measures based on entropy function. IEEE Trans. Inf. Theory **28**(3), 489–495 (1982)
9. Cerone, P., Dragomir, S.S.: A refinement of the Grüss inequality and applications. Tamkang J. Math. **38**(1), 37–49 (2007). Preprint RGMIA. Res. Rep. Coll. **5**(2), Art. 14 (2002). Online http://rgmia.org/v5n2.php.
10. Chen, C.H.: Statistical Pattern Recognition. Hoyderc Book Co., Rocelle Park, New York (1973)
11. Chow, C.K., Lin, C.N.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**(3), 462–467 (1968)
12. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
13. Csiszár, I.: Information measures: a critical survey. In: Trans. 7th Prague Conf. on Info. Th., Statist. Decis. Funct., Random Processes and 8th European Meeting of Statist., vol. B, pp. 73–86. Academia Prague, Crechoslovakia (1978)
14. Csiszár, I.: Information-type measures of difference of probability functions and indirect observations. Studia Sci. Math. Hungar **2**, 299–318 (1967)
15. Csiszár, I., Körner, J.: Information Theory: Coding Theorem for Discrete Memoryless Systems. Academic Press, New York (1981)
16. Dacunha-Castelle, D.: Ecole d'ete de Probabilité s de Saint-Fleour, III-1997. Springer, Berlin/Heidelberg (1978)
17. Dragomir, S.S.: An improvement of Jensen's inequality. Bull. Math. Soc. Sci. Math. Roum. **34**(82)(4), 291–296 (1990)

18. Dragomir, S.S.: Some refinements of Ky Fan's inequality. J. Math. Anal. Appl. **163**(2), 317–321 (1992)
19. Dragomir, S.S.: Some refinements of Jensen's inequality. J. Math. Anal. Appl. **168**(2), 518–522 (1992)
20. Dragomir, S.S.: A further improvement of Jensen's inequality. Tamkang J. Math. **25**(1), 29–36 (1994)
21. Dragomir, S.S.: A new improvement of Jensen's inequality. Indian J. Pure Appl. Math. **26**(10), 959–968 (1995)
22. Dragomir, S.S.: Some inequalities for (m, M)-convex mappings and applications for the Csiszár f-divergence in information theory. Math. J. Ibaraki Univ. **33**, 35–50 (2001)
23. Dragomir, S.S.: Some inequalities for two Csiszár divergences and applications. Mat. Bilten **25**, 73–90 (2001)
24. Dragomir, S.S.: On a reverse of Jessen's inequality for isotonic linear functionals. J. Inequal. Pure Appl. Math. **2**(3), Article 36 (2001)
25. Dragomir, S.S.: An upper bound for the Csiszár f-divergence in terms of variational distance and applications. Pan Am. Math. J. **12**(4), 165–173 (2002)
26. Dragomir, S.S.: Upper and lower bounds for Csiszár f-divergence in terms of Hellinger discrimination and applications. Nonlinear Anal. Forum **7**(1), 1–13 (2002)
27. Dragomir, S.S.: Bounds for f-divergences under likelihood ratio constraints. Appl. Math. **48**(3), 205–223 (2003)
28. Dragomir, S.S.: A Grüss type inequality for isotonic linear functionals and applications. Demonstratio Math. **36**(3), 551–562 (2003). Preprint RGMIA. Res. Rep. Coll. **5**, Suplement, Art. 12 (2002). Online http://rgmia.org/v5(E).php
29. Dragomir, S.S.: Some inequalities for the Csiszár f -divergence. J. KSIAM (Korea), **7**(1), 63–77 (2003)
30. Dragomir, S.S.: New inequalities for Csiszár divergence and applications. Acta Math. Vietnam. **28**(2), 123–134 (2003)
31. Dragomir, S.S.: A converse inequality for the Csisz ár f-divergence. Tamsui Oxf. J. Math. Sci. (Taiwan) **20**(1), 35–53 (2004)
32. Dragomir, S.S.: A converse inequality for the Csiszár Φ-divergence. Tamsui Oxf. J. Math. Sci. **20**(1), 35–53 (2004). Preprint in S.S. Dragomir (ed.) Inequalities for Csiszár f-Divergence in Information Theory. RGMIA Monographs, Victoria University (2000). http://rgmia.org/monographs/csiszar_list.html#chap1
33. Dragomir, S.S.: Some inequalities for the Csiszár f-divergence when f is an L-Lipschitzian function and applications. Ital. J. Pure Appl. Math. **15**, 57–76 (2004)
34. Dragomir, S.S.: Semi-inner Products and Applications. Nova Science, New York (2004)
35. Dragomir, S.S.: Discrete Inequalities of the Cauchy-Bunyakovsky-Schwarz Type. Nova Science Publishers, New York (2004)
36. Dragomir, S.S.: Bounds for the normalized Jensen functional. Bull. Aust. Math. Soc. **74**(3), 471–476 (2006)
37. Dragomir, S.S.: Bounds for the deviation of a function from the chord generated by its extremities. Bull. Aust. Math. Soc. **78**(2), 225–248 (2008)
38. Dragomir, S.S.: A refinement of Jensen's inequality with applications for f-divergence measures. Taiwanese J. Math. **14**(1), 153–164 (2010)
39. Dragomir, S.S.: A new refinement of Jensen's inequality in linear spaces with applications. Math. Comput. Model. **52**(9–10), 1497–1505 (2010)
40. Dragomir, S.S.: Inequalities in terms of the Gâteaux derivatives for convex functions on linear spaces with applications. Bull. Aust. Math. Soc. **83**(3), 500–517 (2011)
41. Dragomir, S.S.: A refinement and a divided difference reverse of Jensen's inequality with applications. Preprint RGMIA. Res. Rep. Coll. **14**, Art. 74 (2011). Online http://rgmia.org/papers/v14/v14a74.pdf

42. Dragomir, S.S.: Some Slater's type Inequalities for convex functions defined on linear spaces and applications. Abstr. Appl. Anal. 2012, 1–16 (2012). doi:10.1155/2012/168405. http://projecteuclid.org/euclid.aaa/135549564.MR2889076: http://www.ams.org/mathscinet-getitem?mr=2889076
43. Dragomir, S.S.: Some reverses of the Jensen inequality with applications. Bull. Aust. Math. Soc. **87**(2), 177–194 (2013)
44. Dragomir, S.S.: Reverses of the Jensen inequality in terms of first derivative and applications. Acta Math. Vietnam. **38**(3), 429–446 (2013)
45. Dragomir, S.S., Goh, C.J.: A counterpart of Jensen's discrete inequality for differentiable convex mappings and applications in information theory. Math. Comput. Model. **24**(2), 1–11 (1996)
46. Dragomir, S.S., Goh, C.J.: Some bounds on entropy measures in information theory. Appl. Math. Lett. **10**, 23–28 (1997)
47. Dragomir, S.S., Goh, C.J.: Some counterpart inequalities in for a functional associated with Jensen's inequality. J. Inequal. Appl. **1**, 311–325 (1997)
48. Dragomir, S.S., Goh, C.J.: A counterpart of Jensen's continuous inequality and applications in information theory. Anal. St. Univ. "Al. I. Cuza", Iaşi, **XLVII**, 239–262 (2001)
49. Dragomir, S.S., Ionescu, N.M.: Some converse of Jensen's inequality and applications. Rev. Anal. Numér. Théor. Approx. **23**(1), 71–78 (1994)
50. Dragomir, S.S., Pečarić, J., Persson, L.E.: Properties of some functionals related to Jensen's inequality. Acta Math. Hung. **70**(1–2), 129–143 (1996)
51. Dragomir, S.S., Scholz, M., Šunde, J.: Some upper bounds for relative entropy and applications. Comput. Math. Appl. **39**, 91–100 (2000)
52. Frieden, B.R.: Image enhancement and restoration. In: Huang, T.S. (ed.) Picture Processing and Digital Filtering. Springer, Berlin (1975)
53. Gallager, R.G.: Information Theory and Reliable Communications. Wiley, New York (1968)
54. Gokhale, D.V., Kullback, S.: Information in Contingency Tables. Merul Dekker, New York (1978)
55. Havrda, J.H., Charvat, F.: Quantification method classification process: concept of structural α-entropy. Kybernetika **3**, 30–35 (1967)
56. Hellinger, E.: Neue Bergrüirdung du Theorie quadratisher Formerus von uneudlichvieleu Veränderlicher. J. für reine Augeur. Math. **36**, 210–271 (1909)
57. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proc. Roy. Soc. Lond. Ser. A **186**, 453–461 (1946)
58. Jistice, J.H. (ed.): Maximum Entropy and Bayesian Methods in Applied Statistics. Cambridge University Press, Cambridge (1986)
59. Kadota, T.T., Shepp, L.A.: On the best finite set of linear observables for discriminating two Gaussian signals. IEEE Trans. Inf. Theory **13**, 288–294 (1967)
60. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. Commun. Technol. **15**, 52–60 (1967)
61. Kapur, J.N.: A comparative assessment of various measures of directed divergence. Adv. Manage. Stud. **3**(1), 1–16 (1984)
62. Kapur, J.N.: On the roles of maximum entropy and minimum discrimination information principles in statistics. In: Technical Address of the 38th Annual Conference of the Indian Society of Agricultural Statistics, pp. 1–44 (1984)
63. Kullback, S.: Information Theory and Statistics. Wiley, New York (1959)
64. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
65. Kazakos, D., Cotsidas, T.: A decision theory approach to the approximation of discrete probability densities. IEEE Trans. Perform. Anal. Mach. Intell. **1**, 61–67 (1980)
66. Kemperman, J.H.B.: On the optimum note of transmitting information. Ann. Math. Stat. **40**, 2158–2177 (1969)
67. Kraft, C.: Some conditions for consistency and uniform consistency of statistical procedures. Univ. Calif. Publ. Stat. **1**, 125–142 (1955)

68. Leahy, R.M., Goutis, C.E.: An optimal technique for constraint-based image restoration and mensuration. IEEE Trans. Acoust. Speech Signal Process. **34**, 1692–1642 (1986)
69. Lecam, L.: Asymptotic Methods in Statistical Decision Theory. Springer, New York (1986)
70. Liese, F., Vajda, I.: Convex Statistical Distances. Teubner Verlag, Leipzig (1987)
71. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theroy **37**(1), 145–151 (1991)
72. Lin, J., Wong, S.K.M.: A new directed divergence measure and its characterization. Int. J. Gen. Syst. **17**, 73–81 (1990)
73. Matić, M., Pearce, C.E.M., Pečarić, J.: Improvements of some bounds on entropy measures in information theory. Math. Inequal. Appl. **1**, 295–304 (1998)
74. Matić, M., Pečarić, J., Ujević, N.: On new estimation of the remainder in generalised Taylor's formula. Math. Inequal. Appl. **2**(3), 343–361 (1999)
75. Mckean, H.P. Jr.: Speed of approach to equilibrium for Koc's caricature of a Maximilian gas. Arch. Ration. Mech. Anal. **21**, 343–367 (1966)
76. Mei, M.: The theory of genetic distance and evaluation of human races. Jpn. J. Hum. Genet. **23**, 341–369 (1978)
77. Pielou, E.C.: Ecological Diversity. Wiley, New York (1975)
78. Pinsker, M.S.: Information and Information Stability of Random variables and processes (in Russian). Izv. Akad. Nouk, Moscow (1960)
79. Rao, C.R.: Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. **21**, 24–43 (1982)
80. Rényi, A.: On measures of entropy and information. In: Proc. Fourth Berkeley Symp. Math. Stat. and Prob., vol. 1, pp. 547–561. University of California Press, Berkeley (1961)
81. Sen, A.: On Economic Inequality. Oxford University Press, London (1973)
82. Sharma, B.D., Mittal, D.P.: New non-additive measures of relative information. J. Comb. Inf. Syst. Sci. **2**(4), 122–132 (1977)
83. Shioya, H., Da-Te, T.: A generalisation of Lin divergence and the derivative of a new information divergence. Electron. Commun. Jpn. **78**(7), 37–40 (1995)
84. Simić, S.: On a global upper bound for Jensen's inequality. J. Math. Anal. Appl. **343**, 414–419 (2008)
85. Taneja, I.J.: Generalised Information Measures and Their Applications. http://www.mtm.ufsc.br/~taneja/bhtml/bhtml.html (2001)
86. Theil, H.: Economics and Information Theory. North-Holland, Amsterdam (1967)
87. Theil, H.: Statistical Decomposition Analysis. North-Holland, Amsterdam (1972)
88. Topsoe, F.: Some inequalities for information divergence and related measures of discrimination. Res. Rep. Coll. RGMIA **2**(1), 85–98 (1999)
89. Toussaint, G.T.: Sharper lower bounds for discrimination in terms of variation. IEEE Trans. Inf. Theory **21**, 99–100 (1975)
90. Vajda, I.: Theory of Statistical Inference and Information. Kluwer, Boston (1989)
91. Vajda, I.: Note on discrimination information and variation. IEEE Trans. Inf. Theory **16**, 771–773 (1970)
92. Volkonski, V.A., Rozanov, J.A.: Some limit theorems for random function -*I* (English Trans.). Theory Probab. Appl. **4**, 178–197 (1959)

On Geometry of the Zeros of a Polynomial

N.K. Govil and Eze R. Nwaeze

Abstract Let $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$ be a polynomial of degree n, where the coefficients a_k may be complex. The problem of locating the zeros of a polynomial $p(z)$ is a long-standing classical problem which has frequently been investigated. These problems, besides being of theoretical interest, have important applications in many scientific specialization areas, such as coding theory, cryptography, combinatorics, number theory, mathematical biology, engineering, signal processing, communication theory, and control theory, and for this reason there is always a need for better and sharper results.

This paper is expository in nature, and here we make an attempt to provide a systematic study of these problems by presenting some results starting from the results of Gauss and Cauchy, who we believe were the earliest contributors in this subject, to some of the most recent ones. When possible, we have tried to present the proofs of some of the theorems. Also, included here are some results on evaluating the quality of bounds by using numerical methods or MATLAB.

Keywords: Complex polynomials • Location of zeros of polynomials • Complex zeros • Inequalities • Trinomials and quadrinomials

1 Introduction

Let $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$ be a polynomial of degree n. By the Fundamental Theorem of Algebra (historically, the first important result concerning the roots of an algebraic equation), $p(z)$ has exactly n zeros in the complex plane, counting multiplicity. But this Theorem does not say anything regarding the location of zeros of polynomial, that is, the region which contains some or all of the zeros of a polynomial. Problems involving location of the zeros of a polynomial, besides being of theoretical interest, find applications in many areas of applied mathematics such as coding theory, cryptography, combinatorics, number

N.K. Govil (✉) • E.R. Nwaeze
Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA
e-mail: govilnk@auburn.edu; ern0002@auburn.edu

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_10

theory, mathematical biology, and engineering [5, 12, 37, 53, 58, 60]. Especially, the polynomial zeros play an important role, for example, in solving digital audio signal processing problems [75], control engineering problems [11], and eigenvalue problems in mathematical physics [64]. Since Abel and Ruffini proved that there is no general algebraic solution to polynomial equations of degree five or higher, the problem of finding a region containing all the zeros of a polynomial became much more interesting, and over a period a large number of results have been provided in this direction.

It may be remarked that there are methods, for example Ehrlich–Aberth's type (see [1, 29, 57]) for the simultaneous determination of the zeros of algebraic polynomials, and there are studies to accelerate convergence and increase computational efficiency of these methods (for example, see [51, 54]). These methods which are of course very useful, because they give approximations to the zeros of a polynomial can possibly become more efficient when combined with the results dealing with the region containing all the zeros of a polynomial, because an accurate estimate of the region or annulus containing all the zeros of a polynomial can considerably reduce the amount of work needed to find exact zeros, and so there is always a need for better and better estimates for the region containing all the zeros of a polynomial. Several books, and monographs have been written on this subject and related subject of approximation theory (for example, see [49, 52, 53, 61]).

The problems concerning the location of the zeros of a polynomial can mainly be divided into two categories, namely:

- Given an integer k, $1 \leq k \leq n$, find a region $R = R(a_0, a_1, a_2, \ldots, a_n)$ containing at least k or exactly k zeros of $p(z)$. In other words, one would like to find the smallest circle $|z| = r$ which will enclose the k zeros of the polynomial. Such results are very useful for solving practical problems in numerical analysis, for example in finding the roots of an algebraic equation by using Newton–Raphson Method, and in finding eigenvalues. Note that when dealing with the problems of finding eigenvalues often one is not interested in computing all eigenvalues precisely.
- Given a region R, to find the number $k = k(a_0, a_1, a_2, \ldots, a_n)$ such that k number of zeros lie in the region R. In particular, to find the number k of zeros whose moduli do not exceed some prescribed value, say r.

The subject of location of zeros of a polynomial has been studied extensively dating back to Gauss and Cauchy to some of the more recent ones. Due to the limited space, it would not be possible to include all the results in this subject, and therefore many important results in this area, which we would have liked to include, had to be excluded (for a more detailed study of the subject, we refer, in particular, to the monograph and books written by Dieudonné [27], Marden [49], Milovanović et al. [53], and Rahman and Schmeisser [61]).

This paper contains five sections. Section 1 being on Introduction where we present a brief introduction and justification to study this subject of Geometry of the Zeros of Polynomials. In Sect. 2, we give a brief history of the subject of the Geometry of Zeros of Polynomials starting with the earliest results of Gauss

and Cauchy on this subject and then develop it by presenting some results in this direction. In Sect. 3, presented are some results concerning the Location of the Zeros of Composite Polynomials, on Linear Combination of Polynomials, and also some results of Peretz and Rassias in this direction. Section 4 deals with the Location of the Zeros of Lacunary Polynomials, along with some recent results concerning the Zeros of Trinomials and Quadrinomials, and finally in Sect. 5 some recent results concerning Cauchy Theorem on the Location of the Zeros of a Polynomial, both in terms of disks and annuli containing all the zeros of a polynomial, have been presented. When possible, we have tried to present proofs of some of the results presented in this paper.

Also, presented in this section are some examples of polynomials to compare bounds obtained by different results, and this has been done by using numerical methods or MATLAB.

2 Results Due to Gauss, Cauchy, and Bounds for Zeros as Functions of All the Coefficients

The earliest result concerning the location of the zeros of a polynomial is probably due to Gauss who incidental to his proofs of the Fundamental Theorem of Algebra showed in 1816 that a polynomial

$$p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n,$$

with all a_j real, has no zeros outside certain circles $|z| = R$, where

$$R = \max_{1 \le j \le n} (n 2^{1/2} |a_j|)^{1/j}.$$

However, in the case of arbitrary real or complex a_j, Gauss [33] in 1849 showed that R may be taken as the positive root of the equation

$$z^n - 2^{1/2}(|a_1| z^{n-1} + \cdots + |a_n|) = 0.$$

As a further indication of Gauss' interest in the location of the zeros of a polynomial, we have his letter (see collected works of Gauss) to Schumacher dated April 2, 1833, in which he tells of having written enough on this topic to fill several volumes, but the only results he published are those in Gauss [33]. Even, his important result, Theorem 2.1 stated below on the mechanical interpretation of the zeros of the derivative of a polynomial comes to us only by a brief entry he made presumably around 1836 in a notebook otherwise devoted to astronomy.

Theorem 2.1. *The zeros of the function* $F(z) = \sum_{j=1}^{k} \frac{m_j}{z - z_j}$, *where all* m_j *are real, are the points of the equilibrium in the field of force due to the system of k masses* m_j *at the fixed points* z_j *repelling a unit movable mass at z according to the inverse distance law.*

Around 1829, Cauchy [14] (also, see the book of Marden [49, Theorem 27.1, p. 122]) derived more exact bounds for the moduli of the zeros of a polynomial than those given by Gauss, by proving the following:

Theorem 2.2. *Let* $p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$ *be a complex polynomial, then all the zeros of* $p(z)$ *lie in the disc*

$$\{z : |z| \leq \eta\} \subset \{z : |z| < 1 + A\}, \tag{1}$$

where

$$A = \max_{0 \leq j \leq n-1} |a_j|,$$

and η *is the unique positive root of the real coefficient equation*

$$z^n - |a_{n-1}|z^{n-1} - |a_{n-2}|z^{n-2} - \cdots - |a_1|z - |a_0| = 0. \tag{2}$$

The result is best possible and the bound is attained when $p(z)$ *is the polynomial on the left-hand side of (2).*

The proof follows easily from the inequality

$$|p(z)| \geq |z|^n - (|a_{n-1}||z|^{n-1} + |a_{n-2}||z|^{n-2} + \cdots + |a_1||z| + |a_0|) = 0, \tag{3}$$

which can be derived easily on applying Triangle Inequality to $p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$.

If one applies the above Theorem 2.2 of Cauchy to the polynomial $P(z) = z^n p(1/z)$ and combine it with Theorem 2.2, one easily gets

Theorem 2.3 (Cauchy). *All the zeros of the polynomial* $p(z) = a_0 + a_1 z + \cdot + a_n z^n$, $a_n \neq 0$, *lie in the annulus* $r_1 \leq |z| \leq r_2$, *where* r_1 *is the unique positive root of the equation*

$$|a_n|z^n + |a_{n-1}|z^{n-1} + \cdot + |a_1|z - |a_0| = 0, \tag{4}$$

and r_2 *is the unique positive root of the equation*

$$|a_0| + |a_1|z + \cdot + |a_{n-1}|z^{n-1} - |a_n|z^n = 0. \tag{5}$$

Although the above result of Cauchy gives an annulus containing all the zeros of a polynomial, it is implicit, in the sense, that in order to find the annulus containing all the zeros of a polynomial, one needs to compute the zeros of two other polynomials. From the inequality (3) as well follows the following result which is also due to Cauchy [14].

Theorem 2.4. *Let* $p(z) = \sum_{j=0}^{n} a_j z^j$ *be a complex polynomial with* $a_n \neq 0$, *then all the zeros of* $p(z)$ *lie in the disc*

$$T = \left\{ z : |z| < 1 + \max_{0 \leq j \leq n-1} \left| \frac{a_j}{a_n} \right| \right\}.$$

To prove Theorem 2.4, note that if $M = \max_{0 \leq j \leq n-1} \left| \frac{a_j}{a_n} \right|$, and if $|z| > 1$, we get from inequality (3) that

$$\begin{aligned} |f(z)| &\geq |a_n||z|^n \left\{ 1 - M \sum_{j=1}^{n} |z|^{-j} \right\} \\ &> |a_n||z|^n \left\{ 1 - M \sum_{j=1}^{\infty} |z|^{-j} \right\} \\ &> |a_n||z|^n \left\{ 1 - \frac{M}{|z| - 1} \right\} \\ &= |a_n||z|^n \left\{ \frac{|z| - 1 - M}{|z| - 1} \right\}. \end{aligned}$$

Hence, if $|z| \geq 1 + M$, then $|f(z)| > 0$, implying that the only zeros of $f(z)$ in $|z| > 1$ are those in T (as defined in Theorem 2.4). But, as all the zeros of $f(z)$ in $|z| \leq 1$ belong to T also, we have fully established Theorem 2.4.

The inequality (3) also yields the following result due to Birkhoff [10], which was later proved independently by Cohn [15] and by Berwald [7].

Theorem 2.5. *The zero* z_1 *of largest modulus of* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, *satisfies the inequalities*

$$(2^{1/n} - 1)r \leq \alpha \leq |z_1| \leq r \leq \frac{\alpha}{(2^{1/n} - 1)}, \tag{6}$$

where r *is the positive root of (2) and* α *is defined as:*

$$\alpha = \max_{1\le j\le n} \left| \frac{a_{n-j}}{a_n C_j^n} \right|^{1/j} \le |z_1|.$$

Here, as usual, C_j^n are the binomial coefficients defined by

$$C_j^n = \frac{n!}{j!(n-j)!}, \quad 0! = 1. \tag{7}$$

The following result is due to Kuniyeda [45] (also, see [22]), Montel [55], and Tôya [68].

Theorem 2.6. *For any p and q such that*

$$p > 1, \ q > 1, \ \frac{1}{p} + \frac{1}{q} = 1, \tag{8}$$

the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, has all its zeros in the circle

$$|z| < \left\{ 1 + \left[\sum_{j=0}^{n-1} \left| \frac{a_j}{a_n} \right|^p \right]^{q/p} \right\}^{1/q} \le (1 + n^{q/p} M^q)^{1/q}, \tag{9}$$

where $M = \max_{0\le j\le n-1} |a_j/a_n|$.

In particular, if we take $p = q = 2$ in inequality (9), we get that $p(z)$ in Theorem 2.6 has all its zeros in

$$|z| < \left\{ 1 + \sum_{j=0}^{n-1} \left| \frac{a_j}{a_n} \right|^2 \right\}^{1/2}. \tag{10}$$

The above inequality (10) has been derived in Carmichael–Mason [13], Kelleher [42], and Fujiwara [32].

Note that as $p \to \infty$, the right side of (9) approaches the limit $(1 + M)$ and thus Theorem 2.4 can be obtained as a special case of Theorem 2.6. If we apply inequality (10) to the polynomial $(1 - z)(a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n)$, $a_n \neq 0$, we easily get the following result of Williams [70].

Theorem 2.7. *All the zeros of the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, lie in the disk*

$$|z| \le \left[1 + \left| \frac{a_0}{a_n} \right|^2 + \left| \frac{a_1 - a_0}{a_n} \right|^2 + \cdots + \left| \frac{a_n - a_{n-1}}{a_n} \right|^2 \right]^{1/2}. \tag{11}$$

In the paper that contains Theorem 2.6, Kuniyeda [45] also proved.

Theorem 2.8. *For any $p > 0$, the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, has all its zeros in the disk*

$$|z| \leq \left\{ 1 + \frac{1}{|a_n|^{\frac{p+1}{p}}} \left[\sum_{j=1}^{n} |a_{n-j}|^{1+p} \right]^{1/p} \right\}^{\frac{p}{p+1}}. \tag{12}$$

Next, we mention the following result due to Walsh [69], Markovitch [50], Kojima [44], and Joyal et al. [40].

Theorem 2.9 (Walsh [69]). *All the zeros of the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, lie in the disk*

$$|z| \leq \sum_{j=1}^{n} \left| \frac{a_{j-n}}{a_n} \right|^{1/j}. \tag{13}$$

Theorem 2.10 (Markovitch [50]). *All the zeros of the polynomial $h(z) = \sum_{j=0}^{n} a_j b_j z^j$ lie in the disk $|z| \leq Mr$, where r is the positive root of the equation*

$$|a_0| + |a_1| z + \cdots + |a_{n-1}| z^{n-1} - |a_n| z^n = 0, \tag{14}$$

and $M = \max_{0 \leq j \leq n-1} \left| \frac{b_j}{b_{j-1}} \right|^{1/n-j}$.

Theorem 2.11 (Kojima [44]). *All the zeros of the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, lie in the disk*

$$|z| \leq \max_{1 \leq j \leq n-1} \left(\frac{|a_0|}{|a_1|}, 2 \frac{|a_j|}{|a_{j+1}|} \right). \tag{15}$$

Theorem 2.12 (Joyal et al. [40]). *For any p and q such that*

$$p > 1, \; q > 1, \; \frac{1}{p} + \frac{1}{q} = 1, \tag{16}$$

the polynomial $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, has all its zeros in the circle

$$|z| < \left\{ \frac{1}{2} \left[1 + \sqrt{1 + 4 M_p^q} \right] \right\}^{1/q}, \tag{17}$$

where $M_p^q = \left(\sum_{j=1}^{n} \left| \frac{a_{n-1} a_{n-j} - a_n a_{n-j-1}}{a_n^2} \right|^p \right)^{1/p}$, $a_{-1} = 0$.

3 Zeros of Composite Polynomials, Linear Combination of Polynomials, and Some Results of Peretz and Rassias

3.1 Grace's Apolarity Theorem and Its Applications to Zeros of Polynomials

In the beginning of the last century, Grace [36] introduced the following concept of apolar polynomials.

Definition 3.1. Two polynomials $p(z) = \sum_{j=0}^{n} a_j C_j^n z^j$ and $q(z) = \sum_{j=0}^{n} b_j C_j^n z^j$ are said to be apolar if their coefficients satisfy the apolarity condition

$$\sum_{j=0}^{n} (-1)^j C_j^n a_j b_{n-j} = 0. \tag{18}$$

In the same paper, Grace [36] also proved the following result, known as Grace's Apolarity Theorem, or simply Grace's Theorem, which has been found to be of great use, and applications. Before we state this result of Grace we would like to introduce the definition of circular domain used in this paper.

Definition 3.2 (See, Rahaman and Schmeisser [61, p. 96]). Let $\hat{\mathbb{C}}$ be the extended complex plane. The rational functions of degree one, usually called Möbius transformations or linear transformations,

$$\psi : \begin{cases} \hat{\mathbb{C}} \to \hat{\mathbb{C}} \\ z \mapsto \dfrac{\alpha z + \beta}{\gamma z + \delta} \end{cases} \quad (\alpha, \beta, \gamma, \delta \in \mathbb{C}, \alpha\delta - \gamma\beta \neq 0)$$

are also bijective. Let $\mathcal{C}$ denote the set of all circles and all straight lines in the plane. Then every ψ maps an element of $\mathcal{C}$ onto an element of $\mathcal{C}$. Since the mappings ψ are bijective, it follows that, in $\hat{\mathbb{C}}$, every domain whose boundary belongs to $\mathcal{C}$ is mapped onto a domain of the same type. Such domains are called circular domains provided that they are open or closed. *Thus, not only a disc (open or closed), but also its compliment with respect to $\hat{\mathbb{C}}$, is also a circular domain, and so is a half-plane.*

Theorem 3.1. *Let the polynomials $p(z) = \sum_{j=0}^{n} a_j C_j^n z^j$ and $q(z) = \sum_{j=0}^{n} b_j C_j^n z^j$ be apolar. Then any circular domain that contains all the zeros of the polynomial $p(z)$ must contain at least one zero of the polynomial $q(z)$.*

Szegö [67] gave an alternative proof of the above theorem of Grace [36], and also gave several applications. Another proof of this theorem was given by Goodman and Schoenberg [34] (also, see Milovanović et al. [53, p. 188]) for which they use induction on n.

The following applications of Grace's Theorem can be found in Szegö [67] (also in the book of Marden [49], Milovanović et al. [53, p. 191] and in paper of Schur [65]).

Theorem 3.2. *If all the zeros of the polynomial* $p(z) = \sum_{j=0}^{n} a_j C_j^n z^j$ *lie in* $|z| < r$ *and all the zeros of the polynomial* $q(z) = \sum_{j=0}^{n} b_j C_j^n z^j$ *lie in* $|z| \le \rho$, *then all the zeros of the polynomial* $\sum_{j=0}^{n} C_j^n a_j b_j z^j$ *are in* $|z| < r\rho$.

Theorem 3.3 (Schur-Szegö Composite Theorem). *If all the zeros of the polynomial* $p(z) = \sum_{j=0}^{n} a_j C_j^n z^j$ *lie in a closed and bounded convex domain D and all the zeros of the polynomial* $q(z) = \sum_{j=0}^{n} b_j C_j^n z^j$ *lie in* $[-1, 0]$, *then all the zeros of the zeros of the polynomial* $\sum_{j=0}^{n} C_j^n a_j b_j z^j$ *are in D.*

By using Theorem 3.1 of Grace, Szegö [67] in his paper, has obtained.

Theorem 3.4. *Suppose the polynomial* $p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$ *has no zeros in the disk* $|z| \le R$. *Then the "section"* $q(z) = p(z) - z^n = \sum_{j=0}^{n-1} a_j z^j$ *has no zeros in the circular region* $|z| \le R/2$.

3.2 Zeros of Linear Combination of Polynomials

By using Grace Theorem [36], Rubinstein [63] proved several results for the linear combination of polynomials with complex coefficients, and here we begin with the following.

Theorem 3.5. *Let the polynomials* $f(z) = z^n + \cdots$, *and* $g(z) = z^r + \cdots$, $n = 2r$, *have zeros in the circles* $|z-a| \le r_1$ *and* $|z-b| \le r_2$, *respectively, then all the zeros of the polynomial*

$$f(z) - \lambda g(z) \tag{19}$$

are in the union of the n circles

$$\left|z-a-\frac{1}{2}\lambda^{2/n}+\lambda^{1/n}\left(a-b+\frac{1}{4}\lambda^{2/n}\right)^{1/2}\right| \leq (r_1+r_2)^{1/2}|\lambda|^{1/n}+r_1, \tag{20}$$

where $\lambda^{1/n}$ assumes all the nth roots of λ.

Proof. The equation $f(z)-\lambda g(z)=0$ can be replaced by Grace's theorem by the equation $(z-\alpha)^n-\lambda(z-\beta)^{n/2}=0$, where $|\alpha-a|\leq r_1$, and $|\beta-b|\leq r_2$.

Solving for z we obtain

$$z=\alpha+\frac{1}{2}\lambda^{2/n}\pm\lambda^{1/n}\left[(\alpha-\beta)+\frac{1}{4}\lambda^{2/n}\right]^{1/2}.$$

Denoting generically the region $|z-c|\leq R$ by $C(c,R)$ we have

$$\alpha-\beta\in C(a-b,r_1+r_2),$$

implying

$$\left(\alpha-\beta+\frac{1}{4}\lambda^{2/n}\right)^{1/2}\in C\left(\pm\left(a-b+\frac{1}{4}\lambda^{2/n}\right),(r_1+r_2)^{1/2}\right).$$

Hence

$$z\in C\left(a+\frac{1}{2}\lambda^{2/n}\pm\lambda^{1/n}\left(a-b+\frac{1}{4}\lambda^{2/n}\right)^{1/2},(r_1+r_2)^{1/2}|\lambda|^{2/n}+r_1\right).$$

and (20) follows, since by assumption n is an even number. The result is sharp for $\lambda=0$, and for $a=b$. □

For the general case Rubinstein proved, in the same paper

Theorem 3.6. [1]*Let $f(z)=z^n+\cdots$, and $g(z)=z^r+\cdots$, $n>r$, have zeros in the circles $|z-a|\leq r_1$ and $|z-b|\leq r_2$, respectively. Then all the zeros of the polynomial $f(z)-\lambda g(z)$ are in the circle*

$$|z-a|\leq r_1+d,$$

where d is the positive root of the equation

$$d^{n/r}-Md-N=0 \tag{21}$$

with $M=|\lambda|^{1/r}$, $N=|\lambda|^{1/r}(|a-b|+r_1+r_2)$.

Proof. Consider the equation

$$(z-\alpha)^n=\lambda(z-\beta)^r,\qquad |a-\alpha|\leq r_1,\qquad |b-\beta|\leq r_2.$$

[1]Theorem 3.6 was proved independently and by a different method by Mishael Zedek [72].

For z_0 satisfying $(z_0-\alpha)^n = \lambda(z_0-\alpha)^r$, $(z_0-\alpha)^{n/r-1} = \lambda^{1/r}[(z_0-\beta)/(z_0-\alpha)]$. Let d_1 be a positive number satisfying

$$d_1^{n/r} - Md_1 - N > 0.$$

For $|z_0-\alpha| \geq d_1$, $(z_0-\beta)/(z_0-\alpha)$ belongs to the circle $|z-1| \leq |\alpha-\beta|/d_1$; hence

$$\left|\lambda^{1/r}\frac{z_0-\beta}{z_0-\alpha}\right| \leq |\lambda|^{1/r}\left(1+\frac{|\alpha-\beta|}{d_1}\right),$$

but

$$|z_0-\alpha|^{n/r-1} \geq d_1^{n/r-1} > |\lambda|^{1/r}\left(1+\frac{|\alpha-\beta|}{d_1}\right),$$

for all α, β such that $|\alpha-a| \leq r_1$, and $|\beta-b| \leq r_2$. We get a contradiction, which proves that $|z_0-\alpha| < d_1$.

It is worthwhile to remark that if $M+N > 1$ an estimate for the positive zero d is the expression

$$\frac{(n-r)(M+N)^{n/n-r}+rN}{(n-r)(M+N)+rN} \leq (M+N)^{r/n-r}.$$

For $M+N < 1$ a bound for the same is $[(n-r+rN)/(n-rM)] \leq 1$. □

Different estimates can be obtained by means of estimates similar to those used in the proof of Theorem 3.6, which are sharp for $\lambda = 0$ or asymptotically for $\lambda \to \infty$. We indicate some of them which are of a relatively simple form.

Theorem 3.7. *Let $f(z)$ and $g(z)$ be as in Theorem 3.6. All the zeros of the polynomial $f(z)-\lambda g(z)$ are in each of the following regions:*

$$|z| \leq \frac{|a|-r_1}{d(|a|-r_1)-1}\big[(|b|+r_2)d+1\big], \tag{22}$$

where $r > n$, $d = |\lambda|^{1/r}(r_1+|a|)^{-n/r}$, and $d(|a|-r_1)-1 > 0$.

$$|z-b| \leq r_2 + 2\max\Big[|\lambda|^{-(1/r-n)}, (|a-b|+r_1+r_2)^{n/r}|\lambda|^{-(1/r)}\Big], \tag{23}$$

where $r = nk$, $k \geq 2$.

$$\left|z-\frac{\delta_k b}{\delta_k-1}\right| \leq \frac{m+|\delta_k|(r_2+1)}{|\delta_k-1|}, \quad k = 1,\ldots,n \tag{24}$$

where $n > r$, $w_k^n = \lambda$, $\delta_k^n = \lambda/(1-\lambda)$, $k = 1, \ldots, n$;

$$m = \max_{1\le k\le n} \frac{1}{|1-w_k|}(|a - w_k b| + r_1 + |w_k| r_2).$$

For the proof of inequalities (22)–(24), see [63].

We continue presenting some more results due to Rubinstein [63] concerning zeros of the linear combination of polynomials.

Theorem 3.8. *At least n zeros of the polynomial* $(z-\alpha)^n - \lambda(z-\beta)^r$ *are in the circle*

$$|z-\alpha| \le \begin{cases} \dfrac{n}{r-n}|\alpha-\beta|, & n < r \le 2n, \\ |\alpha-\beta|, & r \ge 2n, \end{cases}$$

and at most n zeros of the above polynomial are in the circle

$$|z-\alpha| \le \begin{cases} |\alpha-\beta|, & n < r \le 2n, \\ \dfrac{n}{r-n}|\alpha-\beta|, & r \ge 2n, \end{cases}$$

for all complex λ.

The following theorem, which is also due to Rubinstein [63], generalizes a result due to Biernacki and Jankowski (see [9, 39]).

Theorem 3.9. *Let* $P(z) = a_p z^p + a_{p-s} z^{p-s} + \cdots + a_0$, $Q(z) = b_q z^q + b_{q-t} z^{q-t} + \cdots + b_0$, $a_p b_q \ne 0$, $q > p$, $s \ge 1$, $t \ge 1$ *have all their zeros in the circles* $|z| \le R_1$ *and* $|z| \le R_2$, *respectively. Let* $r = \min(s,t) \ge 1$. *Then at least p zeros of the polynomial*

$$P(z) + \lambda Q(z)$$

are in the circle

$$|z| \le \max\left\{ \left(\frac{qR_1^r + pR_2^r}{q-p} \right)^{1/r}, R_2 \right\}.$$

We conclude this section by stating the following result, which can be found in the book of Milovanović et al. [53].

Theorem 3.10. *If all the zeros of a polynomial* $p(z) = \sum_{j=0}^{n} a_j z^j$ *lie in a circle* $|z| \le R$, *then for any a all the zeros of the polynomial* $p(z) - a$ *lie in the disk* $|z| \le R + |a/a_n|^{1/n}$.

3.3 Some Results of Peretz and Rassias

In his book, Marden [49, p. 68–70] states two theorems which are supposed to be restatements of his results in Marden [48].

Theorem 3.11. *Let* $P(z) = \sum_{j=0}^{m} a_j z^j$, $Q(z) = \sum_{j=0}^{n} b_j z^j$, *and* $R(z) = \sum_{j=0}^{m} a_j \times Q(j) z^j$. *If all the zeros of the polynomial* $P(z)$ *lie in the ring*

$$R_0 = \{z : 0 \le r_1 \le |z| \le r_2 \le \infty\}, \tag{25}$$

and if all the zeros of the polynomial $Q(z)$ *lie in the ring*

$$A = \{z : 0 \le \rho_1 \le |z|/|z - m| \le \rho_2 \le \infty\}, \tag{26}$$

then all the zeros of the polynomial $R(z)$ *lie in the ring*

$$R_n = \{z : 0 \le r_1 \min(1, \rho_1^n) \le |z| \le r_2 \max(1, \rho_1^n)\}. \tag{27}$$

Theorem 3.12. *Let* $P(z) = \sum_{j=0}^{m} a_j z^j$, $Q(z) = \sum_{j=0}^{n} b_j z^j$, *and* $R(z) = \sum_{j=0}^{m} a_j \times Q(j) z^j$. *If all the zeros of the polynomial* $P(z)$ *lie in the ring* $R_0 = \{z : 0 \le r_1 \le |z| \le r_2 \le \infty\}$, *then all the zeros of the polynomial* $R(z)$ *lie in the ring*

$$r_1 \min\big[1, |Q(0)/Q(m)|\big] \le |z| \le r_2 \max\big[1, |Q(0)/Q(m)|\big]. \tag{28}$$

Theorem 3.11 is a part of Marden's corollary in [48] whereas Theorem 3.12 is not included there.

In 1992, Peretz and Rassias [59] proved that Theorem 3.12 is, in fact, false. For this, they constructed a counterexample, by taking $P(z) = 1 + 2z + z^2 = (1 + z)^2$ and $Q(z) = 1 + 2z - z^2$. For these polynomials $n = m = 2$, $Q(0) = 1$, $Q(1) = 2$, and $Q(2) = 1$, and therefore $R(z) = 1 + 4z + z^2$. Note that $P(z)$ has a double zero at $z = -1$ and so we can take $r_1 = r_2 = 1$. Since $Q(0)/Q(2) = 1$, by Theorem 3.12 all the zeros of the polynomial $R(z)$ should lie on $|z| = 1$ while, as can be easily seen, its zeros are $-2 + \sqrt{3}$ and $-2 - \sqrt{3}$, which obviously do not lie on $|z| = 1$.

After establishing that Theorem 3.12 is false, in the same paper Peretz and Rassias [59] prove a correct version of Theorem 3.12, for which they introduced the following definition.

Definition 3.3. Let $Q(z) = (\beta_1 - z)\cdots(\beta_n - z)$ and m a positive integer. Then

$$Q^+(z) = \prod_{\substack{1\le j\le n,\\ Re(\beta_j)\ge m/2}} (\beta_j - z), \quad Q^-(z) = \prod_{\substack{1\le j\le n,\\ Re(\beta_j)<m/2}} (\beta_j - z), \tag{29}$$

with the understanding that Q^+ or Q^- takes the value 1, if one of the products is empty.

Note that $Q(z) = Q^+(z)Q^-(z)$, and the zeros of Q^+ are those zeros of Q for which $|\beta/(\beta - m)| \ge 1$.

Now, with the above definition, the following theorem of Peretz and Rassias [59] (also see [53, Theorem 1.4.26 on p. 202]) provides a correct version of Theorem 3.12.

Theorem 3.13. *Let* $P(z) = \sum_{j=0}^{m} a_j z^j$, $Q(z) = \sum_{j=0}^{n} b_j z^j$, *and* $R(z) = \sum_{j=0}^{m} a_j \times Q(j) z^j$. *If all the zeros of the polynomial* $P(z)$ *lie in the ring* $R_0 = \{z : 0 \le r_1 \le |z| \le r_2 \le \infty\}$, *then all the zeros of the polynomial* $R(z)$ *lie in the ring*

$$r_1\left|Q^-(0)/Q^-(m)\right| \le |z| \le r_2\left|Q^+(0)/Q^+(m)\right|. \tag{30}$$

4 Location of Zeros for Lacunary Polynomials, and Trinomials and Quadrinomials

4.1 Results Due to Dehmer Concerning Special Lacunary Polynomial

Dehmer and Mowshowitz [21] proved the following results for special classes of lacunary polynomials.

Theorem 4.1. *If the real polynomial*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1,\ a_0 > 0,\ \ n > 2, \tag{31}$$

has two positive zeros, then its largest positive zero δ *satisfies*

$$\delta < 1 + \sqrt{a_1}. \tag{32}$$

The following is an immediate consequence of the theorem above

Corollary 4.2. *If the real polynomial*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1,\ a_0 > 0,\ \ n > 2, \tag{33}$$

has two positive zeros, then its largest positive zero δ satisfies

$$\delta < 2. \tag{34}$$

Theorem 4.3. *If the real polynomial*

$$p(z) = z^n - a_1 z + a_0, \quad a_1,\ a_0 > 0,\ \ n > 2, \tag{35}$$

has two positive zeros, then its largest positive zero δ satisfies

$$\delta < \frac{1}{2} + \frac{\sqrt{4a_1 + 1}}{2}. \tag{36}$$

Theorem 4.4. *If the real polynomial*

$$p(z) = z^n - z + a_0, \quad a_0 > 0,\ \ n > 2, \tag{37}$$

has two positive zeros, then its largest positive zero δ satisfies

$$\delta < \frac{1}{2} + \frac{\sqrt{5}}{2}. \tag{38}$$

Based on the foregoing, we can determine the locations of all zeros of complex lacunary polynomials.

Theorem 4.5. *Let*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1 a_0 \neq 0,\ \ n > 2, \tag{39}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq \delta,$$

where $\delta > 1$ is the largest positive root of the equation

$$z^{n+1} - 2z^n - |a_1| z^2 + (|a_1| - |a_0|)z + |a_0| = 0. \tag{40}$$

Theorem 4.6. *Let*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1 a_0 \neq 0,\ \ n > 2, \tag{41}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq \delta,$$

where $\delta > 1$ is the largest positive root of the equation

$$z^{n+1} - 2z^n - M_4 z^2 + M_4 = 0, \tag{42}$$

where $M_4 := \max(|a_1|, |a_0|)$.

Applying a classical result of Cauchy [49], one can obtain the following explicit zeros bounds, which are given in Dehmer and Mowshowitz [21].

Theorem 4.7. *Let*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1 a_0 \neq 0, \ \ n > 2, \tag{43}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq 1 + |a_0| + |a_1|. \tag{44}$$

Theorem 4.8. *Let*

$$p(z) = z^n - z^{n-1} - a_1 z + a_0, \quad a_1 a_0 \neq 0, \ \ n > 2, \tag{45}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq \frac{1}{2} + \frac{\sqrt{1 + 4|a_1| + 4|a_0|}}{2}. \tag{46}$$

Theorem 4.9. *Let*

$$p(z) = z^n - a_1 z + a_0, \quad a_1 a_0 \neq 0, \ \ n > 2, \tag{47}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq \max(1, \delta),$$

where δ is the unique positive root of the equation

$$z^n - |a_1| z - |a_0| = 0. \tag{48}$$

Theorem 4.10. *Let*

$$p(z) = z^n - a_1 z + a_0, \quad a_1 a_0 \neq 0, \ \ n > 2, \tag{49}$$

be a complex polynomial with arbitrary coefficients. All zeros of $p(z)$ lie in

$$|z| \leq \max(1, \delta),$$

where δ is the unique positive root of the equation

$$z^n - M_4 z - M_4 = 0, \tag{50}$$

where $M_4 := \max(|a_1|, |a_0|)$.

Theorem 4.11. *Let*

$$p(z) = z^n - a_1 z + a_0, \quad a_1 a_0 \neq 0, \ \ n > 2, \tag{51}$$

be a complex polynomial. All zeros of $p(z)$ lie in

$$|z| \leq \frac{|a_1|}{2} + \frac{\sqrt{|a_1|^2 + 4|a_0| + 4}}{2}. \tag{52}$$

4.2 Location of Zeros of Trinomials and Quadrinomials

Quite a few results giving bound for all the zeros of a polynomial $p(z) = \sum_{j=0}^{n} a_j z^j$ were expressed (see [49, 61]) as functions of all the coefficients. It seems natural to ask whether there exist some bounds for the k zeros of smallest modulus, $k < n$, which would be independent of certain coefficients a_j. Laudau first raised this question in connection with his study of the Picard's Theorem. In [46, 47], Laudau proved that every trinomial

$$a_n z^n + a_1 z + a_0, \quad a_1 a_n \neq 0, \quad n \geq 2,$$

has at least one zero in

$$|z| \leq 2\left|\frac{a_0}{a_1}\right| \tag{53}$$

and every quadrinomial

$$a_n z^n + a_m z^m + a_1 z + a_0, \quad a_1 a_m a_n \neq 0, \quad 0 \leq m < n,$$

has at least one zero in

$$|z| \leq \frac{17}{3}\left|\frac{a_0}{a_1}\right|. \tag{54}$$

For every $n \geq 2$, as a refinement of (53) the trinomial

$$a_n z^n + a_1 z + a_0, \quad a_1 a_n \neq 0,$$

is well known [30] to have a zero in both the regions

$$\left|z+\frac{a_0}{a_1}\right| \leq \left|\frac{a_0}{a_1}\right| \quad \text{and} \quad \left|z+\frac{a_0}{a_1}\right| \geq \left|\frac{a_0}{a_1}\right|. \tag{55}$$

Joyal et al. [40] gave an alternative proof of this fact by using Gauss-Lucas theorem. In literature, there exist several results about zeros distribution of trinomials equations, for example see [4, 31]. In 2013, Aziz and Rather [6] proved certain results for quadrinomials and gave a simpler proof of (55), independent of Gauss-Lucas theorem. Here are their results

Theorem 4.12. *At least one zero of the quadrinomial*

$$a_n z^n + a_m z^m + a_1 z + a_0, \quad a_1 a_m a_n \neq 0, \quad 2 \leq m < n,$$

lie in

$$|z| \leq \frac{2n}{n-1}\left|\frac{a_0}{a_1}\right| \leq 3\left|\frac{a_0}{a_1}\right|. \tag{56}$$

Applying this result to the polynomial $z^n p(1/z)$ where $p(z) = a_0 + a_p z^p + a_{n-1}z^{n-1} + z^n$, they obtained the following:

Corollary 4.13. *At least one zero of the quadrinomial*

$$a_0 + a_p z^p + a_{n-1}z^{n-1} + z^n, \quad a_0 a_p a_{n-1} \neq 0, \quad 1 \leq p \leq n-2,$$

lie in

$$|z| \geq \frac{n-1}{2n}|a_{n-1}|. \tag{57}$$

Theorem 4.14. *For every $n \geq 3$, the quadrinomial*

$$a_n z^n + a_2 z^2 + a_1 z + a_0, \quad a_2 a_n \neq 0,$$

has at least one zero in both

$$|z| \leq \left[\frac{n}{n-2}\left|\frac{a_0}{a_2}\right|\right]^{1/2} \tag{58}$$

and

$$\left|z+\frac{a_1}{2a_2}\right| \geq \left|\frac{a_1}{2a_2}\right|. \tag{59}$$

5 Some Recent Results Concerning Cauchy Theorem on the Location of Zeros of a Polynomial

5.1 Disk Containing All the Zeros of a Polynomial

In recent times there have been many improvements and generalizations of the Cauchy Theorem. We start by stating the following result due to Jain [38].

Theorem 5.1. *All the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n \neq 0$, *lie in the disk*

$$|z| \leq \frac{\max\left(\frac{|a_{n-1}|}{|a_n|}, 2\frac{|a_{n-2}|}{|a_{n-1}|}, 3\frac{|a_{n-3}|}{|a_{n-2}|}, \ldots, n\frac{|a_0|}{|a_1|}\right)}{\ln 2}. \tag{60}$$

Further recent results concerning upper bounds have been obtained by Kalantari [41]. He has found a family of zeros bounds for analytic functions that has been proven powerful when comparing the resulting bounds with classical ones by using complex polynomials. In this regard, he proved the following results

Theorem 5.2. *Let* $m \geq 2$ *and let* $r_m \in [1/2, 1)$ *be the positive root of the polynomial*

$$q(t) := t^{m-1} + t - 1.$$

For $m = 2$ *and* $r_2 = 1/2$, *all the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *lie in the disk*

$$|z| \leq 2 \max_{1 \leq j \leq n} \left(\left| \frac{a_{n-j}}{a_n} \right| \right)^{1/j}. \tag{61}$$

Theorem 5.3. *Let* $m \geq 2$ *and let* $r_m \in [1/2, 1)$ *be the positive root of the polynomial*

$$q(t) := t^{m-1} + t - 1.$$

For $m = 3$ *and* $r_3 = \dfrac{2}{\sqrt{5}+1}$, *all the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *lie in the disk*

$$|z| \leq \frac{\sqrt{5}+1}{2} \max_{2 \leq j \leq n+1} \left(\left| \frac{a_{n-1}a_{n-j+1} - a_n a_{n-j}}{a_n^2} \right| \right)^{1/j}, \quad a_{-1} = 0. \tag{62}$$

Dehmer [20, 21] proved the following implicit zero bound results

Theorem 5.4. *All the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *lie in the disk*

$$|z| \leq \max(1, \delta_2), \tag{63}$$

where δ_2 *(besides* $\delta_1 = 1$*) denotes the positive root of the equation*

$$z^{n+1} - (1 + M_2)z^n + M_2 = 0,$$

and $M_2 := \max\limits_{0 \leq j \leq n-1} \left|\frac{a_j}{a_n}\right|$. *The bound is sharp for all polynomials of the form* $p(z) = az^n - b[z^{n-1} + \cdots + z + 1]$, $a, b > 0$.

Theorem 5.5. *All the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *lie in the disk*

$$|z| \leq \max(1, \delta_2), \tag{64}$$

where δ_2 *(besides* $\delta_1 = 1$*) denotes the positive root of the equation*

$$z^{n+1} - \left(1 + \left|\frac{a_{n-1}}{a_n}\right|\right)z^n + \left(\left|\frac{a_{n-1}}{a_n}\right| - M_1\right)z^{n-1} + M_1 = 0,$$

and $M_1 := \max\limits_{0 \leq j \leq n-2} \left|\frac{a_j}{a_n}\right|$. *The bound is sharp for all polynomials of the form* $p(z) = az^n - bz^{n-1} - c[z^{n-2} + \cdots + z + 1]$, $a, b > 0$, $c \geq 0$.

Theorem 5.6. *Let*

$$M_3 := \max_{2 \leq j \leq n} \left|\frac{a_{n-1}a_{n-j} - a_{n-1}a_{n-j-1}}{a_n^2}\right|, \quad a_{-1} = 0,$$

and

$$\phi_1 = \frac{|a_{n-1}^2 - a_n a_{n-2}|}{|a_n|^2}.$$

In addition, let $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *be a complex polynomial. Then all the zeros of* $p(z)$ *lie in the closed disk*

$$|z| \leq \delta,$$

where $\delta > 1$ *is the largest positive root of the equation*

$$z^3 - z^2 - (M_3 + \phi_1)z + \phi_1 = 0.$$

Moreover,

$$1 < \delta < 1 + \sqrt{M_3 + \phi_1}.$$

In 2011, Dehmer and Mowshowitz [21] proved the following explicit bounds.

Theorem 5.7. *All the zeros of the polynomial* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$, *lie in the disk*

$$|z| \leq \frac{1 + \phi_2}{2} + \frac{\sqrt{(\phi_2 - 1)^2 + 4M_1}}{2}, \tag{65}$$

where $\phi_2 := \left|\frac{a_{n-1}}{a_n}\right|$ *and* $M_1 := \max_{0 \leq j \leq n-2} \left|\frac{a_j}{a_n}\right|$.

The next Theorem gives a bound for polynomials with restrictions on the coefficients. Dehmer [20] has shown that such bounds can be more precise and often lead to better results when locating the zeros of polynomials.

Theorem 5.8. *Let* $p(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots + a_n z^n$, $a_n a_{n-1} \neq 0$. *Suppose that* $|a_j| < 1$, $0 \leq j \leq n-2$. *All zeros of* $p(z)$ *lie in the disk*

$$|z| \leq \frac{1 + \phi_2}{2} + \frac{\sqrt{(\phi_2 - 1)^2 + \frac{4}{|a_n|}}}{2}, \tag{66}$$

where $\phi_2 := \left|\frac{a_{n-1}}{a_n}\right|$. *The bound is sharp for all polynomials of the form* $p(z) = az^n - bz^{n-1} - [z^{n-2} + \cdots + z + 1]$, $a, b > 0$.

Dehmer and Mowshowitz [21] also considered examples to illustrate the quality of bounds given by Theorems 5.4–5.8, and here we present just one of them.

Example 5.1. Consider the following polynomials

$$p_1(z) = -z^5 + 5z^4 - 0.2230z^3 + 9.548 \times 10^{-5} z^2 + 7.851 \times 10^{-5} z - 6.515 \times 10^{-7},$$
$$p_2(z) = z^5 - 4z^3 + 3z.$$

For $p_1(z)$, the conditions $|a_j| \leq 1$, $0 \leq j \leq n-2$ are fulfilled to apply the bound given by Theorem 5.8. The zeros of $p_1(z)$ and $p_2(z)$ are

$$z_1 = 4.9550,\ z_2 = 0.0187,\ z_3 = 0.0339,\ z_4 = 0.0110,\ z_5 = -0.0187$$

and

$$z_1 = 0,\ z_2 = 1.7321,\ z_3 = 1,\ z_4 = -1.7321,\ z_5 = -1,$$

respectively. We see that these polynomials only possess real zeros. Now consider the polynomials

$$p_3(z) = z^3 + 5z^2 - 15z + 1,\ p_4(z) = z^3 + 4z^2 + 1000z + 99.$$

The zeros of $p_3(z)$ and $p_4(z)$ are

$$z_1 = 2.0567,\ z_2 = -7.1249,\ z_3 = 0.0682$$

and

$$z_1 = -0.0990,\ z_2 = -1.950 - 31.556i,\ z_3 = -1.950 + 31.556i,$$

respectively. To gain additional insight into the determination of bounds, we examine two polynomials whose coefficients have, respectively, small and large moduli. The polynomial $p_4(z)$ has two complex zeros, so let $\sigma := \max_{1\le j\le n} |z_j|$ for a given polynomial. Let A, B, C, D, and E represent the bound obtained by Theorems 5.4–5.8, respectively. The values of σ, A, B, C, D, and E for all polynomials defined above can be found in Table 1 below. Notice that Theorem 5.8 does not apply to $p_2,\ p_3,\ p_4$ since the special conditions for the coefficients are not satisfied.

We close this part by giving the following results due to Zeheb [73], and Žilović et al. [74]

Theorem 5.9. *All the zeros of the real polynomial* $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ *lie in the circle* $|z| < 1 + \max\{A_{ij}\}$, *where*

$$A_{ij} = \frac{|a_i a_{j-1} - a_j a_{i-1}|}{|a_i| + |a_j|}, \qquad i, j = 0, \cdots, n; j > i$$
$$a_n \equiv 1, \qquad a_{-1} \equiv 0.$$

Table 1 Evaluation of zero bounds for $p_1(z),\ p_2(z),\ p_3(z)$, and $p_4(z)$

	$p_1(z)$	$p_2(z)$	$p_3(z)$	$p_4(z)$
A	5.9994	4.9987	15.9963	1000.9999
B	5.0549	2.7728	7.3267	34.1447
C	5.1581	2.9254	11.1137	70.8927
D	5.0550	2.5615	7.3588	34.1583
E	5.2361	–	–	–
σ	4.955	1.7321	7.1249	31.6160

Theorem 5.10. *All the zeros of the complex polynomial* $p(z) = z^n + \sum_{k=0}^{n-1} a_k z^k$ *lie in the disk*

$$\{z \in \mathbb{C} : |z| < \sqrt{1 + A}\},$$

where $A = \max_{0 \le k \le n-1} \left\{ |a_k^2 + 2(-1)^k (B - C)| \right\}$, $B = \sum_{\substack{0 \le i < j \le [n/2] \\ i+j=k}} a_{2i} a_{2j}$,

and $C = \sum_{\substack{0 \le i < j \le [(n-1)/2] \\ i+j=k-1}} a_{2i+1} a_{2j+1}$.

Here $a_n = 1$ *and as usual* $[k]$ *denotes the integer part of* k.

5.2 Annuli Containing All the Zeros of a Polynomial

We begin by presenting the result of Datt and Govil [19] (see also Dewan [23]).

Theorem 5.11. *Let* $p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$, *be a polynomial of degree* n *and*

$$A = \max_{0 \le j \le n-1} |a_j|.$$

Then $p(z)$ *has all its zeros in the ring shaped region*

$$\frac{|a_0|}{2\,(1 + A)^{n-1}\,(An + 1)} \le |z| \le 1 + \lambda_0 A, \tag{67}$$

where λ_0 *is the unique positive root of the equation* $x = 1 - 1/\,(1 + Ax)^n$ *in the interval* $(0, 1)$. *The upper bound* $1 + \lambda_0 A$ *in the above given region (68) is best possible and is attained for the polynomial* $p(z) = z^n - A\left(z^{n-1} + \cdots + z + 1\right)$.

In case we do not wish to solve the equation $x = 1 - 1/\,(1 + Ax)^n$, then in order to apply the above result of Datt and Govil [19], we can apply the following result also due to Datt and Govil [19], which in every case clearly gives an improvement over Theorem 2.2 of Cauchy [14].

Theorem 5.12. *Let* $p(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$, *be a polynomial of degree* n *and*

$$A = \max_{0 \le j \le n-1} |a_j|.$$

Then $p(z)$ *has all its zeros in the ring shaped region*

$$\frac{|a_0|}{2\,(1+A)^{n-1}\,(An+1)} \leq |z| \leq 1 + \left(1 - \frac{1}{(1+A)^n}\right)A. \tag{68}$$

Since, always $\left(1 - \frac{1}{(1+A)^n}\right) < 1$, the above Theorem 5.12 in every situation sharpens Theorem 2.2 due to Cauchy.

Although, since the beginning, binomial coefficients defined by $C_k^n = \frac{n!}{k!(n-k)!}, 0! = 1$ have appeared in the derivation or as a part of closed expressions of bounds, the Fibonacci's numbers defined by $F_0 = 0$, $F_1 = 1$, and for $j \geq 2$, $F_j = F_{j-1} + F_{j-2}$ have not appeared either in implicit bounds or explicit bounds for the moduli of the zeros. Diaz-Barrero [25] proved the following result, which gives circular domains containing all the zeros of a polynomial where binomial coefficients and Fibonacci's numbers appear. He also gives an example of a polynomial for which the above theorem gives a better bound than the bound obtainable from Theorem 2.2 of Cauchy [14].

Theorem 5.13. *Let $p(z) = \sum_{j=0}^{n} a_j z^j$ $(a_j \neq 0,\ 0 \leq j \leq n)$ be a complex monic polynomial. Then all its zeros lie in the disk $C_1 = \{z \in \mathbb{C} : |z| \leq r_1\}$ or $C_2 = \{z \in \mathbb{C} : |z| \leq r_2\}$, where*

$$r_1 = \max_{1\leq k\leq n} \left\{ \sqrt[k]{\frac{2^{n-1}C_2^{n+1}}{k^2 C_k^n}|a_{n-k}|} \right\},$$

$$r_2 = \max_{1\leq k\leq n} \left\{ \sqrt[k]{\frac{F_{3n}}{C_k^n 2^k F_k}|a_{n-k}|} \right\}.$$

The proof of the above theorem depends on the identities

$$\sum_{k=1}^{n} k^2 C_k^n = 2^{n-2} n(n+1) \tag{69}$$

and

$$\sum_{k=1}^{n} C_k^n 2^k F_k = F_{3n}, \tag{70}$$

where F_j are the Fibonacci's numbers, and C_k^n the binomial coefficients.

The following result, which provides an annulus region containing all the zeros of a polynomial is also due to Diaz-Barrero [24].

Theorem 5.14. *Let $p(z) = \sum_{j=0}^{n} a_j z^j$ $(a_j \neq 0,\ \ 0 \leq j \leq n)$ be a nonconstant complex polynomial. Then all its zeros lie in the annulus $C = \{z \in \mathbb{C} : r_1 \leq |z| \leq r_2\}$, where*

$$r_1 = \frac{3}{2} \min_{1\le j\le n} \left\{ \frac{2^n F_j C_j^n}{F_{4n}} \left| \frac{a_0}{a_j} \right| \right\}^{1/j},$$

$$r_2 = \frac{2}{3} \max_{1\le j\le n} \left\{ \frac{F_{4n}}{2^n F_j C_j^n} \left| \frac{a_{n-j}}{a_n} \right| \right\}^{1/j}.$$

Here F_j being the Fibonacci's numbers, and C_j^m the binomial coefficients.

The following result of Kim [43], whose proof depends on the use of the identity

$$\sum_{k=0}^{n} C_k^n = 2^n - 1 \tag{71}$$

also provides an annulus containing all the zeros of a polynomial.

Theorem 5.15. *Let $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, \quad 0 \le k \le n)$ be a nonconstant polynomial with complex coefficients. Then all the zeros of $p(z)$ lie in the annulus $A = \{z : r_1 \le |z| \le r_2\}$, where*

$$r_1 = \min_{1\le k\le n} \left\{ \frac{C_k^n}{2^n - 1} \left| \frac{a_0}{a_k} \right| \right\}^{1/k}, \quad r_2 = \max_{1\le k\le n} \left\{ \frac{2^n - 1}{C_k^n} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{72}$$

Here again, as usual, C_k^n denote the binomial coefficients.

Theorem 2.2 of Cauchy has also been refined by Sun and Hsieh [66], who proved

Theorem 5.16. *All the zeros of the complex polynomial*

$$p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$$

lie in the disk

$$\{z : |z| < \eta\} \subset \{z : |z| < 1 + \delta_3\} \subseteq \{z : |z| < 1 + A\},$$

where δ_3 is the unique positive root of the equation,

$$Q_3(x) \equiv x^3 + (2 - |a_{n-1}|)x^2 + (1 - |a_{n-1}| - |a_{n-2}|)x - A = 0, \tag{73}$$

and

$$A = \max_{0\le j\le n-1} |a_j|.$$

Using the method similar to that of Sun and Hsieh [66], Jain [38] refined the above result of Sun and Hsieh [66], and proved.

Theorem 5.17. *All the zeros of the complex polynomial*

$$p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$$

lie in the disk

$$\{z : |z| < \eta\} \subset \{z : |z| < 1 + \delta_4\} \subseteq \{z : |z| < 1 + \delta_3\} \subseteq \{z : |z| < 1 + A\},$$

where δ_4 is the unique positive root of the equation,

$$\begin{aligned} Q_4(x) \equiv x^4 &+ (3 - |a_{n-1}|)\, x^3 + (3 - 2|a_{n-1}| - |a_{n-2}|)\, x^2 \\ &+ (1 - |a_{n-1}| - |a_{n-2}| - |a_{n-3}|)\, x - A = 0, \end{aligned} \tag{74}$$

and $A = \max\limits_{0 \le j \le n-1} |a_j|$, is same as in Theorem 5.16.

In 2009, Affane-Aji, Agarwal, and Govil [2] proved the following result which not only includes the above results of Cauchy [14], Sun and Hsieh [66], and Jain [38] as special cases but also provides a tool for obtaining sharper bounds for the location of the zeros of a polynomial.

Theorem 5.18. *All the zeros of the polynomial*

$$p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$$

lie in the disks

$$\begin{aligned} \{z : |z| < 1 + \delta_k\} &\subseteq \{z : |z| < 1 + \delta_{k-1}\} \cdots \\ &\subseteq \{z : |z| < 1 + \delta_1\} \subseteq \{z : |z| < 1 + A\}, \end{aligned}$$

where δ_k is the unique positive root of the kth degree equation

$$Q_k(x) \equiv x^k + \sum_{\nu=2}^{k} \left[C_{k-\nu}^{k-1} - \sum_{j=1}^{\nu-1} C_{k-\nu}^{k-j-1} |a_{n-j}| \right] x^{k+1-\nu} - A = 0. \tag{75}$$

Here

$$A = \max_{0 \le j \le n-1} |a_j|, \quad a_j = 0 \text{ if } j < 0,$$

and for k, a positive integer, C_k^m are the binomial coefficients.

As is easy to verify, for $k = 1$ the above theorem reduces to Theorem 2.2 due to Cauchy [14], for $k = 3$ to the result of Sun and Hsieh [66], and for $k = 4$ it reduces to the result due to Jain [38]. Further, by choosing k sufficiently large we can make $1 + \delta_k$ in the bound close to the actual bound.

Note that by combining the above Theorem 5.18 with Theorem 5.12 of Datt and Govil [19] one can easily obtain the following result, which is a refinement of the above Theorem 5.18.

Theorem 5.19. *All the zeros of the polynomial*

$$p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$$

lie in the annulus

$$\frac{|a_0|}{2(1+A)^{n-1}(nA+1)} \leq |z| \leq \{z : |z| < 1 + \delta_k\} \subseteq \{z : |z| < 1 + \delta_{k-1}\} \cdots$$
$$\subseteq \{z : |z| < 1 + \delta_1\} \subseteq \{z : |z| < 1 + A\},$$

where δ_k is as defined in Theorem 5.18, and $A = \max\limits_{0 \leq j \leq n-1} |a_j|$.

Similarly, one can obtain a refinement of Theorem 5.18 by combining Theorem 5.18 with Theorem 5.14 of Diaz-Barrero [24].

Later in 2010, Affane-Aji, Biaz, and Govil [3] proved the following refinement of Theorem 5.18, and constructed examples to show that for some polynomials their theorem, stated below, gives much better bounds than obtainable from Theorem 5.19. More precisely, their result is

Theorem 5.20. *All the zeros of the polynomial*

$$p(z) = z^n + \sum_{j=0}^{n-1} a_j z^j$$

lie in the disks

$$R_1 \leq |z| \leq \{z : |z| < 1 + \delta_k\} \subseteq \{z : |z| < 1 + \delta_{k-1}\} \cdots$$
$$\subseteq \{z : |z| < 1 + \delta_1\} \subseteq \{z : |z| < 1 + A\},$$

where δ_k is as defined in Theorem 5.18, and

$$R_1 = \frac{-R^2|a_1|(M - |a_0|) + \sqrt{4R^2M^3|a_0| + \{R^2|a_1|(M - |a_0|)\}^2}}{2M^2}. \tag{76}$$

Here $M = \frac{R^{n+1} + (A-1)R^n - AR}{(R-1)}$ with $R = 1 + \delta_k$ and $A = \max\limits_{0 \leq j \leq n-1} |a_j|$.

Note that $R = 1 + \delta_k > 1$, so for every positive integer k, we have $M > 0$ and $R > 0$. It is obvious that, in general, Theorem 5.20 sharpens Theorem 5.18.

In the same paper Affane-Aji, Biaz, and Govil [3] prove some more refinements of Theorem 5.18, which in some cases gives bounds that are sharper than obtainable from Theorems 5.12, 5.14, and 5.19. This they have shown by constructing some examples of polynomials.

The following two results by Diaz-Barrero and Egozcue [26] also provide annuli containing all the zeros of a polynomial.

Theorem 5.21. *Let $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, 1 \leq k \leq n)$ be a non-constant complex polynomial. Then for $j \geq 2$, all its zeros lie in the annulus $C = \{z : r_1 \leq |z| \leq r_2\}$ where*

$$r_1 = \min_{1\leq k\leq n} \left\{ \frac{C(n,k)A_k B_j^k (bB_{j-1})^{n-k}}{A_{jn}} \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{77}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{A_{jn}}{C(n,k)A_k B_j^k (bB_{j-1})^{n-k} A_{jn}} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{78}$$

Here, $B_n = \sum_{k=0}^{n-1} r^k s^{n-1-k}$ and $A_n = cr^n + ds^n$, where c, d are real constants and r,s are the roots of the equation $x^2 - ax - b = 0$ in which a,b are strictly positive real numbers. For $j \geq 2$, we have $A_{jn} = \sum_{k=0}^{n} C(n,k)(bB_{j-1})^{n-k} B_j^k A_k$. Furthermore, $C(n,k)$ is the binomial coefficient.

Theorem 5.22. *Let $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, 1 \leq k \leq n)$ be a non-constant polynomial with complex coefficients. Then, all its zeros lie in the ring shaped region $C = \{z : r_1 \leq |z| \leq r_2\}$ where*

$$r_1 = \min_{1\leq k\leq n} \left\{ \frac{2^n P_k C(n,k)}{P_{3n}} \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{79}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{P_{3n}}{2^n P_k C(n,k)} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{80}$$

Here P_k is the kth Pell number, namely, $P_0 = 0$, $P_1 = 1$ and for $k \geq 2$, $P_k = 2P_{k-1} + P_{k-2}$. Furthermore, $C(n,k) = \frac{n!}{k!(n-k)!}$ are the binomial coefficients.

Recently, Dalal and Govil [16] unified the above results by proving the following.

Theorem 5.23. *Let $A_k > 0$ for $1 \leq k \leq n$, and be such that $\sum_{k=1}^{n} A_k = 1$. If $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0,\ 1 \leq k \leq n)$ is a non-constant polynomial with complex*

coefficients, then all the zeros of $p(z)$ *lie in the annulus* $C = \{z : r_1 \leq |z| \leq r_2\}$, *where*

$$r_1 = \min_{1\leq k\leq n} \left\{ A_k \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{81}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{1}{A_k} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{82}$$

The above theorem, by appropriate choice of the numbers $A_k > 0$ for $1 \leq k \leq n$, includes as special case Theorems 5.13–5.15, 5.21 and 5.22, and this has been shown in Table 1 in the paper of Dalal and Govil [16, p. 9612].

Also Dalal and Govil [16, p. 9612] show that their theorem is capable of generating infinite number of results. In particular, as corollaries they obtained

Corollary 5.24. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, 1 \leq k \leq n)$ *be a non-constant polynomial with complex coefficients. Then all the zeros of* $p(z)$ *lie in the annulus* $C = \{z : r_1 \leq |z| \leq r_2\}$, *where*

$$r_1 = \min_{1\leq k\leq n} \left\{ \frac{L_k}{L_{n+2} - 3} \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{83}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{L_{n+2} - 3}{L_k} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{84}$$

Here, L_k *is the kth Lucas number defined by* $L_0 = 2,\ L_1 = 1$, *and* $L_{n+2} = L_n + L_{n+1}$, *if* $n \geq 0$.

Corollary 5.25. *Let* $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, 1 \leq k \leq n)$ *be a non-constant polynomial with complex coefficients. Then all the zeros of* $p(z)$ *lie in the annulus* $C = \{z : r_1 \leq |z| \leq r_2\}$, *where*

$$r_1 = \min_{1\leq k\leq n} \left\{ \frac{C_{k-1} C_{n-k}}{C_n} \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{85}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{C_n}{C_{k-1} C_{n-k}} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}, \tag{86}$$

where C_k *is the kth Catalan number.*

Also, Dalal and Govil [16] developed MATLAB code, and use this to construct some examples of polynomials for which the annuli containing all the zeros of the polynomials obtainable by their Corollaries 5.24 and 5.25 are considerably sharper than the annuli obtainable from the known results, Theorems 5.14 and 5.15. Due to limitation in space here we mention only two of their examples.

Example 5.2. Let $p(z) = z^4 + 0.01z^3 + 0.1z^2 + 0.2z + 0.4$

Result	r_1	r_2	Area of annulus
Theorem 5.14	0.1945	1.1289	3.8835
Theorem 5.15	0.4041	1.5650	7.1786
Corollary 5.24	0.1333	0.9621	2.8512
Corollary 5.25	0.6147	1.1186	2.7427
Actual bound	0.7190	0.8801	0.8093

Although both the Corollaries 5.24 and 5.25 give bounds better than those obtainable from Theorems 5.14 and 5.15 but, as is evident from the above table, Corollary 5.25 gives the best bounds in terms of area, with over 29 % improvement in the area obtainable by Theorem 5.14, and with over 61 % improvement in the area obtainable by Theorem 5.15. In fact, if one combines the results of Corollaries 5.24 and 5.25 one obtains that all the zeros lie in the annulus $0.6147 \leq |z| \leq 0.9621$, and the annulus obtained this way is quite close to the actual annulus $0.7190 \leq |z| \leq 0.8801$, in terms of radii r_1 and r_2. Note that the area in this case comes out to be 1.7209 which is an improvement of about 56 % over the area obtainable by Theorem 5.14, and improvement of about 76 % over the area obtainable by Theorem 5.15.

Example 5.3. Let $p(z) = z^5 + 0.006z^4 + 0.01z^3 + 0.2z^2 + 0.3z + 1$

Result	r_1	r_2	Area of annulus
Theorem 5.14	0.1182	1.4097	6.1964
Theorem 5.15	0.5031	1.9873	11.6064
Corollary 5.24	0.1282	1.1877	4.3779
Corollary 5.25	0.7715	1.2805	3.2801
Actual bound	0.9526	1.0607	0.6873

Again, here also, although both the corollaries give bounds better than those obtained from Theorems 5.14 and 5.15, but in particular Corollary 5.25 gives the best bounds in terms of area. As can be seen from the above table, in this case there is an improvement of about 47 % in the area obtained by Theorem 5.14, and of about 71 % in the area obtained by Theorem 5.15. Again, if one combines the results of Corollaries 5.24 and 5.25 one gets that all the zeros lie in the annulus $0.7715 \leq |z| \leq 1.1877$. The area in this case is about 2.5617 which is an improvement of about 59 % over the area obtained by Theorem 5.14, and improvement of about 78 % over the area obtained from Theorem 5.15.

Note that Theorem 5.23 of Dalal and Govil [16] implies that infinitely many annuli containing all the zeros of a complex polynomial $P(z) = \sum_{k=0}^{n} a_k z^k$ ($a_k \neq 0$ for $1 \leq k \leq n$) can be obtained from infinitely many sequences of positive numbers, $\{A_k\}_{k=1}^{n}$, such that $\sum_{k=1}^{n} A_k = 1$. Then, it is natural to ask which sequence of positive numbers, $\{A_k\}_{k=1}^{n}$, with $\sum_{k=1}^{n} A_k = 1$ gives the best result. Over the years, mathematicians have compared their bounds with the existing bounds in the literature by generating some examples, and thus showing that in some special cases their bound gives better results than some known results.

In the following two theorems, Dalal and Govil [17] show that no-matter what result one obtains as a corollary of Theorem 5.23, one can always generate examples in which the bound obtained by this obtained corollary is better than the existing ones.

Theorem 5.26. *Let $\{A_k\}_{k=1}^{n}$ and $\{B_k\}_{k=1}^{n}$ be sequences of positive numbers such that $\sum_{k=1}^{n} A_k = 1$ and $\sum_{k=1}^{n} B_k = 1$. Then, there always exists a polynomial for which $r_1^A > r_1^B$ and vice versa, where r_1^A and r_1^B are the inner radii of the annulus obtained from the Theorem 5.23 by using the sequences $\{A_k\}_{k=1}^{n}$ and $\{B_k\}_{k=1}^{n}$, respectively.*

Theorem 5.27. *Let $\{A_k\}_{k=1}^{n}$ and $\{B_k\}_{k=1}^{n}$ be sequences of positive numbers such that $\sum_{k=1}^{n} A_k = 1$ and $\sum_{k=1}^{n} B_k = 1$. Then, there always exists a polynomial for which $r_2^A > r_2^B$ and vice versa, where r_2^A and r_2^B are the outer radii of the annulus obtained from the Theorem 5.23 by using sequences $\{A_k\}_{k=1}^{n}$ and $\{B_k\}_{k=1}^{n}$, respectively.*

In the same paper, Dalal and Govil [17] also proved the following result which always improves any bound r_1, r_2 obtainable from any of the corollaries of Theorem 5.23, if $D_{r_1} > 0$ and $D_{r_2} > 0$.

Theorem 5.28. *Suppose $\sum_{k=1}^{n} B_k = 1$, with $B_k > 0$ for $1 \leq k \leq n$, and $P(z) = \sum_{k=0}^{n} a_k z^k$ ($a_k \neq 0, 1 \leq k \leq n$) be a non-constant polynomial with complex coefficients. Let r_1, r_2 be any positive numbers such that $D_{r_1} \geq 0$ and $D_{r_2} \geq 0$. Then, all the zeros of $P(z)$ lie in the annulus $C = \{z : r_1' \leq |z| \leq r_2'\}$ where*

$$r_1' = \min_{1 \leq k \leq n} \left\{ r_1^k + D_{r_1} B_k \left| \frac{a_0}{a_k} \right| \right\}^{1/k}, \tag{87}$$

and

$$r_2' = \max_{1 \leq k \leq n} \left\{ \frac{1}{r_2^k} + D_{r_2} B_k \left| \frac{a_n}{a_{n-k}} \right| \right\}^{-1/k}. \tag{88}$$

Finally, Dalal and Govil [17] by using MATLAB constructed a polynomial to compare the results obtained by Theorem 5.28 with the existing results in the literature, and subsequently showed the importance of such a theorem. We present here the following example used by Dalal and Govil [17].

Example 5.4. Let $p(z) = z^3 + 0.1z^2 + 0.1z + 0.7$.
In the following table, they used Theorem 5.14, taking $A_k = \frac{C(n,k)2^k F_k}{F_{3n}}$, to get inner radii r_1 and outer radii r_2, and use Theorem 5.28 with $B_k = 1/3$ for $1 \leq k \leq 3$ to compute the corresponding r_1' and r_2'.

Result	Inner radii (r_1 or r_1')	Outer radii (r_2 or r_2')	Area of annulus
Theorem 5.14	0.6402	1.2312	3.4730
Theorem 5.28	0.6576	1.094	2.4027
Actual bound	0.8840	0.8899	0.0328

In the table below, they used Theorem 5.13, taking $A_k = \frac{k^2 C(n,k)}{2^{n-1}C(n+1,2)}$, to get the inner radii r_1 and outer radii r_2, and use Theorem 5.28 with $B_k = 1/3$ for $1 \leq k \leq 3$ to compute the corresponding r_1' and r_2'.

Result	Inner radii (r_1 or r_1')	Outer radii (r_2 or r_2')	Area of annulus
Theorem 5.13	0.4641	1.6984	8.382
Theorem 5.28	0.5002	1.2076	3.7957
Actual bound	0.8840	0.8899	0.0328

It is clear from the above tables that Theorem 5.28 gives the best bounds in terms of inner and outer radii of the annulus, and as well in terms of area of the annulus containing all the zeros of the polynomial.

In 2011, Bidkham and Shashahani [8] made use of t-Fibonacci numbers, defined by $F_{t,n} = tF_{t,n-1} + F_{t,n-2}$, for $n \geq 2$, with $F_{t,0} = 0, F_{t,1} = 1$, where t is any positive real number, and obtained the following result that gives annulus in terms of t-Fibonacci numbers, containing all the zeros of a polynomial.

Theorem 5.29. *Let $p(z) = \sum_{k=0}^{n} a_k z^k$ $(a_k \neq 0, 0 \leq k \leq n)$ be a non-constant complex polynomial of degree n. Then all the zeros of $p(z)$ lie in the annulus $\mathcal{C} = \{z : r_1 \leq |z| \leq r_2\}$, where*

$$r_1 = \min_{1\leq k\leq n} \left\{ \frac{(t^3+2t)^k (t^2+1)^n F_{t,k} \binom{n}{k}}{(t^2+1)^k F_{t,4n}} \left| \frac{a_0}{a_k} \right| \right\}^{1/k} \tag{89}$$

and

$$r_2 = \max_{1\leq k\leq n} \left\{ \frac{(t^2+1)^k F_{t,4n}}{(t^3+2t)^k (t^2+1)^n F_{t,k} \binom{n}{k}} \left| \frac{a_{n-k}}{a_n} \right| \right\}^{1/k}. \tag{90}$$

For $t = 1$, the above theorem reduces to Theorem 5.14 due to Diaz-Barrero [24].

Later in 2013, Rather and Mattoo [62] considered generalized Fibonacci numbers [71], and proved a result, which generalizes the above Theorem 5.29 due to Bidkham and Shashahani [8].

Recently Dalal and Govil [18] further generalized the concept of generalized Fibonacci numbers and obtained result that generalizes result of Rather and

Mattoo [62], and therefore Theorem 5.29 of Bidkham and Shashahani [8], and Theorem 5.14 due to Diaz-Barrero [24].

We conclude this paper by adding that recently by using Theorem 5.23 of Dalal and Govil [16], Govil and Kumar [35] have obtained several results providing annuli containing all the zeros of a polynomial. Their bounds are in terms of Narayana numbers [56], Motzkin numbers (see [28]), and special combination of binomial coefficients. Also, by using MATLAB they construct examples to show that in special cases their results give sharper bounds than obtainable from some of the known results.

References

1. Aberth, O.: Iteration methods for finding all zeros of a polynomial simultaneously. Math. Comput. **27**, 339–344 (1973)
2. Affane-Aji, C., Agarwal, N., Govil, N.K.: Location of zeros of polynomials. Math. Comput. Model. **50**, 306–313 (2009)
3. Affane-Aji, C., Biaz, S., Govil, N.K.: On annuli containing all the zeros of a polynomial. Math. Comput. Model. **52**, 1532–1537 (2010)
4. Ahn, Y.J., Kim, S-H.: Zeros of certain trinomials equations. Math. Inequal. Appl. **9**, 225–232 (2006)
5. Anai, H., Horimoto, K.: Algebraic biology 2005. In: Proceedings of the 1st International Conference on Algebraic on Algebraic Biology, Tokyo, Japan (2005)
6. Aziz, A., Rather, N.A.: Location of zeros of trinomials and quadrinomials. Math. Inequal. Appl. **17**, 823–829 (2014)
7. Berwald, L.: Elementare Sätze uber die Abgrenzung der Wursln einer algebraischen Gleichung. Acta. Sci. Math. Litt. Sci. Szeged **6**, 209–221 (1934)
8. Bidkham, M., Shashahani, E.: An annulus for the zeros of polynomials. Appl. Math. Lett. **24**, 122–125 (2011)
9. Biernacki, M.: Sur les équations algébriques contenant des paramètres arbitraires. Bull. Acad. Polon. Sci. Sér. A, **III**, 541–685 (1927)
10. Birkhoff, G.D.: An elementary double inequality for the roots of an algebraic equation having greatest value. Bull. Am. Math. Soc. **21**, 494–495 (1914)
11. Bissel, C.: Control Engineering, 2nd edn. CRC Press, Boca Raton (2009)
12. Borwein, P., Erdelyi, T.: Polynomials and Polynomial Inequalities. Springer, Berlin (1995)
13. Carmichael, R.D., Mason, T.E.: Note on the roots of algebraic equations. Bull. Am. Math. Soc. **21**, 14–22 (1914)
14. Cauchy, A.L.: Excercises de Mathematiques. IV Annee de Bure Freres, Paris (1829)
15. Cohn, A.: Über die Anzahl der Wurzeln einer algebraischen Gleichung in einen Kreise. Math. Z. **14**, 110–148 (1922)
16. Dalal, A., Govil, N.K.: On region containing all the zeros of a polynomial. Appl. Math. Comput. **219**, 9609–9614 (2013)
17. Dalal, A., Govil, N.K.: Annulus containing all the zeros of a polynomial. Appl. Math. Comput. **249**, 429–435 (2014)
18. Dalal, A., Govil, N.K.: Generalization of some results on the annulus containing all the zeros of a polynomial (preprint)
19. Datt, B., Govil, N.K.: On the location of zeros of polynomials. J. Approx. Theory **24**, 78–82 (1978)
20. Dehmer, M.: On the location of zeros of complex polynomials. J. Inequal. Pure Appl. Math. **7**(1), 1–27 (2006)

21. Dehmer, M., Mowshowitz, A.: Bounds on the moduli of polynomial zeros. Appl. Math. Comput. **218**, 4128–4137 (2011)
22. Dehmer, M., Tsoy, Y.R.: The quality of zero bounds for complex polynomials. PLoS ONE **7**(7) (2012). Doi:10.1371/journal.pone.0039537
23. Dewan, K.K.: On the location of zeros of polynomials. Math. Stud. **50**, 170–175 (1982)
24. Diaz-Barrero, J.L.: An annulus for the zeros of polynomials. J. Math. Anal. Appl. **273**, 349–352 (2002)
25. Diaz-Barrero, J.L.: Note on bounds of the zeros. Mo. J. Math. Sci. **14**, 88–91 (2002)
26. Diaz-Barrero, J.L., Egozcue, J.J.: Bounds for the moduli of zeros. Appl. Math. Lett. **17**, 993–996 (2004)
27. Dieudonné, J.: La théorie analytique des polynômes d'une variable. Mémor. Sci. Math. **93**, 1–71 (1938)
28. Donaghey, R., Shapiro, L.W.: Motzkin numbers. J. Comput. Theor. **23**, 291–301 (1977)
29. Ehrlich, L.W.: A modified Newton method for polynomials. Commun. ACM **10**, 107–108 (1967)
30. Fejér, L.: Üeber Kreisgebiete, in denen eine Wurzel einer algebraischen Glieichung liegt. Jber. Deutsch. Math. Verein. **26**, 114–128 (1917)
31. Fell, H.: The geometry of zeros of trinomials equations. Rend. Circ. Mat. Palermo **28**(2), 303–336 (1980)
32. Fujiwara, M.: A Ueber die Wurzeln der algebraischen Gleichungen. Tôhoku Math. J. **8**, 78–85 (1915)
33. Gauss, K.F.: Beiträge zur Theorie der algebraischen Gleichungen. Abh. Ges. Wiss. Göttingen **4**; Ges. Werke **3**, 73–102 (1850)
34. Goodman, A.W., Schoenberg, I.J.: A proof of Grace's theorem by induction. Honam Math. J. **9**, 1–6 (1987)
35. Govil, N.K., Kumar, P.: On the annular regions containing all the zeros of a polynomial (preprint)
36. Grace, J.H.: The zeros of a polynomial. Proc. Camb. Philos. Soc. **11**, 352–357 (1901)
37. Heitzinger, W., Troch, W.I., Valentin, G.: Praxisnichtlinearer Gleichungen. Carl Hanser Varlag, München-Wien (1985)
38. Jain, V.K.: On Cauchy's bound for zeros of a polynomial. Turk. J. Math. **30**, 95–100 (2006)
39. Jankowski, W.: Sur les zéros dún polynomial contenant un paramètres arbitraires. Ann. Polon. Math. **3**, 304–311 (1957)
40. Joyal, A., Labelle, G., Rahman, Q.I.: On the location of zeros of polynomials. Can. Math. Bull. **10**, 53–63 (1967)
41. Kalantari, B.: An infinite family of bounds on zeros of analytic functions and relationship to smale's bound. Math. Comput. **74**, 841–852 (2005)
42. Kelleher, S.B.: Des limites des zéroes d'une polynome. J. Math. Pures Appl. **2**, 167–171 (1916)
43. Kim, S.-H.: On the moduli of the zeros of a polynomial. Am. Math. Mon. **112**, 924–925 (2005)
44. Kojima, J.: On the theorem of Hadamard and its applications. Tôhoku Math. J. **5**, 54–60 (1914)
45. Kuniyeda, M.: Notes on the roots of algebraic equation. Tôhoku Math. J. **9**, 167–173 (1916)
46. Laudau, E.: Über den Picardschen Satz, Vierteijahrsschrift Naturforsch. Gesellschaft Zürich **51**, 252–318 (1906)
47. Laudau, E.: Sur quelques généralisations du théorème de M. Picard. Ann. École Norm (3) **24**, 179–201 (1907)
48. Marden, M.: The zeros of certain composite polynomials. Bull. Am. Math. Soc. **49**, 93–100 (1943)
49. Marden, M.: Geometry of Polynomials. Mathematical Surveys and Monographs, vol. 3. American Mathematical Society, Providence, RI (1966)
50. Markovitch, D.: On the composite polynomials. Bull. Soc. Math. Phys. Serbie **3**(3–4), 11–14 (1951)

51. Milovanović, G.V., Petković, M.S.: On computational efficieny of the iterative methods for the simultaneous approximation of polynomial zeros. ACM Trans. Math. Softw. **12**, 295–306 (1986)
52. Milovanovic, G.V., Rassias, M.T.: Analytic Number Theory, Approximation Theory, and Special Function. Springer, Berlin (2014)
53. Milovanovic, G.V., Mitrinovic, D.S., Rassias, T.M.: Topics in Polynomials: Extremal Problems, Inequalities, Zeros. World Scientific, Singapore (1994)
54. Miodrag, S.P.: A highly efficient root-solver of very fast convergence. Appl. Math. Comput. **205**, 298–302 (2008)
55. Montel, P.: Sur la limite supéieure des modules des zéros des polynômes. C. R. Acad. Sci. Paris **193**, 974–976 (1931)
56. Narayana, T.V.: Sur les treillis formes par les partitions d'une unties et leurs applications a la theorie des probabilites. Comp. Rend. Acad. Sci. Paris. **240** (1955), 1188–1189
57. Nourein, A.W.M.: An improvement on two iteration methods for simultaneously determination of the zeros of a polynomial. Int. J. Comput. Math. **6**, 241–252 (1977)
58. Pachter, L., Sturmdfels, B.: Algebraic Statistics for Computational Biology. Cambridge University Press, Cambridge (2005)
59. Peretz, R., Rassias, T.M.: Some remarks on theorems of M. Marden concerning the zeros of certain composite polynomials. Complex Variables **18**, 85–89 (1992)
60. Prasolov, V.V.: Polynomials. Springer, Berlin (2004)
61. Rahman, Q.I., Schmeisser, G.: Analytic Theory of Polynomials. Oxford University Press, New York (2002)
62. Rather, N.A., Mattoo, S.G.: On annulus containing all the zeros of a polynomial. Appl. Math. E-Notes **13**, 155–159 (2013)
63. Rubinstein, Z.: Some results in the location of the zeros of linear combinations of polynomials. Trans. Am. Math. Soc. **116**, 1–8 (1965)
64. Sagan, H.: Boundary and Eigenvalue Problems in Mathematical Physics. Dover Publications, Mineola (1989)
65. Schur, J.: Zwei Sätze über algebraische Gleichungen mit lauter rellen Wurzeln. J. Reine Agnew. Math. **144**, 75–88 (1914)
66. Sun, Y.J., Hsieh, J.G.: A note on circular bound of polynomial zeros. IEEE Trans. Circ. Syst. I **43**, 476–478 (1996)
67. Szegö, G.: Bemerkungen zu einem Satz von J. H. Grace über die Wurzeln algebraischer Gleichungen. Math. Z. **13**, 28–55 (1922)
68. Tôya, T.: Some remarks on Montel's paper concerning upper limits of absolute values of roots of algebraic equations. Sci. Rep. Tokyo Bunrika Daigaku **A1**, 275–282 (1933)
69. Walsh, J.L.: An inequality for the roots of an algebraic equation. Ann. Math. **25**, 285–286 (1924)
70. Williams, K.P.: Note concerning the roots of an equation. Bull. Am. Math. Soc. **28**, 394–396 (1922)
71. Yayenie, O.: A note on generalized Fibonacci sequences. Appl. Math. Comput. **217**, 5603–5611 (2011)
72. Zedek, M.: Continuity and location of zeros of linear combination of polynomials. Proc. Am. Math. Soc. **16**, 78–84 (1965)
73. Zeheb, F.: On the largest modulus of polynomial zeros. IEEE Trans. Circ. Syst. **I** 49, 333–337 (1991)
74. Žilović, M.S., Roytman, L.M., Combettes, P.L., Swamy, M.N.S.: A bound for the zeros of polynomials. IEEE Trans. Circ. Syst. **I** 39, 476–478 (1992)
75. Zölzer, U.: Digital Audio Signal Processing. Wiley, New York (1997)

Approximation by Durrmeyer Type Operators Preserving Linear Functions

Vijay Gupta

Abstract In the present article, we propose a new sequence of linear positive operators having different basis, which are generalizations of Bernstein basis functions. We establish some convergence estimates which include link convergence, asymptotic formula, and direct estimates in terms of usual and Ditzian–Totik modulus of continuity.

Keywords: Bernstein polynomials • Modulus of continuity • Rising factorial • Direct results

1 Introduction

For $f \in C[0,1]$ Bernstein polynomials are defined as

$$P_n(f,x) := \sum_{k=0}^{n} p_{n,k}(x) f\left(\frac{k}{n}\right), x \in [0,1], \tag{1}$$

where

$$p_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}.$$

For all $n \geq 1$, $B_n(f,0) = f(0)$, $B_n(f,1) = f(1)$ so that a Bernstein polynomial for f interpolates f at both endpoints of the interval $[0,1]$. In the year 1968, Stancu [21] introduced a generalization of the Bernstein polynomials based on Polya distribution. The generalized operators $P_n^{(\alpha)} : C[0,1] \to C[0,1]$, introduced in [21] are positive linear operators and depend on a non-negative parameter α, which are defined as

V. Gupta (✉)
Department of Mathematics, Netaji Subhas Institute of Technology,
Sector 3 Dwarka, New Delhi 110078, India
e-mail: vijaygupta2001@hotmail.com

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_11

$$P_n^{(\alpha)}(f,x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) p_{n,k}^{(\alpha)}(x), \tag{2}$$

where $p_{n,k}^{(\alpha)}(x)$ is the Polya distribution with density function given by

$$p_{n,k}^{(\alpha)}(x) = \binom{n}{k} \frac{\prod_{\nu=0}^{k-1}(x+\nu\alpha)\prod_{\mu=0}^{n-k-1}(1-x+\mu\alpha)}{\prod_{\lambda=0}^{n-1}(1+\lambda\alpha)}, x \in [0,1].$$

In case $\alpha = 0$ these operators reduce to the classical Bernstein polynomials. For $\alpha = 1/n$ a special case of the operators (2) was considered by Lupaş and Lupaş [16], which can be represented in an alternate form as

$$P_n^{(1/n)}(f,x) = \frac{2(n!)}{(2n)!} \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) (nx)_k (n-nx)_{n-k}, \tag{3}$$

where the rising factorial is given as $(x)_n = x(x+1)(x+2)\ldots(x+n-1)$. In order to approximate Lebesgue integrable functions on $[0,1]$ Durrmeyer [7] in the year 1967 introduced the integral modification of the classical Bernstein polynomials (1), which were also studied in [2]. Twenty years later in the year 1987 Chen [4] and Goodman and Sharma [10] simultaneously introduced the genuine Bernstein polynomials, which preserve linear functions. Suppose $L_B[0,1]$ denote the space of bounded Lebesgue integrable functions on $[0,1]$ and $\prod_n$ the space of polynomials of degree at most $n \in N$. The genuine operators $D_n : L_B[0,1] \to \prod_n, n \geq 1$ are defined as

$$\begin{aligned} D_n(f,x) = (n-1)\sum_{k=1}^{n-1} p_{n,k}(x) \int_0^1 p_{n-2,k-1}(t) f(t)dt \\ + (1-x)^n f(0) + x^n f(1), f \in L_B[0,1]. \end{aligned} \tag{4}$$

Some other generalizations of Bernstein polynomials have been introduced and studied in [1, 8, 11, 13–15, 19] and [18], etc., but they only reproduce constant functions. For $f \in L_B[0,1]$ and $x \in [0,1]$, Păltănea in the year 2007 gave the modification of the operators (4) based on certain parameter $\rho > 0$ as

$$\begin{aligned} D_n^{\rho}(f,x) = \sum_{k=1}^{n-1} p_{n,k}(x) \int_0^1 \mu_{n,k}^{\rho}(t) f(t)dt \\ + (1-x)^n f(0) + x^n f(1), f \in L_B[0,1], \end{aligned} \tag{5}$$

where

$$\mu_{n,k}^{\rho}(t) := \frac{t^{k\rho+m-1}(1-t)^{(n-k)\rho-1}}{B(k\rho,(n-k)\rho)}.$$

These operators were studied in detail by Gonska and Păltănea in [9], where some direct results in simultaneous approximation have also been discussed. We propose

here a genuine integral type modification of the operators (3), with the weights of $\mu_{n,k}^{\rho}(t)$ considered in (5) as

$$D_n^{(1/n,\rho)}(f,x) = \sum_{k=1}^{n-1} p_{n,k}^{(1/n)}(x) \int_0^1 \mu_{n,k}^{\rho}(t) f(t)\, dt + p_{n,0}^{(1/n)}(x) f(0) + p_{n,n}^{(1/n)}(x) f(1), \tag{6}$$

where

$$p_{n,k}^{(1/n)}(x) = \frac{2(n!)}{(2n)!} \binom{n}{k} (nx)_k (n - nx)_{n-k}$$

and for $1 \leq k \leq n-1$,

$$\mu_{n,k}^{\rho}(t) := \frac{t^{k\rho+m-1}(1-t)^{(n-k)\rho-1}}{B(k\rho,(n-k)\rho)}.$$

Obviously the operators (6) are linear positive operators and preserve the linear functions, we may call such operators as genuine operator. In the limiting case of these operators, we recapture the operators (3). In the recent years many approximation properties have been discussed, we mention some of them as [3, 12, 20]. We estimate some direct results for these operators.

2 Basic Results

In the sequel, we shall need the following basic results:

Lemma 1 (Miclaus [17]). *For the operators defined by (3) with* $e_i(t) = t^i, i = 0, 1, 2$ *we have*

$$P_n^{(1/n)}(e_0,x) = 1, P_n^{(1/n)}(e_1,x) = x$$

and

$$P_n^{(1/n)}(e_2,x) = \frac{nx^2 + 2x - x^2}{n+1} = x^2 + \frac{2x(1-x)}{n+1}.$$

Lemma 2. *For the operators defined by (6), we have*

$$D_n^{(1/n,\rho)}(e_0,x) = 1, D_n^{(1/n,\rho)}(e_1,x) = x$$

$$D_n^{(1/n,\rho)}(e_2,x)=\frac{n^2\rho x^2+2n\rho x-n\rho x^2+nx+x}{(n+1)(n\rho+1)}.$$

Proof. Using the definition of Beta functions of first kind, it follows that

$$\int_0^1 \mu_{n,k}^{\rho}(t)\,t^m dt=\frac{1}{B(k\rho,(n-k)\rho)}\int_0^1 t^{k\rho+m-1}(1-t)^{(n-k)\rho-1}dt \tag{7}$$
$$=\frac{B((k\rho+m,(n-k)\rho)}{B(k\rho,(n-k)\rho)}=\frac{\Gamma(k\rho+m)}{\Gamma(n\rho+m)}.\frac{\Gamma(n\rho)}{\Gamma(k\rho)}.$$

Thus

$$\int_0^1 \mu_{n,k}^{\rho}(t)\,tdt=\frac{k}{n}$$

and

$$\int_0^1 \mu_{n,k}^{\rho}(t)\,t^2dt=\frac{(k\rho+1)k}{(n\rho+1)n}=\frac{n\rho}{(n\rho+1)}\frac{k^2}{n^2}+\frac{1}{(n\rho+1)}\frac{k}{n}.$$

Obviously $D_n^{(1/n,\rho)}(e_0,x)=1$. Next using (7) and applying Lemma 1, we have

$$D_n^{(1/n,\rho)}(e_1,x)=\sum_{k=1}^{n-1}p_{n,k}^{(1/n)}(x)\,\frac{k}{n}+p_{n,n}^{(1/n)}(x)=x.$$

Finally

$$D_n^{(1/n,\rho)}(e_2,x)=\sum_{k=0}^{n}p_{n,k}^{(1/n)}(x)\left[\frac{n\rho}{(n\rho+1)}\frac{k^2}{n^2}+\frac{1}{(n\rho+1)}\frac{k}{n}\right]$$
$$=\frac{n\rho}{(n\rho+1)}P_n^{(1/n)}(e_2,x)+\frac{1}{(n\rho+1)}P_n^{(1/n)}(e_1,x)$$
$$=\frac{n\rho}{(n\rho+1)}.\frac{nx^2+2x-x^2}{n+1}+\frac{x}{(n\rho+1)}$$
$$=\frac{n^2\rho x^2+2n\rho x-n\rho x^2+nx+x}{(n+1)(n\rho+1)}.$$

Remark 1. If we denote $T_{n,r}^{\rho}(x)=D_n^{(1/n,\rho)}((t-x)^r,x)$, then we get

$$T_{n,1}^{\rho}(x)=0,T_{n,2}^{\rho}(x)=\frac{(2n\rho+n+1)x(1-x)}{(n+1)(n\rho+1)}.$$

Moreover, we have

$$T_{n,m}^{\rho}(x) = O(n^{-[(m+1)/2]}),$$

where $[a]$ denote the integral part of a.

Lemma 3. *For $f \in C[0,1]$, we have $\left\| D_n^{(1/n,\rho)}(f,x) \right\| \leq \|f\|$, where $||.||$ is the sup-norm on $[0,1]$.*

Proof. From the definition of operator and Lemma 2, we get

$$\left| D_n^{(1/n,\rho)}(f,x) \right| \leq \|f\| D_n^{(1/n,\rho)}(1,x) = \|f\|.$$

Lemma 4. *For $n \in N$, we have*

$$D_n^{(1/n,\rho)}\left((t-x)^2, x\right) \leq \frac{2\rho+1}{n\rho+1}\delta_n^2(x),$$

where $\delta_n^2(x) = \varphi^2(x) + \frac{1}{n+1}$, where $\varphi^2(x) = x(1-x)$.

Proof. By Remark 1, we have

$$D_n^{(1/n,\rho)}\left((t-x)^2, x\right) = \frac{(2n\rho+n+1)x(1-x)}{(n+1)(n\rho+1)} \leq \frac{2\rho+1}{n\rho+1}\left[\varphi^2(x) + \frac{1}{n+1}\right],$$

which is desired.

3 Convergence Estimates

In this section, we present some convergence estimates of the operators $D_n^{(1/n)}(f,x)$.

Theorem 1. *For any $f \in C[0,1]$, we have*

$$\lim_{\rho\to\infty} D_n^{(1/n,\rho)}(f,x) = P_n^{(1/n)}(f,x), \quad \textit{uniformly.}$$

Proof. Let $f \in C[0,1]$ and suppose $n \in N$ be fixed. For fixed k and n with $1 \leq k \leq n-1$, it suffices to show that

$$\lim_{\rho\to\infty} F_{n,k}^{\rho}(f) = \lim_{\rho\to\infty}\left[\int_0^1 \mu_{n,k}^{\rho}(t)f(t)dt + f(0) + f(1)\right] = f\left(\frac{k}{n}\right).$$

But this is a consequence of well-known Korovkin's theorem, applied to the situation $F_{n,k}^{\rho}(f) : C[0,1] \to C[k/n, k/n]$. By Lemma 2, we have $F_{n,k}^{\rho}(e_1) = k/n$ and $F_{n,k}^{\rho}(e_2) = k^2/n^2$ for sufficiently large ρ. Hence $F_{n,k}^{\rho}(f) \to f(k/n)$ as $\rho \to \infty$.

Theorem 2. *Let $f \in C[0,1]$ and if f'' exists at a point $x \in [0,1]$, then*

$$\lim_{n\to\infty} n\left[D_n^{(1/n,\rho)}(f,x) - f(x)\right] = \frac{(2\rho+1)x(1-x)}{2} f''(x).$$

Proof. By Taylor's expansion of f, we have

$$f(t) = f(x) + (t-x)f'(x) + \frac{(t-x)^2}{2} f''(x) + \varepsilon(t,x)(t-x)^2,$$

where $\varepsilon(t,x) \to 0$ as $t \to x$. Applying $D_n^{(1/n,\rho)}$ on above Taylor's expansion and using Remark 1, we have

$$D_n^{(1/n)}(f,x) - f(x) = f'(x) D_n^{(1/n,\rho)}((t-x),x) + \frac{1}{2} f''(x) D_n^{(1/n,\rho)}((t-x)^2,x) + D_n^{(1/n,\rho)}(\varepsilon(t,x)(t-x)^2,x).$$

Thus

$$\begin{aligned}
&\lim_{n\to\infty} n\left[D_n^{(1/n,\rho)}(f,x) - f(x)\right] \\
&= \lim_{n\to\infty} n\frac{1}{2} f''(x)\, D_n^{(1/n,\rho)}((t-x)^2,x) + \lim_{n\to\infty} nD_n^{(1/n,\rho)}(\varepsilon(t,x)(t-x)^2,x) \\
&= \frac{(2\rho+1)x(1-x)}{2} f''(x) + \lim_{n\to\infty} nD_n^{(1/n,\rho)}\left(\varepsilon(t,x)(t-x)^2, x\right) \\
&=: \frac{(2\rho+1)x(1-x)}{2} f''(x) + F.
\end{aligned}$$

In order to complete the proof, it is sufficient to show that $F = 0$. By Cauchy–Schwarz inequality, we have

$$F = \lim_{n\to\infty} nD_n^{(1/n,\rho)}\left(\varepsilon^2(t,x), x\right)^{1/2} D_n^{(1/n,\rho)}\left((t-x)^4, x\right)^{1/2}. \tag{8}$$

Furthermore, since $\varepsilon^2(x,x) = 0$ and $\varepsilon^2(.,x) \in C[0,1]$, it follows that

$$\lim_{n\to\infty} nD_n^{(1/n,\rho)}\left(\varepsilon^2(t,x), x\right) = 0, \tag{9}$$

uniformly with respect to $x \in [0,1]$. Thus from (8), (9) and application of Remark 1, we get

$$\lim_{n\to\infty} nD_n^{(1/n,\rho)}\left(\varepsilon^2(t,x), x\right)^{1/2} D_n^{(1/n,\rho)}\left((t-x)^4, x\right)^{1/2} = 0.$$

Thus, we have

$$\lim_{n\to\infty} n\left[D_n^{(1/n,\rho)}(f,x)-f(x)\right]=\frac{(2\rho+1)x(1-x)}{2}f''(x),$$

which completes the proof.

To prove the next direct result, we need the following auxiliary function viz. Peetre's K-functional which for $W^2=\{g\in C[0,1]: g',g''\in C[0,1]\}$ is defined as:

$$K_2(f,\delta)=\inf\left\{\|f-g\|+\delta\left\|g''\right\|: g\in W^2\right\}(\delta>0),$$

where $\|.\|$ is the uniform norm on $C[0,1]$.

Theorem 3. *For the operators $D_n^{(1/n,\rho)}$, there exists a constant $C>0$ such that*

$$\left|D_n^{(1/n,\rho)}(f,x)-f(x)\right|\le C\omega_2\left(f,(n+1)^{-1}\delta_n(x)\right),$$

where $f\in C[0,1]$, $\delta_n(x)=\left[\varphi^2(x)+\frac{1}{n+1}\right]^{1/2}$, $\varphi(x)=\sqrt{x(1-x)}$ and $x\in[0,1]$ and the second order modulus of continuity is given by

$$\omega_2(f,\eta)=\sup_{0<h\le\eta}\sup_{x,x+2h\in[0,1]}|f(x+2h)-2f(x+h)+f(x)|.$$

Proof. By Taylor's formula, we can write

$$g(t)=g(x)+(t-x)g'(x)+\int_x^t(t-u)g''(u)\,du.$$

Applying the above Taylor's formula, we have

$$D_n^{(1/n,\rho)}(g,x)=g(x)+D_n^{(1/n,\rho)}\left(\int_x^t(t-u)g''(u)\,du,x\right).$$

Hence

$$\begin{aligned}\left|D_n^{(1/n,\rho)}(g,x)-g(x)\right|&\le D_n^{(1/n,\rho)}\left(\int_x^t|t-u|\left|g''(u)\right|du,x\right)\\&\le D_n^{(1/n,\rho)}\left((t-x)^2,x\right)\left\|g''\right\|.\end{aligned}$$

For $f\in C[0,1]$ and $g\in W^2$, using Lemmas 2 and 3 we have

$$\begin{aligned}&\left|D_n^{(1/n,\rho)}(f,x)-f(x)\right|\\&\le\left|D_n^{(1/n,\rho)}(f-g,x)-(f-g)(x)\right|+\left|D_n^{(1/n,\rho)}(g,x)-g(x)\right|\\&\le 2\|f-g\|+\frac{3}{n+1}\delta_n^2(x)\left\|g''\right\|.\end{aligned}$$

Taking infimum over all $g \in W^2$, we obtain

$$\left|D_n^{(1/n,\rho)}(f,x) - f(x)\right| \leq 3K_2\left(f, \frac{1}{n+1}\delta_n^2(x)\right).$$

Using the inequality due to DeVore and Lorentz [5], there exists a positive constant $C > 0$ such that

$$K_2(f,\delta) \leq C\omega_2\left(f, \sqrt{\delta}\right),$$

we get at once

$$\left|D_n^{(1/n,\rho)}(f,x) - f(x)\right| \leq C\omega_2\left(f, (n+1)^{-1}\delta_n(x)\right),$$

so the proof is completed.

Let $f \in C[0,1]$ and $\varphi(x) = \sqrt{x(1-x)}$, $x \in [0,1]$. The second order Ditzian–Totik modulus of smoothness and corresponding K-functional are given by, respectively,

$$\omega_2^{\varphi}\left(f, \sqrt{\delta}\right) = \sup_{0<h\leq\sqrt{\delta}} \sup_{x\pm h\varphi(x)\in[0,1]} \left|f(x+h\varphi(x)) - 2f(x) + f(x-h\varphi(x))\right|,$$

$$\bar{K}_{2,\varphi}(f,\delta) = \inf\left\{\|f-g\| + \delta\left\|\varphi^2 g''\right\| + \delta^2\left\|g''\right\| : g \in W^2(\varphi)\right\} (\delta > 0),$$

where $W^2(\varphi) = \left\{g \in C[0,1] : g' \in AC_{\text{loc}}[0,1], \varphi^2 g'' \in C[0,1]\right\}$ and $g' \in AC_{\text{loc}}[0,1]$ means that g is differentiable and g' is absolutely continuous on every closed interval $[a,b] \subset [0,1]$. We know from Theorem 1.3.1 of [6] that there exists a positive constant $C > 0$, such that

$$\bar{K}_{2,\varphi}(f,\delta) \leq C\omega_2^{\varphi}\left(f, \sqrt{\delta}\right). \tag{10}$$

Our next direct estimate is in terms of the Ditzian–Totik modulus of continuity.

Theorem 4. *Let $f \in C[0,1]$. Then for $x \in [0,1]$, we have*

$$\left\|D_n^{(1/n,\rho)}f - f\right\| \leq C\omega_2^{\varphi}\left(f, (n+1)^{-1/2}\right),$$

where $C > 0$ is an absolute constant and $\varphi(x) = \sqrt{x(1-x)}$.

Proof. By Taylor's formula, we can write

$$g(t) = g(x) + (t-x)\, g'(x) + \int_x^t (t-u)\, g''(u)\, du.$$

Using the definition of the operator $D_n^{(1/n,\rho)}$ and Lemma 2, we obtain

$$\left|D_n^{(1/n,\rho)}(g;x) - g(x)\right| \le D_n^{(1/n,\rho)}\left(\int_x^t |t-u| \left|g''(u)\right| du; x\right).$$

Moreover, δ_n^2 is a concave function on $x \in [0,1]$, for $u = \lambda x + (1-\lambda)\, t$, $\lambda \in [0,1]$, we get

$$\frac{|t-u|}{\delta_n^2(u)} = \frac{\lambda\,|t-x|}{\delta_n^2(\lambda x + (1-\lambda)\, t)} \le \frac{\lambda\,|t-x|}{\lambda \delta_n^2(x) + (1-\lambda)\,\delta_n^2(t)} \le \frac{|t-x|}{\delta_n^2(x)}.$$

Thus we have

$$\left|D_n^{(1/n,\rho)}(g,x) - g(x)\right| \le \frac{1}{\delta_n^2(x)} \left\|\delta_n^2 g''\right\| D_n^{(1/n,\rho)}\left((t-x)^2, x\right).$$

By using Lemma 4, we have

$$\left|D_n^{(1/n,\rho)}(g,x) - g(x)\right| \le \frac{2\rho+1}{n\rho+1} \left\|\delta_n^2 g''\right\|. \tag{11}$$

Applying Lemma 3 and (11), we have for $f \in C[0,1]$,

$$\begin{aligned}\left|D_n^{(1/n,\rho)}(f,x) - f(x)\right| &\le \left|D_n^{(1/n,\rho)}(f-g,x)\right| + \left|D_n^{(1/n,\rho)}(g,x) - g(x)\right| \\ &\quad + |g(x) - f(x)| \\ &\le 4\|f-g\| + \frac{2\rho+2}{n+1}\left\|\varphi^2 g''\right\| + \frac{2\rho+2}{(n+1)^2}\left\|g''\right\|.\end{aligned}$$

Taking infimum over all $g \in W^2$, we obtain

$$\left|D_n^{(1/n,\rho)}(f,x) - f(x)\right| \le C\bar{K}_{2,\varphi}\left(f, \frac{1}{n+1}\right). \tag{12}$$

Therefore, from (10) and (12) we obtain

$$\left\|D_n^{(1/n,\rho)} f - f\right\| \le C\omega_2^{\varphi}\left(f, (n+1)^{-1/2}\right),$$

which is the desired result.

References

1. Abel, U., Gupta, V., Mohapatra, R.N.: Local approximation by a variant of Bernstein Durrmeyer operators. Nonlinear Anal. Theory Methods Appl. **68**(11), 3372–3381 (2008)
2. Agrawal, P.N., Gupta, V.: Simultaneous approximation by linear combination of modified Bernstein polynomials. Bull. Greek Math. Soc. **39**, 29–43 (1989)
3. Aral, A., Gupta, V., Agarwal, R.P.: Applications of q Calculus in Operator Theory, vol. XII, 262 p. Springer, New York (2013)
4. Chen, W.: On the modified Bernstein Durrmeyer operators. In: Report of the Fifth Chinese Conference on Approximation Theory. Zhen Zhou, China (1987)
5. DeVore, R.A., Lorentz, G.G.: Constructive Approximation, Grundlehren der Mathematischen Wissenschaften, Band 303. Springer, Berlin (1993)
6. Ditzian, Z., Totik, V.: Moduli of Smoothness. Springer, New York (1987)
7. Durrmeyer, J.L.: Une formule d' inversion de la Transformee Laplace, Applications a la Theorie des Moments, These de 3e Cycle, Faculte des Sciences de I' Universite de Paris (1967)
8. Finta, Z., Gupta, V.: Approximation by q-Durrmeyer operators. J. Appl. Math. Comput. **29**(1–2), 401–415 (2009)
9. Gonska, H., Păltănea, R.: Simultaneous approximation by a class of Bernstein-Durrmeyer operators preserving linear functions. Czec. Math. J. **60**(135), 783–799 (2010)
10. Goodman, T.N.T., Sharma, A.: A modified Bernstein-Schoenberg operator. In: Sendov, B.I., et al. (eds.) Proceedings of the Conference on Constructive Theory of Functions, Verna 1987, pp. 166–173. Publishing House of the Bulgarian Academy of Sciences, Sofia (1988)
11. Gupta, V.: Some approximation properties on q-Durrmeyer operators. Appl. Math. Comput. **197**(1), 172–178 (2008)
12. Gupta, V., Agarwal, R.P.: Convergence Estimates in Approximation Theory, Springer, Heidelberg (2014)
13. Gupta, V., Maheshwari, P.: Bézier variant of a new Durrmeyer type operators. Riv. Mat. Univ. Parma **7**(2), 9–21 (2003)
14. Gupta, V., Rassias, Th.M.: Lupas-Durmeyer operators based on Polya distribution. Banach J. Math. Anal. **8**(2), 146–155 (2014)
15. Gupta, V., López-Moreno, A., Palacios, J.: On simultaneous approximation of the Bernstein Durrmeyer operators. Appl. Math. Comput. **213**(1), 112–120 (2009)
16. Lupaş, L., Lupaş, A.: Polynomials of binomial type and approximation operators. Stud. Univ. Babes-Bolyai Math. **32**(4), 61–69 (1987)
17. Miclăuş, D.: The revision of some results for Bernstein Stancu type operators. Carpathian J. Math. **28**(2), 289–300 (2012)
18. Morales, D., Gupta, V.: Two families of Bernstein-Durrmeyer type operators. Appl. Math. Comput. **248**, 342–353 (2014)
19. Sinha, T.A.K., Gupta, V., Agrawal, P.N., Gairola, A.R.: Inverse theorem for an iterative combinations of Bernstein-Durrmeyer operators. Stud. Univ. Babes Bolyai Math. **54**(4), 153–164 (2009)
20. Srivastava, H.M., Gupta, V.: A certain family of summation integral type operators. Math. Comput. Model. **37**, 1307–1315 (2003)
21. Stancu, D.D.: Approximation of functions by a new class of linear polynomial operators. Rew. Roum. Math. Pure. Appl. **13**, 1173–1194 (1968)

Revisiting the Complex Multiplication Method for the Construction of Elliptic Curves

Elisavet Konstantinou and Aristides Kontogeorgis

Abstract In this article we give a detailed overview of the Complex Multiplication (CM) method for constructing elliptic curves with a given number of points. In the core of this method, there is a special polynomial called Hilbert class polynomial which is constructed with input a fundamental discriminant $d < 0$. The construction of this polynomial is the most demanding and time-consuming part of the method and thus the use of several alternative polynomials has been proposed in previous work. All these polynomials are called *class polynomials* and they are generated by proper values of modular functions called *class invariants*. Besides an analysis on these polynomials, in this paper we will describe our results about finding new class invariants using the Shimura reciprocity law. Finally, we will see how the choice of the discriminant can affect the degree of the class polynomial and consequently the efficiency of the whole CM method.

Keywords: Elliptic curve cryptography • Computational class field theory • Complex multiplication • Shimura reciprocity

The authors were partially supported by the Project "*Thalis, Algebraic modeling of topological and computational structures*". The Project "THALIS" is implemented under the Operational Project "Education and Life Long Learning"and is co-funded by the European Union (European Social Fund) and National Resources (ESPA).

E. Konstantinou (✉)
Department of Information and Communication Systems Engineering,
University of the Aegean, Karlovassi, Samos 83200, Greece
e-mail: ekonstantinou@aegean.gr

A. Kontogeorgis
Department of Mathematics, University of Athens, Panepistimioupolis, 15784 Athens, Greece
e-mail: kontogar@math.uoa.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_12

1 Introduction

Complex Multiplication (CM) method is a well-known and efficient method for the construction of elliptic curves with a given number of points. In cryptographic applications, it is required that the order of the elliptic curves satisfies several restrictions and thus CM method is a necessary tool for them. Essentially, CM method is a way to use elliptic curves defined over the field of complex numbers in order to construct elliptic curves defined over finite fields with a given number of points. Therefore, we will begin our article by giving a brief introduction to the theory of elliptic curves over a field K, which for our purposes will be either the finite field $\mathbb{F}_p$ or the field of complex numbers $\mathbb{C}$.

We describe the CM method using first the classical j-invariant and its corresponding Hilbert polynomial. Hilbert polynomial is constructed with input a fundamental discriminant $d < 0$. The disadvantage of Hilbert polynomials is that their coefficients grow very large as the absolute value of the discriminant $D = |d|$ increases and thus their construction requires high precision arithmetic and a huge amount of disk space to store and manipulate them.

Supposing that f is a modular function, such that $f(\tau)$ for some $\tau \in \mathbb{Q}(\sqrt{-D})$ generates the Hilbert class field of $\mathbb{Q}(\sqrt{-D})$, then its minimal polynomial can substitute the Hilbert polynomial in the CM method and the value $f(\tau)$ is called *class invariant*. These minimal polynomials are called *class polynomials*, their coefficients are much smaller than their Hilbert counterparts and their use can considerably improve the efficiency of the whole CM method. Some well-known families of class polynomials are: Weber polynomials [15, 23], $M_{D,l}(x)$ polynomials [21], Double eta (we will denote them by $M_{D,p_1,p_2}(x)$) polynomials [6] and Ramanujan polynomials [17]. The logarithmic height of the coefficients of all these polynomials is smaller by a constant factor than the corresponding logarithmic height of the Hilbert polynomials and this is the reason for their much more efficient construction.

In what follows, we will present our contribution on finding alternative class invariants (instead of the classical j-invariant) which can considerably improve the efficiency of the CM method. Also we will see how the choice of the discriminant can affect the efficiency of the class polynomials' construction.

2 Preliminaries

The theory of elliptic curves is a huge object of study and the interested reader is referred to [2, 25] and references within for more information. An *elliptic curve* defined over a field K of characteristic $p > 3$ is the set of all points $(x, y) \in K \times K$ (in affine coordinates) which satisfy an equation

$$y^2 = x^3 + ax + b \tag{1}$$

where $a, b \in K$ satisfy $4a^3 + 27b^2 \neq 0$, together with at special point O_E which is called the point at infinity. The set $E(K)$ of all points can be naturally equipped with a properly defined addition operation and it forms an abelian group, see [3], [38] for more details on this group.

An elliptic curve $E(\mathbb{F}_q)$ defined over a finite field $\mathbb{F}_q$ is then a finite abelian group and as such it is isomorphic to a product of cyclic groups:

$$E(\mathbb{F}_q) \cong \prod_{i=1}^{s} \mathbb{Z}/n_i\mathbb{Z}.$$

The arithmetic complexity of this elliptic curve is reduced to the smallest cyclic factor of the above decomposition. For example, we can have an elliptic curve of huge order which is the product of a large amount of cyclic groups of order 2. The discrete logarithm problem is trivial for this curve. For cryptographic algorithms, we would like to have elliptic curves which do not admit small cyclic factors and even better elliptic curves which have order a large prime number. This forces the curve to consist of only one cyclic factor.

In order to construct an elliptic curve with a proper order, we can either generate random elliptic curves, compute their order and then check their properties or we can use a method which constructs elliptic curves with a given order which we known beforehand that satisfies our restrictions. In this article we will use the second approach and present the method of Complex Multiplication. This method uses the theory of elliptic curves defined over the field of complex numbers in order to construct elliptic curves over finite fields having the desired order.

Definition 1. A lattice L in the field of complex numbers is the set which consists of all linear $\mathbb{Z}$-combinations of two $\mathbb{Z}$-linearly independent elements $e_1, e_2 \in \mathbb{C}$.

Given a lattice L Weierstrass defined a function $\wp$ depending on the lattice L

$$\wp : \mathbb{C} \to \mathbb{C}$$

by the formula:

$$\wp(z, L) = \frac{1}{z^2} + \sum_{\lambda \in L - \{0\}} \left(\frac{1}{(z+\lambda)^2} - \frac{1}{\lambda^2} \right).$$

The function $\wp$ satisfies the differential equation

$$\wp'(z)^2 = 4\wp(z)^3 - g_2(L)\wp(z) - g_3(L).$$

Therefore the pair $(x, y) = (\wp(z), \wp'(z))$ parametrizes the elliptic curve

$$y^2 = 4x^3 - g_2(L)x - g_3(L).$$

Remark 1. The transcendental functions $(x, y) = (\sin(x), \cos(x)) = (\sin(x), \sin'(x))$ satisfy the equation $x^2 + y^2 = 1$, therefore they parametrise the unit circle.

The function $\wp$ is periodic with period the lattice L, i.e.

$$(\wp(z + \lambda), \wp'(z + \lambda)) = (\wp(z), \wp'(z)) \text{ for every } \lambda \in L.$$

At the level of group theory this means that

$$\frac{\mathbb{C}}{L} \cong E(\mathbb{C}).$$

From the topological viewpoint, this means that the fundamental domain of the lattice, i.e. the set

$$z = ae_1 + be_2 : 0 \leq a, b < 1$$

covers the elliptic curve while the border is glued together giving to the elliptic curve the shape of a "donut".

The functions $g_2(L), g_3(L)$ depend on the lattice L, and are given by the formula

$$g_2(L) = 60 \sum_{\lambda \in L - \{0\}} \frac{1}{\lambda^4} \qquad g_3(L) = 140 \sum_{\lambda \in L - \{0\}} \frac{1}{\lambda^6}.$$

2.1 Algebraic Theory of the Equation $y^2 = x^3 + ax + b$

In this paragraph we will study certain invariants of the elliptic curve given by the equation:

$$y^2 = x^3 + ax + b.$$

For every polynomial of one variable $f(x)$ we can define the discriminant. This is a generalization of the known discriminant of a quadratic polynomial and is equal to zero if and only if the polynomial f has a double root.

For the special case of the cubic polynomial $x^3 + ax + b$ the discriminant is given by the formula: $-16(4a^3 + 27b^2)$. We observe that by definition all elliptic curves have non-zero discriminant.

The j-invariant of the elliptic curve is defined by:

$$j(E) = \frac{(4a)^3}{4a^3 + 27b^2} = -\frac{4a^3}{\Delta(E)}.$$

Proposition 1. *Two elliptic curves defined over an algebraically closed field are isomorphic if and only if have the same j-invariant.*

This proposition does not hold if the elliptic curves are considered over a non-algebraically closed field k. They became isomorphic over a quadratic extension of k.

Proposition 2. *For every integer $j_0 \in K$ there is an elliptic curve E defined over K with j-invariant equal to j_0.*

Proof. If $j \neq 0, 1728$, then the elliptic curve defined by

$$E : y^2 + xy = x^3 - \frac{36}{j_0 - 1728}x - \frac{1}{j_0 - 1728}$$

has discriminant

$$\Delta(E) = \frac{j_0^3}{(j_0 - 1728)^3} \text{ and } j(E) = j_0.$$

When $j_0 = 0$ we consider the elliptic curve:

$$E : y^2 + y = x^3, \text{ with } \Delta(E) = -27 \text{ and } j = 0$$

while for j_0 we consider the elliptic curve:

$$E : y^2 = x^3 + x, \text{ with } \Delta(E) = -64 \text{ and } j = 1728.$$

Proposition 3. *Every element in the finite field $\mathbb{F}_p$ is the j-invariant of an elliptic curve defined over $\mathbb{F}_p$. For $j \neq 0, 1728$ this elliptic curve is given by*

$$y^2 = x^3 + 3kc^2x + 2kc^3,$$

for $k = j/(1728-j)$ and c an arbitrary element in $\mathbb{F}_p$. There are two non-isomorphic elliptic curves E, E' over $\mathbb{F}_p$ which correspond to different values of c. They have orders

$$|E| = p + 1 - t \text{ and } |E| = p + 1 + t.$$

In this section we consider the lattices generated by $1, \tau$, where $\tau = a + ib$ is a complex number with $b > 0$. The set of such τ's is called the hyperbolic plane and it is generated by $\mathbb{H}$. In this setting the Eisenstein series, the discriminant and the j-invariant defined above (which depend on L) can be seen as functions of τ.

Proposition 4. *The functions g_2, g_3, Δ, j seen as functions of $\tau \in \mathbb{H}$ remain invariant under transformations of the form:*

$$\tau \mapsto \frac{a\tau + b}{c\tau + d}, \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}).$$

In particular these functions remain invariant under the transformation $\tau \mapsto \tau + 1$ so they are periodic. Hence they admit a Fourier expansion. In the coefficients of the Fourier expansion there is "hidden arithmetic information". For example, the Fourier expansion of the j-invariant function is given by:

$$j(\tau) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + 864299970q^3 + \cdots,$$

where $q = e^{2\pi i \tau}$.

Definition 2. We will say that the function $f : E \to E$ is an endomorphism of the elliptic curve if it can be expressed in terms of rational functions and moreover $f(O_E) = O_E$, where O_E is the neutral element of the elliptic curve.

The set of endomorphisms will be denoted by $\mathrm{End}(E)$ and it has the structure of a ring where addition is the natural addition of functions and multiplication is composition of functions.

If we fix an integer $n \in \mathbb{Z}$, then we can define the endomorphism sending $P \in E$ to nP. In this way $\mathbb{Z}$ becomes a subring of $\mathrm{End}(E)$.

For most elliptic curves defined over fields of characteristic 0, $\mathrm{End}(E) = \mathbb{Z}$. For elliptic curves defined over the finite field $\mathbb{F}_q$, there is always an extra endomorphism the so-called Frobenious endomorphism ϕ, which is defined as follows:

The element $P \in E$ with coordinates (x, y) is mapped to the element $\phi(P)$ with coordinates (x^q, y^q). This endomorphism is interesting because we know that $x \in \bar{\mathbb{F}}_q$ is an element in $\mathbb{F}_q$ if and only if $x^q = x$. So the elements which remain invariant under the action of the Frobenious endomorphism are exactly the points of the elliptic curve over the finite field $\mathbb{F}_p$.

Proposition 5. *The Frobenious endomorphism Φ satisfies the relation*

$$\phi^2 - t\phi + q = 0, \tag{2}$$

where t is an integer called the "trace of Frobenious".

Theorem 1 (H. Hasse). *The trace of Frobenious satisfies*

$$|t| \leq 2\sqrt{q}.$$

Proposition 6. *For a general elliptic curve if there is an extra endomorphism ϕ then it satisfies an equation of the form:*

$$\phi^2 + a\phi + b = 0,$$

with negative discriminant (the term "complex multiplication" owes his name to this fact).

Remark 2. The bound of Hasse is equivalent to the fact that the quadratic equation (2) satisfied by Frobenious has negative discriminant.

Let $\tau \in \mathbb{H}$, for example the one which satisfies the relation

$$\tau^2 - t\tau + q = 0$$

for a negative discriminant D. The theorem of complex multiplication asserts that $j(\tau)$ satisfies an a polynomial $f(x) \in \mathbb{Z}[x]$ end that the elliptic curve E_τ, has j-invariant $j(\tau)$ end endomorphism ring $\mathrm{End}(E_\tau) = \mathbb{Z}[\tau]$.

Moreover, if we reduce the polynomial $f(x)$ modulo p, then the roots of the reduced polynomials are j-invariants which correspond to elliptic curves $\mathbb{F}_p$ with Frobenious endomorphisms ϕ satisfying $\phi^2 - t\phi + q = 0$.

K.F. Gauss in his work Disquisitiones Arithmeticae [8] studied the quadratic forms of discriminant D of the form

$$ax^2 + bxy + cy^2; b^2 - 4ac = -D, a, b, c \in \mathbb{Z} \ \ (a, b, c) = 1,$$

up to the following equivalence relation which in modern language can be defined as: two quadratic forms $f(x, y)$ and $g(x, y)$ are equivalent if there is a transformation $\tau \in \mathrm{SL}(2, \mathbb{Z})$ such that

$$\tau = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ and } f(x, y) = g(ax + by, cx + dy).$$

For more information on this classical subject, we refer to [5].

A full set of representatives $\mathrm{CL}(D)$ of the equivalence classes are the elements (a, b, c) such that

$$|b| \leq a \leq \sqrt{\frac{D}{3}}, a \leq c, (a, b, c) = 1, b^2 - 4ac = -D$$

if $|b| = a$ or $a = c$ then $b \geq 0$.

Theorem 2. *Consider $\tau \in \mathbb{H}$ which satisfies a monic quadratic polynomial in $\mathbb{Z}[x]$. Consider the elliptic curve $E_\tau = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ which has j-invariant $j(\tau)$.*

The complex number $j(\tau)$ satisfies an algebraic equation given by:

$$H_D(x) = \prod_{[a,b,c] \in \mathrm{CL(D)}} \left(x - j\left(\frac{-b + \sqrt{-D}}{2a} \right) \right) \in \mathbb{Z}[x].$$

Moreover a root of the reduction of the polynomial $H_D(x)$ modulo p corresponds to an elliptic curve with Frobenious endomorphism sharing the same characteristic polynomial with τ.

Example. For $D = 491$ we have compute the following equivalence classes for quadratic forms of discriminant -491

$$\mathrm{CL}(D) = [1, 1, 123], [3, \pm 1, 41], [9, \pm 7, 15], [5, \pm 3, 25], [11, \pm 9, 3].$$

For each of the above $[a, b, c]$ we compute the root

$$\rho = \frac{-b + i\sqrt{491}}{2s},$$

of positive imaginary part.

This computation is summarized to the following table:

$[a, b, c]$	Root	j-invariant
$[1, 1, 123]$	$\rho_1 = (-1 + i\sqrt{491})/2$	$-1.7082855E30$
$[3, 1, 41]$	$\rho_2 = (-1 + i\sqrt{491})/6$	5.977095 E9 + 1.0352632 E10I
$[3, -1, 41]$	$\rho_3 = (1 + i\sqrt{491})/6$	5.9770957 E9 − 1.0352632 E10I
$[9, 7, 15]$	$\rho_4 = (-7 + i\sqrt{491})/18$	−1072.7816 + 1418.3793I
$[9, -7, 15]$	$\rho_5 = (7 + i\sqrt{491})/18$	−1072.7816 −1418.3793I
$[5, 3, 25]$	$\rho_6 = (-3 + i\sqrt{491})/10)$	−343205.38 + 1058567.0I
$[5, -3, 25]$	$\rho_7 = (3 + i\sqrt{491})/10$	−343205.38 − 1058567.0I
$[11, 9, 13]$	$\rho_8 = (-9 + i\sqrt{491})/22$	6.0525190 + 170.50800I
$[11, -9, 13]$	$\rho_9 = (9 + i\sqrt{491})/22$	6.0525190 − 170.50800I

We can now compute the polynomial

$$f(x) = \prod_{i=1}^{9} (x - j(\rho_i))$$

with 100-digit precision and we arrive at (computations by magma algebra system [3])

```
x^9 + (17082855199382935607111650508800.0 + 0.E-105*I)*x^8 +
(-20419995943814746224552691418802908299264.0 + 5.527 E-76*I)*x^7 +
(24410449766543274815871531378358311021155670228992 0.0 - 3.203 E-66*I)*x^6 +
(16806109970717648926762170533796935238933528040486 3647744.0 - 8.477 E-61*I)*x^5 +
(30266340622871033999335677742593898488443328160369893 4574743552.0 + 1.179E-53*I)*x^4 +
(64548590008561678492635478603558110892092369718837594939539324928 0.0 + 5.552 E-50*I)*x^3 +
(95704113804639787096552080857655294919888566573818364375039492069785 6.0 - 1.530 E-47*I)*x^2 +
(73228628710337844192365961292732508455291085022217625565074454720020 48.0 + 4.458 E-45*I)*x +
(27831365943253888043128977216106999444228139865055751457267582234307592192.0 - 3.587 E-43*I)
```

which we recognize as a polynomial with integer coefficients (all complex coefficients multiplied by 10^{-40} or a smaller power are considered to be zero and are just floating point approximation garbage).

3 Complex Multiplication Method and Shimura Reciprocity Law

We would like to construct an elliptic curve defined over the finite field $\mathbb{F}_p$ with order $p+1-m$. For this case, we must construct the appropriate $j \in \mathbb{F}_p$. The bound of Hasse gives us that $Z := 4p-(p+1-m)^2 \geq 0$. We write $Z = Dv^2$ as a square v^2 times a number D which is squarefree.

The equation $4p = u^2 + Dv^2$ for some integer u satisfies $m = p+1 \pm u$. The negative integer$-D$ is called the CM-discriminant for the prime p.

We have $x^2 - \mathrm{tr}(\phi)x + p \mapsto \Delta = \phi(F)^2 - 4p = -Dv^2$.

Algorithm:

1. Select a prime p. Select the least D together with $u, v \in \mathbb{Z}$ such that $4p = u^2 + Dv^2$.
2. If one of the values $p+1-u, p+1+u$ is a prime number, then we proceed to the next steps, otherwise we go back to step 1.
3. We compute the Hilbert polynomial $H_D(x) \in \mathbb{Z}[x]$ using floating approximations of the j-invariant.
4. Reduce modulo p and find a root of $H_D(x) mod p$. This root is the desired j-invariant. The elliptic curve corresponding to j-invariant $j \neq 0, 1728$ is

$$y^2 = x^3 + 3kc^2x + 2kc^3, k = j/(1728-j), c \in \mathbb{F}_p.$$

To different values of c correspond two different elliptic curves E, E' which have orders $p+1 \pm t$. One is

$$y^2 = x^3 + ax + b$$

and the other is

$$y^2 = x^3 + ac^2x + bc^3,$$

where c is a quadratic non-residue in $\mathbb{F}_p$. In order to select the elliptic curve with the correct order we choose a point P in one of them and we compute its order, i.e. the natural number n such that $nP = O_E$. This order should divide either $p+1-t$ or $p+1+t$.

The CM method for every discriminant D requires the construction of polynomial $H_D(x) \in \mathbb{Z}[x]$ (called the Hilbert polynomial)

$$H_D(x) = \prod_{\tau}(x - j(\tau)),$$

for all values $\tau = (-b + \sqrt{-D})/2a$ for all integers $[a, b, c]$ running over a set of representatives of the group of equivalent quadratic forms.

Let h be the order of $\mathrm{Cl}(D)$. It is known that the bit precision required of the generation of $H_D(x)$ (see [20]):

$$\mathrm{H}-\mathrm{Prec}(D) \cong \frac{\ln 10}{\ln 2}(h/4+5)+\frac{\pi\sqrt{D}}{\ln(2)}\sum_{\tau}\frac{1}{a}.$$

The most demanding step of the CM-method is the construction of the Hilbert polynomial, as it requires high precision floating point and complex arithmetic. As the value of the discriminant D increases, the coefficients of the grow extremely large and their computation becomes more inefficient.

In order to overcome this difficulty, alternative class functions were proposed by several authors. It was known in the literature [13, 27, 28] that several other complex valued functions can be used in order to construct at special values the Hilbert class field. Usually one tries functions of the form

$$\frac{\eta(p\tau)}{\eta(\tau)} \text{ or } \frac{\eta(p\tau)\eta(q\tau)}{\eta(pq\tau)\eta(\tau)},$$

where η is the Dedekind zeta function defined by

$$\eta(\tau)=e^{2\pi i\tau/24}\prod_{n\geq 1}(1-q^n), \tau\in\mathbb{C}, \mathrm{Im}(\tau)>0, q=e^{2\pi i\tau}.$$

All such constructions have the Shimura reciprocity law as ingredient or can be written in this language. This technique was proposed by Shimura [24], but it was Gee and Stevenhagen [9–11, 26] who put it in form suitable for applications. In order to define Shimura reciprocity law, we have to define some minimum amount of the theory of modular functions.

Consider the group $\mathrm{SL}(2,\mathbb{Z})$ consisted by all 2×2 matrices with integer entries and determinant 1. It is known that an element

$$\sigma := \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$$

acts on the upper complex plane $\mathbb{H} := \{z\in\mathbb{C} : \mathrm{Im}(z) > 0\}$ by Möbious transformations by

$$\sigma z = \frac{az+b}{cz+d}.$$

Moreover it is known that $\mathrm{SL}(2,\mathbb{Z})$ can be generated by the elements $S : z\mapsto -\frac{1}{z}$ and $T : z\mapsto z+1$. Let $\Gamma(N)$ be the kernel of the map $\mathrm{SL}(2,\mathbb{Z})\mapsto \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$.

Let $\mathbb{H}^*$ be the upper plane $\mathbb{H}\cup\mathbb{P}^1(\mathbb{Q})$. One can show that the quotient $\Gamma(N)\backslash\mathbb{H}^*$ has the structure of a compact Riemann surface which can be described as an algebraic curve defined over the field $\mathbb{Q}(\zeta_N)$, where ζ_N is a primitive N-th root of

unity. We consider the function field F_N of this algebraic curve defined over $\mathbb{Q}(\zeta_N)$. The function field F_N is acted on by

$$\Gamma(N)/\{\pm 1\} \cong \mathrm{Gal}(F_N/F_1(\zeta_N)).$$

For an element $d \in \left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)^*$ we consider the automorphism $\sigma_d : \zeta_N \mapsto \zeta_N^d$. Since the Fourier coefficients of a function $h \in F_N$ are known to be in $\mathbb{Q}(\zeta_N)$, we consider the action of σ_d on F_N by applying σ_d on the Fourier coefficients of h. In this way we define an arithmetic action of

$$\mathrm{Gal}(F_1(\zeta_N)/F_1) \cong \mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q}) \cong \left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)^*,$$

on F_N. We have an action of the group $\mathrm{GL}\left(2, \frac{\mathbb{Z}}{N\mathbb{Z}}\right)$ on F_N that fits in the following short exact sequence.

$$1 \to \mathrm{SL}\left(2, \frac{\mathbb{Z}}{N\mathbb{Z}}\right) \to \mathrm{GL}\left(2, \frac{\mathbb{Z}}{N\mathbb{Z}}\right) \xrightarrow{\det} \left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)^* \to 1.$$

The following theorem by A. Gee and P. Stevehagen is based on the work of Shimura:

Theorem 3. *Let $\mathcal{O} = \mathbb{Z}[\theta]$ be the ring of integers of an imaginary quadratic number field K of discriminant $d < -4$. Suppose that a modular function $h \in F_N$ does not have a pole at θ and $\mathbb{Q}(j) \subset \mathbb{Q}(h)$. Denote by $x^2 + Bx + C$ the minimum polynomial of θ over $\mathbb{Q}$. Then there is a subgroup $W_{N,\theta} \subset \mathrm{GL}\left(2, \frac{\mathbb{Z}}{N\mathbb{Z}}\right)$ with elements of the form:*

$$W_{N,\theta} = \left\{ \begin{pmatrix} t - Bs & -Cs \\ s & t \end{pmatrix} \in \mathrm{GL}\left(2, \frac{\mathbb{Z}}{N\mathbb{Z}}\right) : t\theta + s \in (\mathcal{O}/N\mathcal{O})^* \right\}.$$

The function value $h(\theta)$ is a class invariant if and only if the group $W_{N,\theta}$ acts trivially on h.

Proof. [9, cor. 4].

The above theorem can be applied in order to show that a modular function gives rise to a class invariant and was used with success in order to prove that several functions were indeed class invariants. Also A. Gee and P. Stevenhagen provided us with an explicit way of describing the Galois action of $\mathrm{Cl}(\mathcal{O})$ on the class invariant so that we can construct the minimal polynomial of the ring class field.

The authors have used in [16] this technique in order to prove a claim of S. Ramanujan that the function

$$R_2(\tau) = \frac{\eta(3\tau)\eta(\tau/3 + 2/3)}{\eta^2(\tau)}$$

gives rise to class invariants. Ramanujan managed somehow (we are only left with the final result written in his notebook) to compute the first class polynomials corresponding to this class invariant and many years later, Berndt and Chan [4] proved that these first polynomials where indeed class invariants and the class polynomials written by Ramanujan were correct. We would like to notice that these Ramanujan invariants proved to be one of the most efficient invariants for the construction of prime order elliptic curves [17, 18] if one uses the CM method.

We will present now an algorithm which will allow us not only to check that a modular function is a class invariant but also to find bases of vector spaces of them. Let V be a finite dimensional vector space consisting of modular functions of level N so that $\mathrm{GL}(2,\mathbb{Z}/N\mathbb{Z})$ acts on V.

Example 1 (Generalized Weber Functions). An example of such a vector space of modular form is given by the generalized Weber functions defined as:

$$\nu_{N,0} := \sqrt{N}\frac{\eta\circ\begin{pmatrix}N & 0\\ 0 & 1\end{pmatrix}}{\eta} \text{ and } \nu_{k,N} := \frac{\eta\circ\begin{pmatrix}1 & k\\ 0 & N\end{pmatrix}}{\eta}, 0\le k\le N-1. \tag{3}$$

These are known to be modular functions of level $24N$ [10, th5. p.76]. Notice that $\sqrt{N}\in\mathbb{Q}(\zeta_N)\subset\mathbb{Q}(\zeta_{24\cdot N})$ and an explicit expression of $\sqrt{N}$ in terms of ζ_N can be given by using Gauss sums [7, 3.14 p. 228].

The group $\mathrm{SL}(2,\mathbb{Z})$ acts on the $(N+1)$-th dimensional vector space generated by them. In order to describe this action we have to describe the action of the two generators S,T of $\mathrm{SL}(2,\mathbb{Z})$ given by $S: z\mapsto -\frac{1}{z}$ and $T: z\mapsto z+1$. Keep in mind that

$$\eta\circ T(z)=\zeta_{24}\eta(z) \text{ and } \eta\circ S(z)=\zeta_8^{-1}\sqrt{iz}\eta(z).$$

We compute that (see also [10, p.77])

$$\nu_{N,0}\circ S=\nu_{0,N} \text{ and } \nu_{N,0}\circ T=\zeta_{24}^{N-1}\nu_{N,0},$$
$$\nu_{0,N}\circ S=\nu_{N,0} \text{ and } \nu_{0,N}\circ T=\zeta_{24}^{-1}\nu_{1,N},$$

for $1\le k<N-1$ and N is prime

$$\nu_{k,N}\circ S=\left(\frac{-c}{n}\right)i^{\frac{1-n}{2}}\zeta_{24}^{N(k-c)} \text{ and } \nu_{k,N}\circ T=\zeta_{24}^{-1}\nu_{k+1,N},$$

where $c=-k^{-1} \mod N$. The computation of the action of S on η is the most difficult, see [13, eq. 8 p.443].

Notice that every element $a\in\mathrm{GL}(2,\mathbb{Z}/N\mathbb{Z})$ can be written as $b\cdot\begin{pmatrix}1 & 0\\ 0 & d\end{pmatrix}$, $d\in\mathbb{Z}/N\mathbb{Z}^*$ and $b\in\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$. The group $\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$ is generated by

the elements S and T. The action of S on functions $g \in V$ is defined to be $g \circ S = g(-1/z) \in V$ and the action of T is defined $g \circ T = g(z+1) \in V$.

So in order to define the action of $\mathrm{SL}(2, \mathbb{Z}/N\mathbb{Z})$ we first use the decomposition based on Chinese remainder theorem:

$$\mathrm{GL}(2, \mathbb{Z}/N\mathbb{Z}) = \prod_{p|N} \mathrm{GL}(2, \mathbb{Z}/p^{v_p(N)}\mathbb{Z}),$$

where $v_p(N)$ denotes the power of p that appears in the decomposition in prime factors. Working with the general linear group over a field has advantages and one can use lemma 6 in [9] in order to express an element of determinant one in $\mathrm{SL}(2, \mathbb{Z}/p^{v_p(N)}\mathbb{Z})$ as word in elements S_p, T_p where S_p and T_p are 2×2 matrices which reduce to S and T modulo $p^{v_p(N)}$ and to the identity modulo $q^{v_q(N)}$ for prime divisors q of N, $p \neq q$.

This way the problem is reduced to the problem of finding the matrices S_p, T_p (this is easy using the Chinese remainder Theorem), and expressing them as products of S, T. For more details and examples, the reader is referred to the article of the second author [19].

The action of the matrix $\begin{pmatrix} 1 & 0 \\ 0 & d \end{pmatrix}$ is given by the action of the elements

$$\sigma_d \in \mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$$

on the Fourier coefficients of the expansion at the cusp at infinity [9].

4 Class Invariants and Invariant Theory

Since every element in $\mathrm{SL}(2, \mathbb{Z}/N\mathbb{Z})$ can be written as a word in S, T we obtain a function ρ

$$\left(\frac{\mathcal{O}}{N\mathcal{O}}\right)^* \xrightarrow{\phi} \mathrm{GL}(2, \mathbb{Z}/N\mathbb{Z}) \longrightarrow \mathrm{GL}(V), \qquad \rho \text{ (overall)} \tag{4}$$

where ϕ is the natural homomorphism given by Theorem 3.

The map ρ defined above is not a homomorphism but a cocycle. Indeed, if $e_1, \ldots, e_m$ is a basis of V, then the action of σ is given in matrix notation as

$$e_i \circ \sigma = \sum_{\nu=1}^{m} \rho(\sigma)_{\nu,i} e_\nu,$$

and then since $(e_i \circ \sigma) \circ \tau = e_i \circ (\sigma\tau)$ we obtain

$$e_i \circ (\sigma\tau) = \sum_{\nu,\mu=1}^{m} \rho(\sigma)^{\tau}_{\nu,i}\rho(\tau)_{\mu,\nu}e_\mu.$$

Notice that the elements $\rho(\sigma)_{\nu,i} \in \mathbb{Q}(\zeta_N)$ and $\tau \in \mathrm{GL}(2,\mathbb{Z}/N\mathbb{Z})$ acts on them as well by the element $\sigma_{\det(\tau)} \in \mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$. So we arrive at the following:

Proposition 7. *The map ρ defined in Eq. (4) satisfies the cocycle condition*

$$\rho(\sigma\tau) = \rho(\tau)\rho(\sigma)^{\tau} \tag{5}$$

and gives rise to a class in $H^1(G, \mathrm{GL}(V))$, where $G = (\mathcal{O}/N\mathcal{O})^*$. The restriction of the map ρ in the subgroup H of G defined by

$$H := \{x \in G : \det(\phi(x)) = 1\}$$

is a homomorphism.

The basis elements $e_1, \ldots e_m$ are modular functions. There is a natural notion of multiplication for them so we consider them as elements in the polynomial algebra $\mathbb{Q}(\zeta_N)[e_1, \ldots, e_m]$. The group H acts on this algebra in terms of the linear representation ρ (recall that ρ when restricted to H is a homomorphism).

Classical invariant theory provides us with effective methods (Reynolds operator method,linear algebra method [14]) in order to compute the ring of invariants $\mathbb{Q}(\zeta_N)[e_1, \ldots, e_m]^H$. Also there is a well-defined action of the quotient group $G/H \cong \mathrm{Gal}(\mathbb{Q}(\zeta_N)/\mathbb{Q})$ on $\mathbb{Q}(\zeta_N)[e_1, \ldots, e_m]^H$.

Define the vector space V_n of invariant polynomials of given degree n:

$$V_n := \{F \in \mathbb{Q}(\zeta_N)[e_1, \ldots, e_m]^H : \deg F = n\}.$$

The action of G/H on V_n gives rise to a cocycle

$$\rho' \in H^1(\mathrm{Gal}(\mathbb{Q}(\zeta_N))/\mathbb{Q}), \mathrm{GL}(V_n)).$$

The multidimensional Hilbert 90 theorem asserts that there is an element $P \in \mathrm{GL}(V_n)$ such that

$$\rho'(\sigma) = P^{-1}P^{\sigma}. \tag{6}$$

Let $v_1, \ldots, v_\ell$ be a basis of V_n. The elements v_i are by construction H invariant. The elements $w_i := v_iP^{-1}$ are G/H invariant since

$$(v_iP^{-1}) \circ \sigma = (v_i \circ \sigma)(P^{-1})^{\sigma} = v_i\rho(\sigma)(P^{-1})^{\sigma} = v_iP^{-1}P^{\sigma}(P^{-1})^{\sigma} = v_iP^{-1}.$$

The above computation together with Theorem 3 allows us to prove

Proposition 8. *Consider the polynomial ring $\mathbb{Q}(\zeta_N)[e_1, \ldots, e_m]$ and the vector space V_n of H-invariant homogenous polynomials of degree n. If P is a matrix such that Eq. (6) holds, then the images of a basis of V_n under the action of P^{-1} are class invariants.*

For computing the matrix P so that Eq. (6) holds one can use the probabilistic algorithm of Glasby-Howlett [12]. In this method one starts with the sum

$$B_Q := \sum_{\sigma \in G/H} \rho(\sigma) Q^{\sigma}. \tag{7}$$

We have to find 2×2 matrix in $\mathrm{GL}(2, \mathbb{Q}(\zeta_N))$ such that B_Q is invertible then $P := B_Q^{-1}$. Indeed, we compute that

$$B_Q^{\tau} = \sum_{\sigma \in G/H} \rho(\sigma)^{\tau} Q^{\sigma\tau}, \tag{8}$$

and the cocycle condition $\rho(\sigma\tau) = \rho(\sigma)^{\tau}\rho(\tau)$, together with Eq. (8) allows us to write:

$$B_Q^{\tau} = \sum_{\sigma \in G/H} \rho(\sigma\tau)\rho(\tau)^{-1} Q^{\sigma\tau} = B_Q \rho_{\tau}^{-1}$$

i.e.

$$\rho(\tau) = B_Q \left(B_Q^{\tau}\right)^{-1}.$$

We feed Eq. (8) with random matrices Q until B_Q is invertible. Since non invertible matrices form a Zariski closed subset in the space of matrices practice shows that we obtain an invertible B_Q almost immediately. For examples on this construction we refer to [19].

This method does not give us only some class invariants but whole vector spaces of them. For example for the space of the generalized Weber functions $\mathfrak{g}_0, \mathfrak{g}_1, \mathfrak{g}_2, \mathfrak{g}_3$ defined in the work of Gee in [10, p. 73] as

$$\mathfrak{g}_0(\tau) = \frac{\eta(\frac{\tau}{3})}{\eta(\tau)},\ \mathfrak{g}_1(\tau) = \zeta_{24}^{-1}\frac{\eta(\frac{\tau+1}{3})}{\eta(\tau)},\ \mathfrak{g}_2(\tau) = \frac{\eta(\frac{\tau+2}{3})}{\eta(\tau)},\ \mathfrak{g}_3(\tau) = \sqrt{3}\frac{\eta(3\tau)}{\eta(\tau)},$$

which are the functions defined in Example 1 for $N = 3$. We find first that the polynomials

$$I_1 := \mathfrak{g}_0\mathfrak{g}_2 + \zeta_{72}^6 \mathfrak{g}_1\mathfrak{g}_3, \qquad I_2 := \mathfrak{g}_0\mathfrak{g}_3 + (-\zeta_{72}^{18} + \zeta_{72}^6)\mathfrak{g}_1\mathfrak{g}_2$$

are indeed invariants of the action of H. Then using our method

Table 1 Minimal polynomials using the $\mathfrak{g}_0, \ldots, \mathfrak{g}_3$ functions

Invariant	Polynomial
Hilbert	$t^5 + 400497845154831586723701480652800t^4 +$
	$8185208091546130657700382653342904483 84t^3 +$
	$43982507524220948112386894195744223037268951 04t^2 -$
	$16319730975176203906274913715913862844512542392320t +$
	$15283054453672803818066421650036653646232315192410112$
e_1	$t^5 - 936t^4 - 60912t^3 - 2426112t^2 - 40310784t - 3386105856$
e_2	$t^5 - 1512t^4 - 29808t^3 + 979776t^2 + 3359232t - 423263232$

$$e_1 := (-12\zeta_{72}^{18} + 12\zeta_{72}^{6})\mathfrak{g}_0\mathfrak{g}_3 + 12\zeta_{72}^{6}\mathfrak{g}_0\mathfrak{g}_3 + 12\mathfrak{g}_1\mathfrak{g}_2 + 12\mathfrak{g}_1\mathfrak{g}_3,$$
$$e_2 := 12\zeta_{72}^{6}\mathfrak{g}_1\mathfrak{g}_2 + (-12\zeta_{72}^{18} + 12\zeta_{72}^{6})\mathfrak{g}_0\mathfrak{g}_3 + (-12\zeta_{72}^{12} + 12)\mathfrak{g}_1\mathfrak{g}_3 + 12\zeta_{72}^{12}\mathfrak{g}_1\mathfrak{g}_3$$

generate a $\mathbb{Q}$-vector space of class invariants. All $\mathbb{Q}$ linear combinations of the form $\lambda_1 e_1 + \lambda_2 e_2$ also provide class invariants. Finding the most efficient class invariant among them is a difficult problem which we hope to solve in the near future. For comparison we present in Table 1 the polynomials generating the Hilbert class field using the j invariant and the two class functions we obtained by our method.

5 Selecting the Discriminant

We have seen in the previous sections that the original version of the CM method uses a special polynomial called Hilbert class polynomial which is constructed with input a fundamental discriminant $d < 0$. A discriminant $d < 0$ is fundamental if and only if d is free of any odd square prime factors and either $-d \equiv 3 \pmod 4$ or $-d/4 \equiv 1, 2, 5, 6 \pmod 8$. The disadvantage of Hilbert class polynomials is that their coefficients grow very large as the absolute value of the discriminant $D = |d|$ increases and thus their construction requires high precision arithmetic.

According to the first main theorem of complex multiplication, the modular function $j(\theta)$ generates the Hilbert class field over K. However, the Hilbert class field can also be generated by modular functions of higher level. There are several known families of class polynomials having integer coefficients which are much smaller than the coefficients of their Hilbert counterparts. Therefore, they can substitute Hilbert class polynomials in the CM method and their use can considerably improve its efficiency. Some well-known families of class polynomials are: Weber polynomials [23], $M_{D,l}(x)$ polynomials [21], Double eta (we will denote them by $M_{D,p_1,p_2}(x)$) polynomials [6] and Ramanujan polynomials [17]. The logarithmic height of the coefficients of all these polynomials is smaller by a constant factor than the corresponding logarithmic height of the Hilbert class polynomials and this is the reason for their much more efficient construction.

A crucial question is which polynomial leads to the most efficient construction. The answer to the above question can be derived by the precision requirements of the polynomials or (in other words) the logarithmic height of their coefficients. There are asymptotic bounds which estimate with remarkable accuracy the precision requirements for the construction of the polynomials. The polynomials with the smallest (known so far) asymptotic bound are Weber polynomials constructed with discriminants d satisfying the congruence $D = |d| \equiv 7 \pmod 8$. Naturally, this leads to the conclusion that these polynomials will require less precision for their construction than all other class polynomials constructed with values D' close enough to the values of D.

In what follows, we will show that this is not really true in practice. Clearly, the degrees of class polynomials vary as a function of D, but we will see that on average these degrees are affected by the congruence of D modulo 8. In particular, we prove theoretically that class polynomials (with degree equal to their Hilbert counterparts) constructed with values $D \equiv 3 \pmod 8$ have three times smaller degree than polynomials constructed with comparable in size values of D that satisfy the congruence $D \equiv 7 \pmod 8$. Class polynomials with even discriminants (e.g., $D \equiv 0 \pmod 4$) have on average two times smaller degree than polynomials constructed with comparable in size values $D \equiv 7 \pmod 8$. This phenomenon can be generalized for congruences of D modulo larger numbers. This leads to the (surprising enough) result that there are families of polynomials which seem to have asymptotically larger precision requirements for their construction than Weber polynomials with $D \equiv 7 \pmod 8$, but they can be constructed more efficiently than them in practice (for comparable values of D).

The degree of every polynomial generating the Hilbert class field equals the class number h_D which for a fundamental discriminant $-D < 4$ is given by [22, p. 436]

$$h_D = \frac{\sqrt{D}}{2\pi} L(1, \chi) = \frac{\sqrt{D}}{2\pi} \prod_p \left(1 - \frac{\chi(p)}{p}\right)^{-1},$$

where χ is the quadratic character given by the Legendre symbol, i.e. $\chi(p) = \left(\frac{-D}{p}\right)$. Let us now consider the Euler factor

$$\left(1 - \frac{\chi(p)}{p}\right)^{-1} = \begin{cases} 1 & \text{if } p \mid D \\ \frac{p}{p-1} & \text{if } \left(\frac{-D}{p}\right) = 1 \\ \frac{p}{p+1} & \text{if } \left(\frac{-D}{p}\right) = -1. \end{cases} \tag{9}$$

Observe that smaller primes have a bigger influence on the value of h_D. For example, if $p = 2$, then we compute

$$\left(1 - \frac{\chi(2)}{2}\right)^{-1} = \begin{cases} 1 & \text{if } 2 \mid D \\ 2 & \text{if } D \equiv 7 \pmod 8 \\ \frac{2}{3} & \text{if } D \equiv 3 \pmod 8. \end{cases} \tag{10}$$

This leads us to the conclusion that on average the degree of a class polynomial with $D \equiv 3 \pmod 8$ will have three times smaller degree than a polynomial constructed with a comparable value of $D \equiv 7 \pmod 8$. Similarly, the degree of a polynomial constructed with even values of $D \equiv 0 \pmod 4$ will have on average two times smaller degree than a polynomial with $D \equiv 7 \pmod 8$.

Going back to Eq. (9), we can see that for discriminants of the same congruence modulo 8, we can proceed to the next prime $p = 3$ and compute

$$\left(1 - \frac{\chi(3)}{3}\right)^{-1} = \begin{cases} 1 & \text{if } 3 \mid D \\ \frac{3}{2} & \text{if } \left(\frac{-D}{3}\right) = 1 \\ \frac{3}{4} & \text{if } \left(\frac{-D}{3}\right) = -1. \end{cases}$$

This means that for values of D such that $\left(\frac{-D}{3}\right) = -1$ the value of h_D is on average two times smaller than class numbers corresponding to values with $\left(\frac{-D}{3}\right) = 1$. Consider for example, the cases $D \equiv 3 \pmod 8$ and $D \equiv 7 \pmod 8$. If we now include in our analysis the prime $p = 3$, then we can distinguish 6 different subcases $D \equiv 3, 11, 19 \pmod{24}$ and $D \equiv 7, 15, 23 \pmod{24}$. Having in mind the values $\left(1 - \frac{\chi(2)}{2}\right)^{-1}$ and $\left(1 - \frac{\chi(3)}{3}\right)^{-1}$, we can easily see, for example, that the polynomials with $D \equiv 19 \pmod{24}$ will have on average 6 times smaller degrees than the polynomials with $D \equiv 23 \pmod{24}$.

What happens if we continue selecting larger primes p? Equation (9) implies that if we select a discriminant $-D$ such that for all primes $p < N$ we have $\left(\frac{-D}{p}\right) = -1$ then the class number corresponding to D has a ratio that differs from other discriminants by a factor of at most

$$\prod_{p<N} \left(\frac{p-1}{p+1}\right) = \prod_{p<N} \left(1 - \frac{2}{p+1}\right). \tag{11}$$

Since the series $\sum_p \frac{2}{p+1}$ diverges (p runs over the prime numbers), the product in Eq. (11) diverges as well [1, p.192 th. 5]. Therefore, the product in Eq. (11) can have arbitrarily high values for sufficiently large values of N. This also means that if D is sufficiently big we can choose discriminants that correspond to class numbers that have an arbitrarily high ratio with respect to other discriminants of the same size.

6 Conclusions

In this paper, we have given a detailed overview of the CM method for the construction of elliptic curves. We have presented the necessary theoretical background and we have described our published results on finding new class invariants using the Shimura reciprocity law. The proper selection of a suitable discriminant D for the construction of class polynomials, combined with the above results, will hopefully lead us to more efficient constructions in the future using new families of class polynomials.

References

1. Ahlfors, L.V.: Complex Analysis. An introduction to the Theory of Analytic Functions of One Complex Variable, 3rd edn. International Series in Pure and Applied Mathematics. McGraw-Hill, New York (1978)
2. Blake, I.F., Seroussi, G., Smart, N.P.: Elliptic Curves in Cryptography London Mathematical Society Lecture Note Series, vol. 165, New York (1999)
3. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. J. Symb. Comput. **24**(3–4), 235–265 (1997)
4. Bruce, C., Berndt, H., Huat, C.: Ramanujan and the modular j-invariant. Can. Math. Bull. **42**(4), 427–440 (1999). MR MR1727340 (2002a:11035)
5. David, A.C.: Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory and Complex Multiplication. Wiley, New York (1989). MR MR1028322 (90m:11016)
6. Enge, A., Schertz, R.: Constructing elliptic curves over finite fields using double eta-quotients. J. Théor. Nombres Bordeaux **16**, 555–568 (2004). (MR2144957)
7. Fröhlich, A., Taylor, M.J.: Algebraic Number Theory. Cambridge Studies in Advanced Mathematics, vol. 27. Cambridge University Press, Cambridge (1993). xiv+355 pp. ISBN: 0-521-43834-9
8. Gauss, C.F.: Disquisitiones Arithmeticae. Traducida por Arthur A. Clarke. Yale University Press, New Haven and London (1966)
9. Gee, A.: Class invariants by Shimura's reciprocity law, J. Théor. Nombres Bordeaux **11**(1), 45–72 (1999) Les XXèmes Journées Arithmétiques (Limoges, 1997). MR MR1730432 (2000i:11171)
10. Gee, A.: Class fields by Shimura reciprocity, Ph.D. thesis, Leiden University available online at http://www.math.leidenuniv.nl/nl/theses/44 (2001)
11. Gee, A., Stevenhagen, P.: Generating class fields using Shimura reciprocity. In: Buhler, J.P. (ed.) Algorithmic Number Theory (Portland, OR, 1998). Lecture Notes in Computer Science, vol. 1423, pp. 441–453. Springer, Berlin (1998). MR MR1726092 (2000m:11112)
12. Glasby, S.P., Howlett, R.B.: Writting representatations over minimal fields. Commun. Algebra **25**(6), 1703–1711 (1997)
13. Hart, W.B.: Schläfli modular equations for generalized Weber functions. Ramanujan J. **15**(3), 435–468 (2008)
14. Kemper, G., Steel, A.: Some algorithms in invariant theory of finite groups. In: Dräxler, P., Michler, G.O., Ringel, C.M. (eds.) Computational Methods for Representations of Groups and Algebras, Euroconference in Essen. Progress in Mathematics, vol. 173. Birkhäuser, Basel (1997)
15. Konstantinou, E., Kontogeorgis, A., Stamatiou, Y.C., Zaroliagis, C.: Generating prime order elliptic curves: difficulties and efficiency considerations. In: International Conference on Information Security and Cryptology – ICISC 2004. Lecture Notes in Computer Science, vol. 3506, pp. 261–278. Springer, Berlin (2005)
16. Konstantinou, E., Kontogeorgis, A.: Computing polynomials of the Ramanujan t_n class invariants. Can. Math. Bull. **52**(4), 583–597 (2009). MR MR2567152
17. Konstantinou, E., Kontogeorgis, A.: Introducing Ramanujan's class polynomials in the generation of prime order elliptic curves. Comput. Math. Appl. **59**(8), 2901–2917 (2010)
18. Konstantinou, E., Kontogeorgis, A.: Ramanujan invariants for discriminants equivalent to 5 mod 24. Int. J. Number Theory **8**(1), 265–287
19. Kontogeorgis, A.: Constructing class invariants. Math. Comput. **83**(287), 1477–1488 (2014)
20. Lay, G.J., Zimmer, H.G.: Constructing elliptic curves with given group order over large finite fields. In: Algorithmic Number Theory Symposium I. Springer Lecture Notes in Computer Science. Springer, Berlin (1994)
21. Morain, F.: Modular curves and class. LMS Durham Symposium on Computational Number Theory (2000)

22. Narkiewicz, W.: Elementary and Analytic Theory of Algebraic Numbers, 2nd edn. Springer, Berlin (1990)
23. Schertz, R.: Weber's class invariants revisited. J. Théor. Nombres Bordeaux **4**, 325–343 (2002). (MR1926005)
24. Shimura G.: Introduction to the Arithmetic Theory of Automorphic Functions. Publications of the Mathematical Society of Japan, vol. 11, Princeton University Press, Princeton, NJ (1994). Reprint of the 1971 original, Kano Memorial Lectures, 1. MR MR1291394 (95e:11048)
25. Silverman, J.: The Arithmetic of Elliptic Curves. Graduate Texts in Mathematics, vol. 106. Springer, New York (1986)
26. Stevenhagen, P.: Hilbert's 12th problem, complex multiplication and Shimura reciprocity. In: Class Field Theory—Its Centenary and Prospect (Tokyo, 1998). Advanced Studies in Pure Mathematics, vol. 30, pp. 161–176, Mathematical Society of Japan, Tokyo (2001). MR MR 18464571 (2002i:11110)
27. Weber, H.: Lehrbuch der Algebra, Band III, 2nd edition, Chelsea reprint, original edition 1908
28. Yui, N., Zagier, Don.: On the singular values of Weber modular functions. Math. Comput. Am. Math. Soc. **66**(220), 1645–1662 (1997). MR MR1415803 (99i:11046)

Generalized Laplace Transform Inequalities in Multiple Weighted Orlicz Spaces

Jichang Kuang

Abstract In this paper we use quite different new methods to establish some new generalized Laplace transform inequalities in the multiple weighted Orlicz spaces. They are significant improvements and generalizations of many famous results.

Keywords: Laplace transform inequality • Weighted Orlicz space • Norm inequality

1 Introduction

Given a function f on $(0,\infty)$ such that $\mathrm{e}^{-\alpha y}|f(y)|$ is integrable over the interval $(0,\infty)$ for some real α, we define $F(z)$ as

$$F(z)=\int_0^\infty \mathrm{e}^{-zy}f(y)dy, \tag{1}$$

where we require that $Re(z)>\alpha$ so that the integral in (1) converges. F is called the (one-sided) Laplace transform of f. We consider only one-sided Laplace transforms with the real parameter x, that is,

$$F(x)=\int_0^\infty \mathrm{e}^{-xy}f(y)dy, x>\alpha, \tag{2}$$

as these play the most important role in the solution of initial and boundary value problems for partial differential equations (see [1–3]). We use the standard notations:

$$L^p_\omega(0,\infty)=\{f:\|f\|_{p,\omega}<\infty\}, \|f\|_{p,\omega}=\left(\int_0^\infty |f(x)|^p\omega(x)dx\right)^{1/p}.$$

J. Kuang (✉)
Department of Mathematics, Hunan Normal University, Changsha, Hunan 410081, People's Republic of China
e-mail: jckuang@163.com

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_13

If $\omega(x) \equiv 1$, we will denote $\|f\|_{p,1}$by $\|f\|_p$. $\Gamma(\alpha)$ is the Gamma function:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \mathrm{e}^x dx \quad (\alpha > 0).$$

In asymptotic analysis , we often study functions f whose has $N+1$ continuous derivatives while $f^{(N+2)}$ is piecewise continuous on $(0, \infty)$, then by ([4]), we have

$$F(x) = \int_0^\infty e^{-xy} f(y) dy \sim \sum_{k=0}^{N} x^{-(k+1)} f^k(0)$$

represents an asymptotic expansion of F, as $x \to \infty$, to $N+1$ terms. But the following Hardy's result of being neglected ([5]):

Theorem 1. *If $f \in L^p_\omega(0, \infty), 1 < p < \infty, \omega(x) = x^{p-2}$, then F is defined by (2) satisfies*

$$\|F\|_p \le \Gamma(1/p) \|f\|_{p,\omega}.$$

It is important to note that x appears in (2) only through the product xy, this suggests that, as a generalization, we might consider the wider class of integral operators

$$T(f, x) = \int_{\mathbb{R}^n_+} K(\|x\|^{\lambda_1} \cdot \|y\|^{\lambda_2}) f(y) dy, \tag{3}$$

where $x \in \mathbb{R}^n_+ = \{(x_1, \ldots, x_n) : x_k \ge 0, 1 \le k \le n\}$, $\|x\| = (\sum_{k=1}^n |x_k|^2)^{1/2}$, λ_1, λ_2 are real numbers and $\lambda_1 \times \lambda_2 \ne 0$. In particular, if $n = 1, \lambda_1 = \lambda_2 = 1$, then

$$T(f, x) = \int_0^\infty K(xy) f(y) dy, x \in \mathbb{R}^1_+. \tag{4}$$

In this case, we say that $T(f)$ is the generalized Laplace transform of f. In [5], Hardy proved the following two theorems:

Theorem 2. *Let K be a non-negative and measurable function on $(0, \infty)$. If f be a non-negative and not null, $1 < p < \infty$, then the integral operator T is defined by (4): $T : L^p_\omega(0, \infty) \to L^p(0, \infty)$ exists as a bounded operator and*

$$\|Tf\|_p \le C(1/p) \|f\|_{p,\omega}, \tag{5}$$

where $\omega(x) = x^{p-2}$ and

$$C(1/p) = \int_0^\infty K(t) t^{(1/p)-1} dt. \tag{6}$$

This is an immediate consequences of the following Theorem 3:

Theorem 3. *Let K be a non-negative and measurable function on $(0,\infty)$. Let f, g are non-negative and neither f nor g is null,* $1 < p < \infty, \frac{1}{p} + \frac{1}{q} = 1$. *If* $f \in L^p_\omega(0,\infty), g \in L^q(0,\infty)$, *then*

$$\int_0^\infty \int_0^\infty K(xy)f(x)g(y)dxdy < C(1/p)\|f\|_{p,\omega}\|g\|_q, \tag{7}$$

where $\omega(x) = x^{p-2}$ *and* $C(1/p)$ *is defined by (6). If the measurable function* $p : \mathbb{R}^n \to [1,\infty)$ *as exponential function, by* $L^{p(x)}(\mathbb{R}^n)$ *we denote the Banach function space of the measurable function* $f : \mathbb{R}^n \to \mathbb{R}^1$ *such that*

$$\|f\|_{L^{p(\cdot)}(\mathbb{R}^n)} = \inf\{\lambda > 0 : \int_{\mathbb{R}^n} |\frac{f(x)}{\lambda}|^{p(x)}dx \le 1\} < \infty. \tag{8}$$

For the basic properties of spaces $L^{p(x)}(\mathbb{R}^n)$, we refer to [6–11]. The variable exponent Lebesgue spaces $L^{p(x)}(\mathbb{R}^n)$ and the corresponding variable Sobolev spaces $W^{k,p(\cdot)}(\mathbb{R}^n)$ are of interest for their applications to modeling problems in physics and to the study of variational integrals and partial differential equations with nonstandard growth condition. It is well-known that the Orlicz spaces are the generalizations of L^p spaces and play an important role in mathematical physics. In 2008, Kuang and Debnath obtained in [15] the Hilbert's inequalities with the homogeneous kernel on the one-dimensional weighted Orlicz spaces. For the basic results of the one- dimensional Orlicz spaces, we refer to [13, 14]. In 2014, the author [16] introduce the new multiple weighted Orlicz spaces, they are generalizations of the variable exponent Lebesgue spaces $L^{p(x)}(\mathbb{R}^n_+)$. The main aim of this paper is to establish some new generalized Laplace transform inequalities in the new multiple weighted Orlicz spaces. Here we use quite different methods and techniques. They are significant improvements and generalizations of many famous results.

2 Definitions and Statement of the Main Results

In what follows, we write

$$\|f\|_{p,\omega} = \left(\int_{\mathbb{R}^n_+} |f(x)|^p \omega(x)dx\right)^{1/p}, L^p(\omega) = \{f : f \text{ is measurable}, \ ||f||_{p,\omega} < \infty\}.$$

Definition 1 (See [13, 14]). We call φ a Young's function if it is a non-negative increasing convex function on $(0,\infty)$ with $\varphi(0) = 0, \varphi(u) > 0, u > 0$, and

$$\lim_{u\to 0}\frac{\varphi(u)}{u}=0,\ \lim_{u\to\infty}\frac{\varphi(u)}{u}=\infty.$$

To Young's function φ we can associate its convex conjugate function denoted by $\psi=\varphi^*$ and defined by

$$\psi(v)=\varphi^*(v)=\sup\{uv-\varphi(u):u\geq 0\},v\geq 0.$$

We note that $\psi=\varphi^*$ is also a Young's function and $\psi^*=(\varphi^*)^*=\varphi$. From the definition of $\psi=\varphi^*$, we get Young's inequality

$$uv\leq\varphi(u)+\psi(v),u,v>0. \tag{9}$$

Let φ^{-1} be inverse function of φ, we have

$$v\leq\varphi^{-1}(v)\psi^{-1}(v)\leq 2v,v\geq 0. \tag{10}$$

Definition 2 (See [16]). Let φ be a Young's function on $(0,\infty)$, for any measurable function f and non-negative weight function ω on $\mathbb{R}^n_+$, the multiple weighted Luxemburg norm is defined as follows:

$$\|f\|_{\varphi,\omega}=\inf\{\lambda>0:\int_{\mathbb{R}^n_+}\varphi(\frac{|f(x)|}{\lambda})\omega(x)dx\leq 1\}. \tag{11}$$

The multiple weighted Orlicz space is defined as follows:

$$L_\varphi(\omega)=\{f:\|f\|_{\varphi,\omega}<\infty\}. \tag{12}$$

In particular, if $\varphi(u)=u^{p(x)}$, then $L_\varphi(\omega)$ is the weighted variable exponent Lebesgue spaces $L^{p(\cdot)}(\omega)$; if the exponents $p(x),q(x)$ are constant, for example, $\varphi(u)=u^p,1<p<\infty$, then $L_\varphi(\omega)$ is the weighted Lebesgue spaces $L^p(\omega)$ on R^n_+; if $\varphi(u)=u(\log(u+c))^q,q\geq 0,c>o$, then $L_\varphi(\omega)$ is the weighted spaces $L(\omega)(\log L(\omega))^q$ on R^n_+.

Definition 3 (See [13]). We call the Young's function φ on $(0,\infty)$ sub-multiplicative, if

$$\varphi(uv)\leq\varphi(u)\varphi(v) \tag{13}$$

for all $u,v\geq 0$.

Remark 1. If φ satisfies (13), then φ also satisfies Orlicz ∇_2- condition, that is, there exists a constant $C>1$ such that

$$\varphi(2u)\leq C\varphi(u)$$

for all $u\geq 0$.

Our main result is the following theorem:

Theorem 4. *Let K be a non-negative and measurable function on $R_+^n \times R_+^n$. Let the conjugate Young's functions φ, ψ on $(0,\infty)$ sub-multiplicative, and*

$$\omega_1(x) = \|x\|^{-\frac{(n\lambda_1)}{\lambda_2}}, \omega_2(y) = \|y\|^{-\frac{(n\lambda_2)}{\lambda_1}},$$

where λ_1, λ_2 are real numbers and $\lambda_1 \times \lambda_2 \neq 0$. Let $f \in L_\varphi(\omega_1)$, $g \in L_\psi(\omega_2)$ and $\|f\|_{\varphi,\omega_1} > 0, \|g\|_{\psi,\omega_2} > 0$. If

$$C_1 = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)\lambda_2} \times \int_0^\infty K(u)\psi^{-1}(u)u^{(\frac{n}{\lambda_2})-1}du < \infty; \tag{14}$$

$$C_2 = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)\lambda_1} \times \int_0^\infty K(u)\psi\left(\frac{1}{\varphi^{-1}(\psi^{-1}(u))}\right) \times u^{(\frac{n}{\lambda_1})-1}du < \infty, \tag{15}$$

then

$$\int_{\mathbb{R}_+^n}\int_{\mathbb{R}_+^n} K(\|x\|^{\lambda_1} \cdot \|y\|^{\lambda_2}) \times f(x)g(y)dxdy \leq C(\varphi,\psi)\|f\|_{\varphi,\omega_1}\|g\|_{\psi,\omega_2}, \tag{16}$$

where $C(\varphi,\psi) = C_1 + C_2$ is defined by (14) and (15).

We obtain the following Corollary 1 by taking $\varphi(u) = u^{p(x)}x$, $\psi(v) = v^{q(x)}$ in Theorem 4, where $1 < p(x) < \infty$, $\frac{1}{p(x)} + \frac{1}{q(x)} = 1, x \in R_+^n$, and $p_- = \text{essinf}\{p(x) : x \in R_+^n\}, p_+ = \text{esssup}\{p(x) : x \in R_+^n\}, 1 < p_- \leq p_+ < \infty$.

Corollary 1. *Let $K, \lambda_1, \lambda_2, \omega_1$, and ω_2 satisfy the conditions of Theorem 4. If $f \in L^{p(\cdot)}(\omega_1)$, $g \in L^{q(\cdot)}(\omega_2)$, then*

$$\int_{R_+^n}\int_{R_+^n} K(\|x\|^{\lambda_1} \cdot \|y\|^{\lambda_2})f(x)g(y)dxdy \leq c(p,q)\|f\|_{p(\cdot),\omega_1}\|g\|_{q(\cdot),\omega_2}, \tag{17}$$

where

$$c(p,q) = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)}\left\{\frac{1}{\lambda_2} \times \int_0^1 K(u)u^{(\frac{1}{q_+}+\frac{n}{\lambda_2}-1)}du + \int_1^\infty K(u)u^{(\frac{1}{q_-}-\frac{n}{\lambda_2}-1)}du)\right.$$
$$\left. + \frac{1}{\lambda_1}\left(\int_0^1 K(u)u^{(-\frac{1}{p_-}+\frac{n}{\lambda_1}-1)}du + \int_1^\infty K(u)u^{(\frac{1}{p_+}+\frac{n}{\lambda_1}-1)}du\right)\right\}. \tag{18}$$

In particular, if $n = 1$, in Corollary 1, then

$$\int_0^\infty\int_0^\infty K(x^{\lambda_1}y^{\lambda_2})f(x)g(y)dxdy \leq c(p,q)\|f\|_{p(\cdot),\omega_1}\|g\|_{q(\cdot),\omega_2}, \tag{19}$$

where $\omega_1(x) = x^{-\frac{\lambda_1}{\lambda_2}}$, $\omega_2(y) = y^{-\frac{\lambda_2}{\lambda_1}}$, and

$$C(p,q) = \frac{1}{\lambda_2}\left(\int_0^1 K(u)u^{(\frac{1}{q_+}-\frac{1}{\lambda_2}-1)}du + \int_1^\infty k(u)u^{(\frac{1}{q_-}-\frac{1}{\lambda_2}-1)}du\right) + \frac{1}{\lambda_1}\left(\int_0^1 K(u)u^{(-\frac{1}{p_-}+\frac{1}{\lambda_1}-1)}du + \int_1^\infty k(u)u^{(-\frac{1}{p_+}+\frac{1}{\lambda_1}-1)}du\right). \quad (20)$$

We obtain the following Corollary 2 by taking $\varphi(u) = u^p$, $\psi(v) = v^q$, $1 < p, q < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, in Theorem 4:

Corollary 2. *Let* $K, \lambda_1, \lambda_2, \omega_1$, *and* ω_2 *satisfy the conditions of Theorem 4. If* $f \in L^p(\omega_1)$, $g \in L^q(\omega_2)$, $1 < p < \infty$, $\frac{1}{p} + \frac{1}{q} = 1$, *then*

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+} K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})f(x)g(y)dxdy \le C(p,q)\|f\|_{p,\omega_1}\|g\|_{q,\omega_2}, \quad (21)$$

where

$$C(p,q) = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)}\left\{\frac{1}{\lambda_2}\times\int_0^\infty K(u)u^{(\frac{1}{q}+\frac{1}{\lambda_2}-1)}du + \frac{1}{\lambda_1}\int_0^\infty K(u)u^{(-\frac{1}{p}+\frac{n}{\lambda_1}-1)}du\right\} \quad (22)$$

In particular, if $\lambda_1 = \lambda_2 = 1$, in Corollary 2, then

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+} K(\|x\|\cdot\|y\|)f(x)g(y)dxdy \le C(p,q)\|f\|_{p,\omega}\|g\|_{q,\omega}, \quad (23)$$

where $\omega(x) = \|x\|^{-n}$, and

$$C(p,q) = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)}\left\{\int_0^\infty K(u)u^{(n-\frac{1}{p})}du\int_0^\infty K(u)u^{(-\frac{1}{p}+n-1)}du\right\}. \quad (24)$$

If $n = 1$ in (23), then

$$\int_0^\infty\int_0^\infty K(xy)f(x)g(y)dxdy \le \left(\int_0^\infty K(u)(u^{1/q}+u^{-(1/p)})du\right)\|f\|_{p,\omega}\|g\|_{q,\omega}, \quad (25)$$

where $\omega(x) = x^{-1}$.

Remark 2. Take $K(u) = \mathrm{e}^{-u}$ in (21), if $\lambda_1 \in (0, pn)$, $\lambda_2 \in (-\infty, -qn)\bigcup(0,\infty)$, then

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+} \mathrm{e}^{-(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})} f(x)g(y)dxdy \le C(p,q)\|f\|_{p,\omega_1}\|g\|_{q,\omega_2}, \tag{26}$$

where

$$C(p,q) = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)} \times \left\{ \frac{1}{\lambda_2}\Gamma\left(\frac{1}{q}+\frac{n}{\lambda_2}\right) + \frac{1}{\lambda_1}\Gamma\left(\frac{n}{\lambda_1}-\frac{1}{p}\right)\right\}. \tag{27}$$

In particular, if $\lambda_1 = \lambda_2 = 1$, then

$$C(p,q) = \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)} \times \left\{ \Gamma\left(\frac{1}{q}+n\right) + \Gamma\left(n-\frac{1}{p}\right)\right\}. \tag{28}$$

So that if taking $K(u) = \mathrm{e}^{-u}$ in (25), we get

$$\int_0^\infty\int_0^\infty \mathrm{e}^{-xy} f(x)g(y)dxdy \le \left(\Gamma\left(\frac{1}{q}\right) + \Gamma\left(1+\frac{1}{q}\right)\right)\|f\|_{p,\omega}\|g\|_{q,\omega}. \tag{29}$$

Defining other forms of K, we can obtain new results of interest.

3 Proof of Theorem 4

We require the following lemmas to prove our result:

Lemma 1 (See [14, 18]). *If $a_k, b_k, p_k > 0, 1 \le k \le n$, f be a measurable function on [0,1]. Let $D = \{(x_1, x_2, \ldots, x_n) : \sum_{k=1}^n (\frac{x_k}{a_k})^{b_k} \le 1, x_k \ge 0\}$, then*

$$\int_D f\left(\sum_{k=1}^n \left(\frac{x_k}{a_k}\right)^{b_k}\right) x_1^{p_1-1}\cdots x_n^{p_n-1}dx_1\ldots dx_n$$

$$= \frac{\prod_{k=1}^n a_k^{p_k}}{\prod_{k=1}^n b_k} \times \frac{\prod_{k=1}^n \Gamma\left(\frac{p_k}{b_k}\right)}{\Gamma\left(\sum_{k=1}^n \frac{p_k}{b_k}\right)} \times \int_0^1 f(t)t^{(\sum_{k=1}^n \frac{p_k}{b_k}-1)}dt. \tag{30}$$

Let $E = \{(x_1, x_2, \ldots, x_n) : \sum_{k=1}^n (\frac{x_k}{a_k})^{b_k} \ge 1, x_k \ge 0\}$, then

$$\int_E f\left(\sum_{k=1}^n \left(\frac{x_k}{a_k}\right)^{b_k}\right) x_1^{p_1-1}\ldots x_n^{p_n-1}dx_1\ldots dx_n$$

$$= \frac{\prod_{k=1}^n a_k^{p_k}}{\prod_{k=1}^n b_k} \times \frac{\prod_{k=1}^n \Gamma\left(\frac{p_k}{b_k}\right)}{\Gamma\left(\sum_{k=1}^n \frac{p_k}{b_k}\right)} \times \int_1^\infty f(t)t^{(\sum_{k=1}^n \frac{p_k}{b_k}-1)}dt. \tag{31}$$

From (30) and (31), we have the following lemma:

Lemma 2. *Let f be a measurable function on $[0,\infty)$, then*

$$\int_{\mathbb{R}^n_+} f(\|x\|^2)dx = \frac{\pi^{n/2}}{2^n\Gamma(n/2)}\int_0^\infty f(t)t^{(n/2)-1}dt, \tag{32}$$

where $\|x\| = (\sum_{k=1}^n |x_k|^2)^{1/2}$.

Proof of Theorem 4. Applying (10) and Young's inequality (9), we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+} K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times f(x)g(y)dxdy\\
&\quad\le \int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\{|f(x)|\varphi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}\{|g(y)|\psi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}dxdy\\
&\quad= \int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\{|f(x)|\varphi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}\\
&\qquad\times\{|g(y)|\psi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\}dxdy\\
&\quad\le \int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\varphi\{|f(x)|\varphi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}dxdy\\
&\qquad+\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\psi\{|g(y)|\psi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\}dxdy\\
&\quad= I_1+I_2.
\end{aligned}\tag{33}$$

Since φ on $(0,\infty)$ is sub-multiplicative, we have

$$\begin{aligned}
&\varphi\{|f(x)|\varphi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\times\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}\\
&\quad\le\varphi(|f(x)|)\varphi\{\varphi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\times\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\}\\
&\quad\le\varphi(|f(x)|)K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}).
\end{aligned}\tag{34}$$

Then, we have

$$\begin{aligned}
I_1&\le\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\varphi(|f(x)|)\times K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})dxdy\\
&=\int_{\mathbb{R}^n_+}\varphi(|f(x)|)\times\left\{\int_{\mathbb{R}^n_+}K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})dy\right\}dx.
\end{aligned}\tag{35}$$

By (32), we have

$$\int_{\mathbb{R}^n_+} K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})dy$$
$$=\frac{\pi^{n/2}}{2^n\Gamma(n/2)}\int_0^\infty K(\|x\|^{\lambda_1}t^{\frac{\lambda_2}{2}})\times\psi^{-1}(\|x\|^{\lambda_1}\cdot t^{\frac{\lambda_2}{2}})t^{(n/2)-1}dt. \tag{36}$$

Let $u=\|x\|^{\lambda_1}\cdot t^{\frac{\lambda_2}{2}}$, and by (35), (36), and (14), we get

$$I_1\le\frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)\lambda_2}\times\int_{\mathbb{R}^n_+}\int_0^\infty\varphi(|f(x)|)\times\|x\|^{-\frac{n\lambda_1}{\lambda_2}}\times K(u)\psi^{-1}(u)u^{\frac{n}{\lambda_2}-1}dudx$$
$$=\frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)\lambda_2}\times\left\{\int_0^\infty K(u)\psi^{-1}(u)u^{\frac{n}{\lambda_2}-1}du\right\}\times\left\{\int_{\mathbb{R}^n_+}\varphi(|f(x)|)\|x\|^{(-\frac{n\lambda_1}{\lambda_2})}dx\right\}$$
$$=C_1\int_{\mathbb{R}^n_+}\varphi(|f(x)|)\omega_1(x)dx. \tag{37}$$

Similarly, we have

$$\psi\left\{|g(y)|\psi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))\times\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\right\}$$
$$\le\psi(|g(y)|)K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi\left\{\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\right\}. \tag{38}$$

By (32), we have

$$\int_{\mathbb{R}^n_+}K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\psi\left\{\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\right\}dx$$
$$=\frac{\pi^{n/2}}{2^n\Gamma(n/2)}\int_0^\infty K(t^{\frac{\lambda_1}{2}}\|y\|^{\lambda_2})\times\psi\left\{\frac{1}{\varphi^{-1}(\psi^{-1}(t^{\frac{\lambda_1}{2}}\|y\|^{\lambda_2}))}\right\}\times t^{\frac{n}{2}-1}dt. \tag{39}$$

Let $u=t^{(\frac{\lambda_1}{2})}\cdot\|y\|^{\lambda_2}$, and by (38), (39), (32) and (15) we get

$$I_2=\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}\psi\{|g(y)|\times\psi^{-1}(K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))$$
$$\times\frac{1}{\varphi^{-1}(\psi^{-1}(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2}))}\}dxdy$$

$$= \frac{\pi^{n/2}}{2^{n-1}\Gamma(n/2)\lambda_1}\left\{\int_0^\infty K(u)\times\psi\left\{\frac{1}{\varphi^{-1}(\psi^{-1}(u))}\right\}u^{(\frac{n}{\lambda_1})-1}du\right\}$$
$$\times\int_{\mathbb{R}^n_+}\psi(|g(y)|)\|y\|^{(-\frac{n\lambda_2}{\lambda_1})}dy$$
$$= C_2\int_{\mathbb{R}^n_+}\psi(|g(y)|)\omega_2(y)dy. \tag{40}$$

Thus, by (37) and (40), we obtain

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times f(x)g(y)dxdy$$
$$\le C_1\int_{\mathbb{R}^n_+}\varphi(|f(x)|)\omega_1(x)dx + C_2\int_{\mathbb{R}^n_+}\psi(|g(y)|)\omega_2(y)dy. \tag{41}$$

It follows that

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times\left(\frac{f(x)}{\|f\|_{\varphi,\omega_1}}\right)\left(\frac{g(y)}{\|g\|_{\psi,\omega_2}}\right)dxdy$$
$$\le C_1\int_{\mathbb{R}^n_+}\varphi\left(\frac{|f(x)|}{\|f\|_{\varphi,\omega_1}}\right)\omega_1(x)dx + C_2\int_{\mathbb{R}^n_+}\psi\left(\frac{|g(y)|}{\|g\|_{\psi,\omega_2}}\right)\omega_2(y)dy$$
$$\le C_1 + C_2 = C(\varphi,\psi).$$

Hence,

$$\int_{\mathbb{R}^n_+}\int_{\mathbb{R}^n_+}K(\|x\|^{\lambda_1}\cdot\|y\|^{\lambda_2})\times f(x)g(y)dxdy \le C(\varphi,\psi)\|f\|_{\varphi,\omega_1}\|g\|_{\psi,\omega_2}.$$

The proof is complete.

Acknowledgements *Foundation item:* This work is supported by the Natural Science Foundation of China (No. 11271123).

References

1. Widder, D.V.: The Laplace Transform. Princeton University Press, Princeton (1972)
2. Wolff, K.B.: Integral Transforms in Science and Engineering. Plenum, New York (1979)
3. Zauderer, E.: Partial Differential Equations of Applied Mathematics. A Wiley-Interscience. Wiley, New York (1983)
4. Bleistein, N., Handelsman, R.A.: Asymptotic Expansions of Integrals. Dover, New York (1986)
5. Hardy, G.H.: The constants of certain inequalities. J. Lond. Math. Soc. **8**, 114–119 (1933)

6. Kovacik, O., Rakosnik, J.: On spaces $L^{p(x)}$ and $W^{1,p(x)}$. Czechoslo. Math. J. **41**(116), 592–618 (1991)
7. Fhan, X.L., Zhao, D.: On the spaces $L^{p(x)}$ and $W^{k,p(x)}$. J. Math. Anal. Appl. **263**, 424–446 (2001)
8. Edmunds, D.E., Kokilashvili, V., Meskhi, A.: On the boundedness and compactness of the weighted Hardy operators in $L^{p(x)}$ spaces. Georgian Math. J. **12**(1), 126–130 (2005)
9. Diening, L., Samko, S.: Hardy inequality in variable eponent Lebesgue spaces. Fract. Calc. Appl. Anal. **10**(1), 1–17 (2007)
10. Kopaliam, T.: Interpolation theorems for variable exponent Lebesgue space. J. Funct. Anal. **257**, 3541–3551 (2009)
11. Mamedov, F.I., Harman, A.: On a weighted inequality of Hardy type in spaces $L^{p(\cdot)}$. J. Math. Anal. Appl. **353**, 521–530 (2009)
12. Jichang, K., Debnath, L.: On Hilbert's type inequalities on the weighted Orlicz spaces. Pac. J. Appl. Math. **1**(1), 89–97 (2008)
13. Maligranda, L.: Orlicz spaces and interpolation. IMECC (1989)
14. Strömberg, J.O.: Bounded mean oscillations with Orlicz norms and duality of Hardy spaces. Indiana Univ. Math. J. **28**(3), 511–544 (1979)
15. Zwillinger, D.: Handbook of Integration. Springer, New York (1992)
16. Kuang, J.C.: Generalized Hardy-Hilbert type inequalities on multiple weighted Orlicz spaces. In: Handbook of Functional Equations: Functional Inequalities. Springer Optimizations and Its Applications, vol. 95, pp. 273–280. Springer, Berlin (2014)
17. Jichang, K.: Applied Inequalities, 4th edn. Shandong Science and Technology Press, Jinan (2010) (in Chinese)
18. Kuang, J.C.: Real and Functional Analysis (Continuation), vol. 2. Higher Education Press, Beijing (2015) (in Chinese)

Threshold Secret Sharing Through Multivariate Birkhoff Interpolation

Vasileios E. Markoutis, Gerasimos C. Meletiou, Aphrodite N. Veneti, and Michael N. Vrahatis

Abstract Secret sharing schemes have been well studied and widely used in different aspects of real life applications. The original secret sharing scheme was proposed by Adi Shamir in 1979. A similar scheme was also invented independently in the same year by George Blakley. Shamir's scheme is based on Lagrange interpolation while Blakley's approach uses principles of hyperplane geometry. In 2007, Tamir Tassa proposed a hierarchical secret sharing scheme through univariate Birkhoff interpolation (a generalization of Lagrangian and Hermitian interpolation). In the contribution at hand we investigate the idea of generalizing Tassa's scheme through multivariate Birkhoff interpolation. We consider the problem of finding secret sharing schemes with multilevel structures and partially ordered sets of levels of participants. In order to ensure that our scheme meets the necessary requirements, we use totally nonsingular matrices.

Keywords: Secret sharing schemes • Multivariate Birkhoff interpolation • Hierarchies

V.E. Markoutis (✉)
Department of Mathematics, University of Patras, GR-26110 Patras, Greece
e-mail: billmarku@yahoo.gr

G.C. Meletiou
A.T.E.I. of Epirus, P.O. 110, GR-47100 Arta, Greece
e-mail: gmelet@teiep.gr

A.N. Veneti
Department of Informatics, University of Piraeus, 18534 Piraeus, Greece
e-mail: aveneti@unipi.gr

M.N. Vrahatis
Computational Intelligence Laboratory (CILab), Department of Mathematics, University of Patras, GR-26110 Patras, Greece
e-mail: vrahatis@math.upatras.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_14

1 Introduction

A secret sharing scheme is a methodology to distribute appropriately a piece of information of a secret, called ***share***, to each element of a specific set, called ***participant***, so that the secret can be reconstructed after the revelation of the shares of specific subsets of the set of participants. Since these specific subsets of participants depend on the secret sharing problem that has to be solved, a plethora of different schemes have been proposed.

Secret sharing schemes are very important, since they are used in various significant applications including cryptographic key distribution and sharing, e-voting, secure online auctions, information hiding as well as secure multiparty computation, among others. Shamir in [20] and Blakley in [5] invented independently, in 1979, the idea of secret sharing schemes. Shamir's approach is based on Lagrange interpolation while Blakley's method uses principles of hyperplane geometry. Tassa in [22] generalized Shamir's construction for a hierarchical threshold secret sharing scheme. His approach solves the problem of an efficient hierarchical threshold secret sharing scheme with a totally ordered set of levels of participants and is based on univariate Birkhoff interpolation. Birkhoff interpolation is a generalization of the Hermite case, obtained by relaxing the requirement of consecutive derivatives at the nodes.

In the contribution at hand we investigate the idea of generalizing Tassa's scheme through multivariate Birkhoff interpolation. We consider the problem of finding secret sharing schemes with multilevel structures and partially ordered sets of levels of participants. In order to ensure that our scheme meets the necessary requirements, we use totally nonsingular matrices.

In Sect. 2 of the work at hand we present basic concepts and background material related to secret sharing and threshold secret sharing schemes. Also, we briefly describe Blakley's scheme as well as we present Shamir's scheme based on Lagrange interpolation. Subsequently, in Sect. 3 we give some basic definitions related to Birkhoff interpolation. Next, in Sect. 4 we give a brief description of Tassa's secret sharing scheme based on univariate Birkhoff interpolation. In Sect. 5 we detail our ideas for constructing partially ordered secret sharing schemes through multivariate Birkhoff interpolation, we discuss the obtained results and open up some perspectives for our future work. The chapter ends in Sect. 6 with a synopsis.

2 Secret Sharing and Threshold Secret Sharing Schemes

In this section, basic concepts and background material related to secret sharing and threshold secret sharing schemes are given. Also, Blakley's scheme is briefly described. Furthermore, Shamir's scheme based on Lagrange interpolation is presented.

2.1 Secret Sharing Schemes

Stinson in his survey article for secret sharing schemes [21] gives a detailed description of the basic concepts of a secret sharing scheme.

Let $\mathscr{P}$ be a set of n participants that a secret is distributed to and Γ be the set of subsets of $\mathscr{P}$ such as $\Gamma \subseteq 2^{\mathscr{P}}$. The set Γ contains every subset of participants that should be able to compute the secret. Thus, Γ is called an ***access structure*** and the subsets in Γ are called ***authorized*** subsets. An access structure must satisfy the ***monotonicity*** property. Suppose that $B \in \Gamma$ and $B \subseteq C \subseteq \mathscr{P}$. Then the subset C can determine the value of secret key K. Formally we can say that [3, 21]:

$$\text{if} \quad B \in \Gamma \quad \text{and} \quad B \subseteq C \subseteq \mathscr{P}, \quad \text{then} \quad C \in \Gamma.$$

If Γ is an access structure, then $B \in \Gamma$ is a minimal authorized subset of $A \notin \Gamma$ whenever $A \subset B$. The set of minimal authorized subsets of Γ is denoted by Γ_0 and is called the ***basis*** of Γ.

Let D be a participant, called ***dealer***, who does not belong to the set $\mathscr{P}$. The dealer chooses the value of the secret and distributes the shares of the secret secretly so that no participant knows the share given to another participant. Also, let $\mathscr{K}$ be the ***key set*** and $\mathscr{S}$ be the ***share set***. When the dealer D wants to share a ***secret key*** $K \in \mathscr{K}$ he gives each participant a share from $\mathscr{S}$.

A simple approach for the definition of a secret sharing scheme is given in [6]. Given a set of n participants and an access structure Γ, a ***secret sharing scheme*** for Γ is a method of distributing shares to each of the participants such that:

1. Any subset of the participants in Γ can determine the secret.
2. Any subset of the participants that does not belong in Γ cannot determine the secret.

The share of a participant refers specifically to the information that the dealer D sends in private to the participant. If any subset of participants that does not belong in Γ cannot determine any information about the secret, then the secret sharing scheme is said to be ***perfect***. Given a secret sharing scheme we define the ***information rate*** ρ of the scheme as follows:

$$\rho = \frac{\log_2 |\mathscr{K}|}{\log_2 |\mathscr{S}|}. \tag{1}$$

If $\rho = 1$, then the scheme is called ***ideal***.

Remark 1. The first property implies that the shares given to an authorized subset uniquely determine the value of the secret. ***Accessibility*** and ***correctness*** are terms that are used alternatively to describe this property. The second property ensures that the shares given to an unauthorized subset reveal no information as to the value of the secret. ***Perfect security*** and ***privacy*** are terms that are used alternatively to describe this property.

The construction of a secret sharing scheme can be divided into the following three phases:

1. **Initialization phase:** The dealer chooses the secret key K.
2. **Secret sharing phase:** The dealer shares the secret key K among the set $\mathscr{P}$ of n participants giving each participant a share from $\mathscr{S}$ secretly.
3. **Secret reconstruction phase:** At a later time, a subset B of participants with $B \subseteq \mathscr{P}$ will pull their shares in an attempt to recompose the secret key K.

2.2 Threshold Secret Sharing Schemes

One of the most common class of secret sharing schemes is the class of threshold secret sharing schemes which implies that the reconstruction of the secret can be achieved by the contribution of a minimum number of participants of the set which we call ***threshold***.

Threshold secret sharing schemes were initially proposed for key management purposes. Let us recall an example from [21]:

Example 1. Assume that there is a vault in a bank that must be opened every day. The bank employs three senior tellers, but it is not desirable to entrust the combination to a unique person. We want to design a system whereby any two of the three senior tellers can gain access to the vault, but no individual can do so.

According to Shamir [20] a threshold secret sharing scheme can be defined as follows:

Definition 1. A (k, n) ***threshold secret sharing scheme*** is a method which gives efficient solution to the problem of the division of a piece of data K into n pieces $K_1, K_2, \ldots, K_n$ with the following two constraints:

1. K can be easily retrieved with the knowledge of k or more K_i pieces.
2. No information can be revealed about K with the knowledge of any $k-1$ or fewer K_i pieces.

Remark 2. In other words, a (k, n) threshold secret sharing scheme is a method of sharing a secret key K among a finite set $\mathscr{P}$ of n participants in such a way that any k participants can compute the value of K, but no one group of $k-1$ participants can do so.

A (k, n) threshold secret sharing scheme realizes the access structure:

$$\mathscr{A} = \{B \subseteq \mathscr{P} : |B| \geqslant k\}.$$

Such an access structure is called a ***threshold access structure***. It is obvious that in the case of a threshold access structure, the basis of the structure consists of all subsets of exactly k participants.

According to ***Blakley's scheme*** [5, 13] the secret is a point in a k-dimensional subspace over a finite field and the coefficients of the hyperplanes that intersect at this point are used to construct the shares. For the implementation of a (k, n) threshold secret sharing scheme, to each one of the n participants is given a hyperplane equation. In order to obtain the secret, a system of linear equations $Ax = y$ must be solved, where the matrix A and the vector y are derived from the hyperplane equations. When k participants come together, they can solve the system to find the intersection point of the hyperplanes in order to obtain the secret.

As we have mentioned before, Shamir [20] constructed a threshold secret sharing scheme using Lagrange interpolation. Also, Tassa [22] generalized Shamir's construction for a hierarchical threshold secret sharing scheme. His approach was based on univariate Birkhoff interpolation which solves the problem of an efficient hierarchical threshold secret sharing scheme with totally ordered set of levels of participants.

In our approach we investigate the construction of secret sharing schemes with the usage of multivariate Birkhoff interpolation. In this case, the structure that results is multilevel but the set of levels of participants is partially ordered.

Various threshold secret sharing schemes have been applied in many fields of information science [2] including threshold cryptography [10] and ad-hoc networks [1] among others.

2.3 Shamir's Scheme Through Lagrange Interpolation

As we have already mentioned, Shamir in [20] introduced the idea of a threshold secret sharing scheme through polynomial interpolation. His idea was based on Lagrange interpolation. More specifically, he exploited the fact that given k points on a 2-dimensional plane $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$ with distinct x_i, there exists one and only one polynomial $g(x)$ of $k-1$ degree such that $g(x_i) = y_i$ for all $i = 1, 2, \ldots, k$.

Thus, in order to divide and share a secret S he considered a random polynomial $g(x)$ of $k-1$ degree as following:

$$g(x) = a_1 x + a_2 x^2 + \cdots + a_{k-1} x^{k-1} + S. \tag{2}$$

A polynomial interpolating value $g(x_i) = y_i$ is a share that can be given to a participant. A set of k shares are enough to define the unique polynomial $g(x)$ and obviously reveal S while $k-1$ or less shares do not suffice for the calculation of S.

Shamir's (k, n) threshold secret sharing scheme can be described by the following algorithm:

Algorithm 1

1. **Initialization phase:** The dealer chooses n distinct nonzero elements from a finite field $\mathbb{F}_q$, $\{x_1, x_2, \ldots, x_n\}$, and gives x_i to the i-th participant p_i. In other terms participant p_i is identified to the field element x_i.
2. **Secret sharing phase:** The dealer secretly chooses $k-1$ elements from $\mathbb{F}_q$, $\{a_1, a_2, \ldots, a_{k-1}\}$, and considers the following polynomial:

$$g(x) = \sum_{i=1}^{k-1} a_i x^i + S, \tag{3}$$

   where S is the constant term of the polynomial which represents the secret. The dealer computes the n shares $y_i = g(x_i)$ and gives each share to the corresponding participant.
3. **Secret reconstruction phase:** A subset B of k participants $\{p_{i_1}, p_{i_2}, \ldots, p_{i_k}\}$ will pull their shares and attempt to reconstruct S. Suppose that the k shares $y_{i_j} = g(x_{i_j})$, $1 \leqslant j \leqslant k$ are revealed. Then, the coefficients of polynomial $g(x)$ can be evaluated by Lagrange interpolation. Consequently secret S is obtained by the evaluation $S = g(0)$.

3 Birkhoff Interpolation

The problem of interpolating a function $f : \mathbb{R} \to \mathbb{R}$ by a univariate polynomial from the values of f and some of its derivatives on a set of sample points is one of the main questions in Numerical Analysis and Approximation Theory [18].

Birkhoff interpolation [4, 15, 17, 19] is a generalization of Lagrange and Hermite polynomial interpolation. It amounts to the problem of finding a polynomial $f(x)$ of degree $k-1$ such that certain derivatives have specified values at specified points:

$$f^{(n_i)}(x_i) = y_i, \quad \text{for} \quad i = 1, 2, \ldots, k, \tag{4}$$

where the data points (x_i, y_i) as well as the nonnegative integers n_i are given.

Remark 3. In contrast to Lagrange and Hermite interpolation problems which are well posed, Birkhoff interpolation problems do not always have unique solution.

Definition 2. Let $\mathscr{X} = \{x_1, x_2, \ldots, x_n\}$ be an ordered set of real numbers such that $x_1 < x_2 < \cdots < x_n$ and $\mathscr{I} \subset \{1, 2, \ldots, n\} \times \{0, 1, \ldots, r\}$ be the set of pairs (i, j) such that the value $f_{i,j} = f^{(j)}(x_i)$ is known. The problem of determining the existence and uniqueness of a polynomial Q in $\mathbb{R}[X]$ of degree bounded by r such that:

$$\forall\, (i, j) \in \mathscr{I}, \quad Q^{(j)}(x_i) = f_{i,j}\,, \tag{5}$$

is called the ***Birkhoff interpolation problem***.

The multivariate Birkhoff interpolation problem is more complicated. A formal definition of this problem can be given as follows [8, 14]:

Definition 3. A ***multivariate Birkhoff interpolation scheme***, $(E, \mathbb{W}_s)$, consists of three components:

1. A **set of nodes** $\mathscr{Z}$:

$$\mathscr{Z} = \{z_t\}_{t=1}^{m} = \{(x_{t,1}, x_{t,2}, \ldots, x_{t,d})\}_{t=1}^{m} . \tag{6}$$

2. An **interpolation space** $\mathbb{W}_s$:

$$\mathbb{W}_s = \left\{W \ : \ W(z) = W(x_1, x_2, \ldots, x_d) = \sum_{i \in S} a_i x_1^{i_1}, \ldots, x_d^{i_d}\right\}, \tag{7}$$

where S is a lower subset of $\mathbb{N}_0^d$. A subset A of $\mathbb{N}_0^d$ is a lower set if $0 \leqslant j_k \leqslant i_k$, $k = 1, 2, \ldots, d$ and $i \in S$ implies that $j \in S$.

3. An **incidence** $(d+1)$**-dimensional matrix** E:

$$E = \{e_{t,\alpha}\}, \quad t = 1, 2, \ldots, m, \quad \alpha \in S, \tag{8}$$

where $e_{t,\alpha} = 0$ or $e_{t,\alpha} = 1$.

Given these components, the ***multivariate Birkhoff interpolation problem*** is, for given real numbers $c_{t,\alpha}$ for those t, α with $e_{t,\alpha} = 1$, to find a polynomial $W \in \mathbb{W}_s$ satisfying the interpolation conditions:

$$\frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_d}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}} W(z_t) = c_{t,\alpha} , \tag{9}$$

for those t, α with $e_{t,\alpha} = 1$.

Remark 4. The aforementioned schemes are interpolations over the real numbers. In cryptographic applications finite fields are used and derivatives (ordinary and partial) are replaced with formal derivatives of polynomials. Since we deal with polynomials it is always true that $\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}$.

4 Tassa's Scheme Through Univariate Birkhoff Interpolation

Tassa in [22] proposed a perfect and ideal secret sharing scheme for a multilevel totally ordered structure. His approach is based on univariate Birkhoff interpolation. Since Tassa's scheme is the basis of our scheme, we detail his following definition for a hierarchical threshold secret sharing scheme.

Definition 4. Let $\mathscr{P}$ be a set of n participants and assume that $\mathscr{P}$ is composed of levels, i.e., $\mathscr{P} = \cup_{i=0}^{m} \mathscr{P}_i$ where $\mathscr{P}_i \cap \mathscr{P}_j = \emptyset$ for all $0 \leqslant i < j \leqslant m$. Let $\boldsymbol{\kappa} = \{k_i\}_{i=0}^{m}$ be a monotonically increasing sequence of integers, $0 < k_0 < k_1 < \cdots < k_m$. Then, the $(\boldsymbol{\kappa}, n)$ ***hierarchical threshold access structure*** is given as follows:

$$\Gamma = \left\{B \subset \mathscr{P} : \ \left|B \cap \left(\cup_{j=0}^{i} \mathscr{P}_j\right)\right| \geqslant k_i , \quad \forall \ i \in \{0, 1, \ldots, m\}\right\}. \tag{10}$$

Algorithm 2

1. **Initialization phase:** The dealer chooses n distinct nonzero elements from a finite field $\mathbb{F}_q$, $\{x_1, x_2, \ldots, x_n\}$, and gives x_i to the i-th participant p_i. In other terms the dealer identifies each participant $p \in \mathscr{P}$ with an element of the field $\mathbb{F}_q$. For simplicity, the field element is also denoted by p.
2. **Secret sharing phase:** The dealer secretly chooses $k-1$ elements from $\mathbb{F}_q$, $\{a_1, a_2, \ldots, a_{k-1}\}$, and considers the following polynomial:

$$g(x) = \sum_{i=1}^{k-1} a_i x^i + S, \tag{11}$$

   where S is the constant term of the polynomial which represents the secret. Every participant p of the i-th level of the hierarchy receives the share:

$$y = \left(\frac{\mathrm{d}^{k_{i-1}} g}{\mathrm{d}x^{k_{i-1}}} \right)_p = g^{(k_{i-1})}(p),$$

   where $g^{(k_{i-1})}(p)$ is the k_{i-1}-th formal derivative of $g(x)$ at $x = p$ with $k_{-1} = 0$.
3. **Secret reconstruction phase:** An authorized subset B of k participants will pull their shares and attempt to reconstruct S. Then, the coefficients of polynomial $g(x)$ can be evaluated by univariate Birkhoff interpolation. Consequently secret S is obtained by the evaluation $S = g(0)$.

A corresponding $(\boldsymbol{\kappa}, n)$ ***hierarchical secret sharing scheme*** is a scheme that realizes the above access structure; namely, a method of assigning each participant $P_l \in \mathscr{P}$, with $0 \leqslant l < n$, a share $\sigma(P_l)$ of a given secret S such that authorized subsets $B \in \Gamma$ may recover the secret from the shares possessed by their participants, $\sigma(B) = \{\sigma(P_l) : P_l \in B\}$, while the shares of unauthorized subsets $B \notin \Gamma$ do not reveal any information about the value of the secret.

Remark 5. For the construction of a hierarchical threshold secret sharing scheme, Tassa used k-order derivatives and constructed shares for each level according to the order of the derivative. In this way he ensured that the participants of an upper hierarchically level possess more amount of information to their share than the participants of a lower level. The calculation of the polynomial coefficients during the secret reconstruction phase was based on the univariate Birkhoff interpolation.

Tassa's $(\boldsymbol{\kappa}, n)$ hierarchical threshold secret sharing scheme with $\boldsymbol{\kappa} = \{k_i\}_{i=0}^m$ and $k = k_m$ can be described by the following algorithm:

5 The Proposed Approach

As we have already mentioned, in our approach we investigate the construction of secret sharing schemes with the usage of multivariate Birkhoff interpolation. In this case, the structure that results is multilevel but the set of levels of participants is partially ordered. In Tassa's scheme, the shares of two participants p_a and p_b of different levels have, by necessity, at least one of the following two properties:

(a) The share of p_a can substitute the share of p_b.
(b) The share of p_b can substitute the share of p_a.

In our case this is not always true due to the partial order of the levels of participants and this is the main difference with Tassa's scheme.

We illustrate our ideas through some examples and we propose a construction for the simple linear case of a threshold multilevel partially ordered secret sharing scheme. However, a generalized case of a partially ordered set should be an object of a much more complicated effort. At this point, it must be noted that a partially ordered secret sharing scheme is not hierarchical, since a hierarchical structure presupposes a totally ordered set of participants [11, 12].

5.1 The Main Idea of Our Approach

We consider the multivariate polynomial $g(x_1, x_2, \ldots, x_d)$ with coefficients from a finite field. The constant term of the polynomial denotes the ***secret*** S, that is $g(0, 0, \ldots, 0) = S$. Some participants receive ***shares*** of the following form:

$$y_t = g(x_{t,1}, x_{t,2}, \ldots, x_{t,d}) = g(z_t),$$

and they consist of the (top) level $\mathscr{P}_0$ (the d-tuples are nodes as they are described in Definition 3). Some participants receive shares of the form $\frac{\partial}{\partial x_1} g(z_t)$ and they belong to the level $\mathscr{P}_1$. In a similar way we define $\mathscr{P}_2, \mathscr{P}_3, \ldots$ The level $\mathscr{P}_{1,1}$ is related to the shares of the form $\frac{\partial^2}{\partial x_1^2} g(z_t)$ while the level $\mathscr{P}_{1,2}$ is related to the shares of the form $\frac{\partial^2}{\partial x_1 \partial x_2} g(z_t)$. Since $\frac{\partial^2}{\partial x_1 \partial x_2} g(z_t) = \frac{\partial^2}{\partial x_2 \partial x_1} g(z_t)$, the level $\mathscr{P}_{1,2}$ coincides with the level $\mathscr{P}_{2,1}$. Thus, an ordered set of levels is derived which have the form $\mathscr{P}_{j_1, j_2, \ldots, j_n}$.

For $d = 2$ and $d = 3$ the obtained multilevel structures are exhibited in Figs. 1 and 2, respectively.

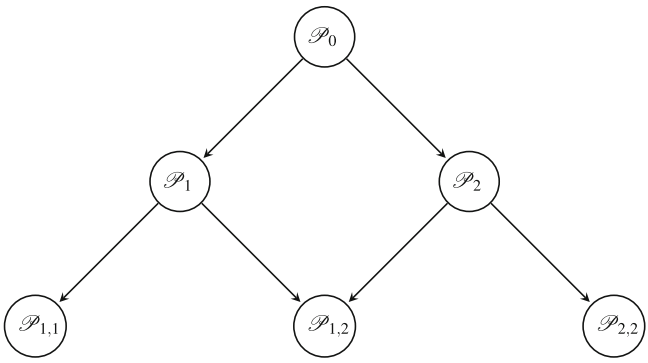


Fig. 1 The structure of a secret sharing scheme that can be constructed from a polynomial $g(x_1, x_2)$

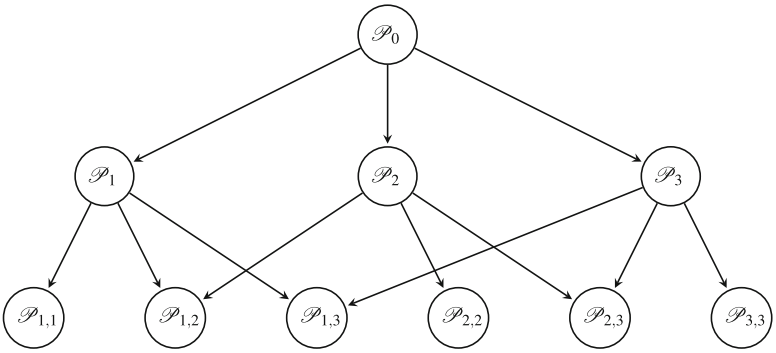


Fig. 2 The structure of a secret sharing scheme that can be constructed from a polynomial $g(x_1, x_2, x_3)$

Table 1 The distributed shares for the participants of the scheme that can be constructed from the polynomial g_1

Level of participant	Type of share
$\mathscr{P}_0$	$g_1(x_1, x_2) = a_1x_1^2 + a_2x_2^2 + a_3x_1x_2 + S$
$\mathscr{P}_1$	$\frac{\partial g_1}{\partial x_1} = 2a_1x_1 + a_3x_2$
$\mathscr{P}_2$	$\frac{\partial g_1}{\partial x_2} = 2a_2x_2 + a_3x_1$
$\mathscr{P}_{1,1}$	$\frac{\partial^2 g_1}{\partial x_1^2} = 2a_1$
$\mathscr{P}_{1,2}$	$\frac{\partial^2 g_1}{\partial x_1 \partial x_2} = a_3$
$\mathscr{P}_{2,2}$	$\frac{\partial^2 g_1}{\partial x_2^2} = 2a_2$

Remark 6. In order to reconstruct S from the shares we have to tackle the multivariate Birkhoff interpolation problem. The set of levels is a partially ordered set, namely an upper semilattice. The level P is "greater" (or "higher") than the level Q, $P \succ Q$ means that a participant from P can replace a participant from Q.

The main idea of our approach is illustrated in the following examples.

5.2 Illustrative Examples

We consider the following polynomial:

$$g_1(x_1, x_2) = a_1x_1^2 + a_2x_2^2 + a_3x_1x_2 + S. \tag{12}$$

By taking the first-order partial derivatives of the polynomial g_1 we get the polynomials that give the shares of the $\mathscr{P}_i$, $i = 1, 2$ level participants. Subsequently, by taking the second-order partial order derivatives we get the values of the shares of the $\mathscr{P}_{i,j}$, $i, j = 1, 2$ level participants. The shares that are distributed to the participants are exhibited in Table 1 while the consequent structure is the same as exhibited in Fig. 1.

Working in the same manner, we are able to construct a plethora of structures that represent the hierarchical relationship between participants in a secret sharing scheme. For example, we consider the following polynomial:

$$g_2(x_1, x_2, x_3) = a_1x_1x_2 + a_2x_2x_3 + a_3x_1x_3 + S. \tag{13}$$

The partial derivatives of the polynomial g_2 are used as the distributed shares of Table 2. The structure of the resulted secret sharing scheme is exhibited in Fig. 3.

Next, we present two additional illustrative examples by considering the polynomials:

$$g_3(x_1, x_2, x_3) = \lambda(x_1^2 + x_2^2 + x_3^2) + a_1x_1 + a_2x_2 + a_3x_3 + S, \tag{14}$$

and

$$g_4(x_1, x_2) = ax_1^3 + bx_1^2 + cx_1 + ax_2^2 + dx_2 + S. \tag{15}$$

Table 2 The distributed shares for the participants of the scheme that can be constructed from the polynomial g_2

Level of participant	Type of share
$\mathscr{P}_0$	$g_2(x_1, x_2, x_3) = a_1x_1x_2 + a_2x_2x_3 + a_3x_1x_3 + S$
$\mathscr{P}_1$	$\frac{\partial g_2}{\partial x_1} = a_1x_2 + a_3x_3$
$\mathscr{P}_2$	$\frac{\partial g_2}{\partial x_2} = a_1x_1 + a_2x_3$
$\mathscr{P}_3$	$\frac{\partial^2 g_2}{\partial x_3} = a_2x_2 + a_3x_1$
$\mathscr{P}_{1,2}$	$\frac{\partial^2 g_2}{\partial x_1 \partial x_2} = a_1$
$\mathscr{P}_{2,3}$	$\frac{\partial^2 g_2}{\partial x_2 \partial x_3} = a_2$
$\mathscr{P}_{1,3}$	$\frac{\partial^2 g_2}{\partial x_1 \partial x_3} = a_3$

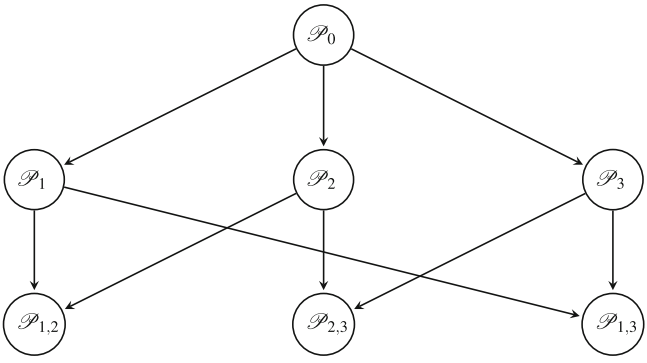


Fig. 3 The structure of a secret sharing scheme that can be constructed from the polynomial g_2

Table 3 The distributed shares for the participants that can be constructed from the polynomial g_3

Level of participant	Type of share
$\mathscr{P}_0$	$g_3(x_1, x_2, x_3) = \lambda(x_1^2 + x_2^2 + x_3^2) + a_1x_1 + a_2x_2 + a_3x_3 + S$
$\mathscr{P}_1$	$\frac{\partial g_3}{\partial x_1} = 2\lambda x_1 + a_1$
$\mathscr{P}_2$	$\frac{\partial g_3}{\partial x_2} = 2\lambda x_2 + a_2$
$\mathscr{P}_3$	$\frac{\partial g_3}{\partial x_3} = 2\lambda x_3 + a_3$
$\mathscr{P}_{xx} = \mathscr{P}_{1,1} = \mathscr{P}_{2,2} = \mathscr{P}_{3,3}$	$\frac{\partial^2 g_3}{\partial x_1^2} = \frac{\partial^2 g_3}{\partial x_2^2} = \frac{\partial^2 g_3}{\partial x_3^2} = 2\lambda$

Table 4 The distributed shares for the participants that can be constructed from the polynomial g_4

Level of participant	Type of share
$\mathscr{P}_0$	$g_4(x_1, x_2) = ax_1^3 + bx_1^2 + cx_1 + ax_2^2 + dx_2 + S$
$\mathscr{P}_1$	$\frac{\partial g_4}{\partial x_1} = 3ax_1^2 + 2bx_1 + c$
$\mathscr{P}_{11}$	$\frac{\partial^2 g_4}{\partial x_1^2} = 6ax_1 + 2b$
$\mathscr{P}_2$	$\frac{\partial g_4}{\partial x_2} = 2ax_2 + d$
$\mathscr{P}' = \mathscr{P}_{1,1,1} = \mathscr{P}_{2,2}$	$\frac{\partial^3 g_4}{\partial x_1^3} = 3\frac{\partial^2 g_4}{\partial x_2^2} = 6a$

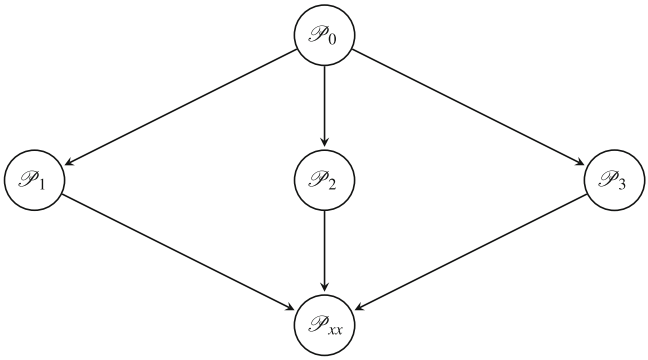


Fig. 4 The structure of a secret sharing scheme that can be constructed from the polynomial g_3

In Tables 3 and 4 we present, respectively, the shares that are distributed to the participants. The corresponding structures are exhibited in Figs. 4 and 5.

5.3 The Linear Polynomial Case

In this subsection we present a ***threshold*** $(n + 1)$***-level partially ordered secret sharing scheme***. To this end, we consider the scheme that is derived from an n-variable linear polynomial of the following form:

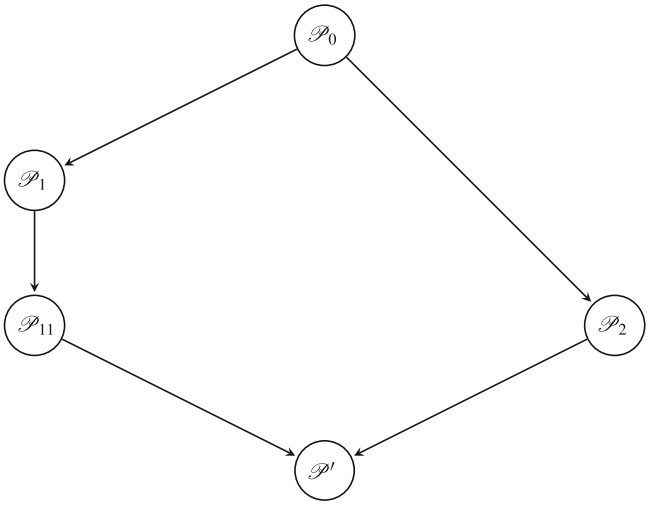


Fig. 5 The structure of a secret sharing scheme that can be constructed from the polynomial g_4

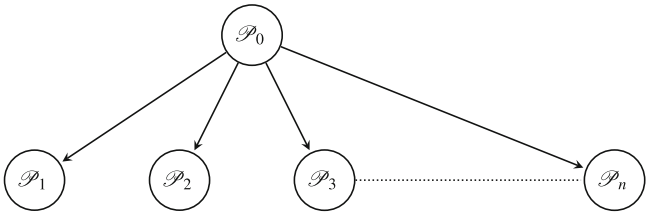


Fig. 6 The structure of a partially ordered $(\boldsymbol{\kappa}, 2n+1)$ threshold secret sharing scheme with $\boldsymbol{\kappa} = (1, n+1)$

$$g(x_1, x_2, \ldots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + S, \tag{16}$$

where $a_1, a_2, \ldots, a_n$ are the coefficients of the polynomial g and S is the constant term of the polynomial that represents the secret key. The partial order which is defined has the structure exhibited in Fig. 6.

The information piece (share) for participants from $\mathscr{P}_0$ is unique. Therefore, without loss of generality we assume that $\mathscr{P}_j$ has exactly one participant $|\mathscr{P}_j| = 1$. Also, without loss of generality we assume that $\mathscr{P}_0$ contains $n+1$ participants, which determines the minimal number for reconstructing the secret S.

The specific structure has two important properties:

(a) None of the participants of a level $\mathscr{P}_j, j \neq 0$ can replace a participant of a level $\mathscr{P}_i, i \neq 0, j$. Thus, we say that we have a ***partially ordered structure***.
(b) ***Participants of the level $\mathscr{P}_0$ can replace whichever participant*** of the structure such that an authorized subset can be constructed.

The following algorithm describes the corresponding secret sharing scheme:

Algorithm 3

1. **Initialization phase:** The dealer selects the following polynomial:

$$g(x_1, x_2, \ldots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + S, \tag{17}$$

where a_i are elements that are chosen randomly from a finite field $\mathbb{F}_q$ and q is a large prime power. The participants are identified by the dealer so that each participant from $\mathscr{P}_0$ is identified with the n-tuple $(x_{1,k}, x_{2,k}, \ldots, x_{n,k}) \in \mathbb{F}_q^n$ after a suitable selection of the $x_{i,k}$ and each participant from $\mathscr{P}_j$, $1 \leqslant j \leqslant n$ is identified with the index j.

2. **Secret sharing phase:** The dealer distributes the shares so that each participant from $\mathscr{P}_0$ receives the value:

$$y_k = g(x_{1,k}, x_{2,k}, \ldots, x_{n,k}) \in \mathbb{F}_q, \tag{18}$$

and each participant from $\mathscr{P}_j$ receives the value:

$$a_j = \frac{\partial g}{\partial x_j}. \tag{19}$$

3. **Secret reconstruction phase:** A subset B of participants will pull their shares and attempt to reconstruct S. This can be done by solving a system of linear equations. The unknowns are the coefficients a_i as well as the element S. The participants from $\mathscr{P}_0$ will pull their equation:

$$y_k = \sum_{i=1}^{n} a_i x_{i,k} + S. \tag{20}$$

The participants from $\mathscr{P}_j$, $1 \leqslant j \leqslant n$ will pull the value:

$$a_j = \frac{\partial g}{\partial x_j}. \tag{21}$$

If the $x_{i,k}$ with $1 \leqslant i \leqslant n$ and $1 \leqslant k \leqslant n+1$ are suitably chosen from a finite field, then a unique solution exists for S if B is an authorized subset, $|B| \geqslant n+1$.

Next, we define a class of matrices on which our scheme is based.

Definition 5. An $n \times n$ matrix is called a ***principally nonsingular matrix*** if every principal submatrix is nonsingular. Also, an $n \times n$ matrix is said to be a ***totally nonsingular matrix*** if all its square submatrices are nonsingular.

Remark 7. This class of matrices contains the ***totally negative matrices***, whose the determinant of the corresponding minors is strictly negative, and the ***totally positive matrices*** whose the determinant of the corresponding minors is strictly positive. If we allow the existence of null minors, these classes can be extended to the ***totally nonpositive matrices*** as well as to the ***totally nonnegative matrices***.

The following theorem gives necessary and sufficient conditions for accessibility and perfect security of the scheme:

Theorem 1. *Consider the following* $(n+1)\times(n+1)$ *matrix:*

$$X_1 = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{n,1} & 1 \\ x_{1,2} & x_{2,2} & \cdots & x_{n,2} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{n,n} & 1 \\ x_{1,n+1} & x_{2,n+1} & \cdots & x_{n,n+1} & 1 \end{pmatrix}. \tag{22}$$

Then, the accessibility and perfect security are satisfied iff the matrix X_1 *is totally nonsingular.*

Proof. Let us denote by X the following matrix:

$$X = \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,n} & x_{2,n} & \cdots & x_{n,n} \\ x_{1,n+1} & x_{2,n+1} & \cdots & x_{n,n+1} \end{pmatrix}. \tag{23}$$

We consider the following cases:

Case 1: *All participants belong to the level* $\mathscr{P}_0$.

According to the assumptions of the theorem all the rows of X and X_1 are linearly independent. Retrieving the secret S amounts to the solution of the following system:

$$\sum_{i=1}^{n} a_i x_{i,j} + S = y_j, \quad j = 1, 2, \ldots, n+1. \tag{24}$$

The matrix X_1 is the coefficient matrix of the system and the elements $a_1, a_2, \ldots, a_n, S$ are the unknowns. The condition $\det(X_1) \neq 0$ implies existence of a unique solution and the secret S can be retrieved. Thus, the **accessibility is satisfied**.

Let B be a set of participants with $|B| = n$. It corresponds to a set of n rows of X, namely $\{(x_{1,j_k}, x_{2,j_k}, \ldots, x_{n,j_k})\}$, $k = 1, 2, \ldots, n$. The unknown S can be treated as a parameter. For any randomly chosen value S_0 of S we derive the following linear system:

$$\sum_{i=1}^{n} a_i x_{i,j_k} = y_{j_k} - S_0, \quad k = 1, 2, \ldots, n. \tag{25}$$

Its coefficient matrix is an $n \times n$ minor of X. According to the assumptions it has exactly one solution for all S_0, therefore no information can be revealed about S. Thus, the **perfect security is satisfied**.

For a set B of m participants, $|B| = m \leqslant n$ the same technique can be used. The corresponding set of rows of X is $\{(x_{1,j_k}, x_{2,j_k}, \ldots, x_{n,j_k})\}$, $k = 1, 2, \ldots, m$, which are linearly independent due to the assumption. The following linear system of m equations:

$$\sum_{i=1}^{n} a_i x_{i,j_k} = y_{j_k} - S_0, \quad k = 1, 2, \ldots, m, \tag{26}$$

has exactly q^{n-m} solutions (where q is the cardinality of the finite field $\mathbb{F}_q$) and no information can be obtained about S.

For the inverse part of the proof, let us assume that the conditions of accessibility and perfect security are satisfied. On the contrary, assume that $\det(X_1) = 0$. Also, let us assume that the linear system (24) has more than one solutions and a unique value can be found for the unknown S, which is possible. This implies that at least one of the equations of the system (24) can be removed and that n or less than n participants can reveal the secret S, which is a contradiction to the assumption of perfect security. Therefore, the rows of X_1 are linearly independent.

Again, on the contrary, assume that m rows of X, $m < n + 1$ are linearly dependent, namely the rows $\{(x_{1,j_k}, x_{2,j_k}, \ldots, x_{n,j_k})\}$, $k = 1, 2, \ldots, m$. However the corresponding rows $\{(x_{1,j_k}, x_{2,j_k}, \ldots, x_{n,j_k}1, 1)\}$, $k = 1, 2, \ldots, m$ of X_1 are linearly independent and S can be retrieved from m participants which is a contradiction to the assumption of perfect security.

We conclude that matrix X_1 is invertible and that all submatrices $n \times n$ minors of X are invertible.

Case 2: *Some of the participants belong to the levels* $\mathscr{P}_1, \mathscr{P}_2, \ldots, \mathscr{P}_n$.

This case can be treated as the Case 1. Assume that r participants, $r \leqslant n$, belong to the levels $\mathscr{P}_{t_1}, \mathscr{P}_{t_2}, \ldots, \mathscr{P}_{t_r}$, where $\{t_1, t_2, \ldots, t_r\} \subseteq \{1, 2, \ldots, n\}$ and that the remaining $n + 1 - r$ participants belong to the level $\mathscr{P}_0$. The share of the participant of the level $\mathscr{P}_{t_l}$, $1 \leqslant l \leqslant r$, is $a_{t_l} = \frac{\partial g}{\partial x_{t_l}}$ and the t_l-th column has to be deleted from the matrix X_1. The new obtained matrix X_1' has $n + 1 - r$ rows corresponding to the $n + 1 - r$ participants from $\mathscr{P}_0$, and $n + 1 - r$ columns after the deletion of r columns. The new matrix X' is $(n + 1 - r) \times (n - r)$. The rest of the proof is similar to the Case 1.

Thus the theorem is proved. □

Remark 8. Obviously, the scheme is also ***ideal***, since every participant receives a field element, just like the secret.

Remark 9. For the implementation of the scheme a totally nonsingular matrix is required which can be obtained from a totally positive matrix [7] over the reals. The well-known ***Hilbert matrix***:

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1} \end{pmatrix}, \tag{27}$$

is totally positive [16]. Also, totally nonsingular matrices can be derived from the ***Vandermonde matrix*** under specific conditions [9].

Next, we present an illustrative example. To this end, we consider the structure exhibited in Fig. 7 with which we represent a $(\boldsymbol{\kappa}, 7)$ four-level threshold partially ordered secret sharing scheme with $\boldsymbol{\kappa} = (1, 4)$. In order to construct the scheme we use the Hilbert matrix. For this case the corresponding algorithm of our approach is the following:

5.4 Perspectives for Future Work

Multivariate Birkhoff interpolation over large degree polynomials is a challenge to build multilevel threshold secret sharing schemes with partially ordered sets of levels. The ordered set, exhibited in Fig. 8, represents the structure of a multilevel partially ordered threshold secret scheme.

Observing this special structure a general question has to be answered: *Given a scheme with a partially ordered set of levels as above, is it always feasible to find a multivariate polynomial, such that the order is derived from the polynomial ?*

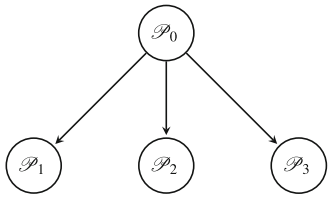


Fig. 7 The structure of a partially ordered $(\boldsymbol{\kappa}, t)$ four-level threshold secret sharing scheme with $\boldsymbol{\kappa} = (1, 4)$

Algorithm 4

1. **Initialization phase:** The dealer selects a polynomial of the form:

$$g(x_1, x_2, x_3) = \sum_{i=1}^{3} a_i x_i + S, \tag{28}$$

where a_i are chosen randomly over a finite field $\mathbb{F}_q$. Let us assume that $a_1 = 2, a_2 = 4, a_3 = 5$ and $q = 11$. Suppose further that the secret S is 8. Participants are identified by the dealer so that each participant from $\mathscr{P}_0$ is identified with the first 3 elements of a row of the 4×4 matrix which has been resulted after the transformation, with row multiplication, of the last column of the 4×4 Hilbert matrix to a vector of ones, and each participant from $\mathscr{P}_j$, $1 \leqslant j \leqslant 3$ with the index j.

2. **Secret sharing phase:** The dealer distributes the shares so that each participant from $\mathscr{P}_0$ receives the value:

$$y_k = g(x_{1,k}, x_{2,k}, x_{3,k}) \in \mathbb{F}_{11}, \tag{29}$$

and each participant from $\mathscr{P}_j$ receives the value:

$$a_j = \frac{\partial g}{\partial x_j}. \tag{30}$$

The distributed shares are shown in Table 5.

3. **Secret reconstruction phase:** Suppose now that we have an authorized set which consists of 2 participants of the level $\mathscr{P}_0$ and 2 participants of the levels $\mathscr{P}_j$, $1 \leqslant j \leqslant 3$. For example we assume that we have the subset $\{p_0^1, p_0^3, p_1, p_2\}$. Since p_1, p_2 are elements of the specific subset, a_1 and a_2 are the coefficients that we can obtain directly. Due to the presence of participants p_0^1 and p_0^3 in the set, we obtain the following linear system:

$$a_1 x_{1,1} + a_2 x_{2,1} + a_3 x_{3,1} + S = y_1,$$

$$a_1 x_{1,3} + a_2 x_{2,3} + a_3 x_{3,3} + S = y_3,$$

$$a_1 = 2,$$

$$a_2 = 4.$$

By substituting $x_{i,k}$ with the corresponding elements of the transformed Hilbert and y_j with the values of the shares, we rewrite the system as follows:

$$4a_1 + 2a_2 + 5a_3 + S = 5,$$

$$2a_1 + 7a_2 + 10a_3 + S = 2,$$

$$a_1 = 2,$$

$$a_2 = 4.$$

By computing the inverses over finite field $\mathbb{F}_{11}$ we finally get $S = 8$ which is the correct value.

Table 5 The distributed shares for the participants of the scheme of Fig. 7

Participant	Value of share
p_0^1	5
p_0^2	3
p_0^3	2
p_0^4	9
p_1	2
p_2	4
p_3	5

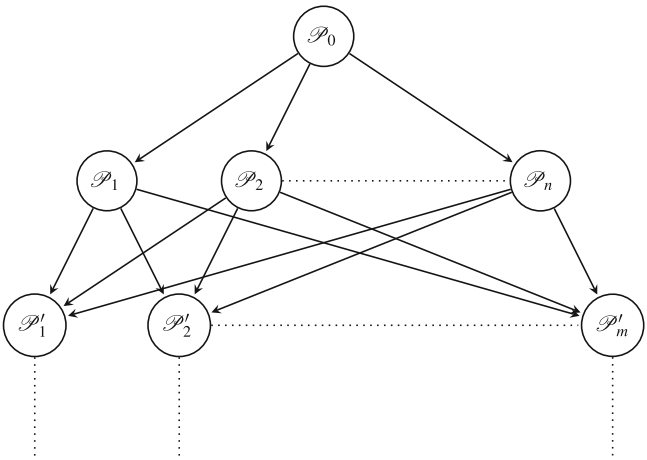


Fig. 8 The structure of a partially ordered threshold secret sharing scheme

6 Synopsis

In the work at hand, we investigated the adaptation of multivariate Birkhoff interpolation problem for the construction simple secret sharing schemes. The resulted structures consist of partially ordered levels of participants. For a simple linear polynomial with n variables, the secret sharing scheme that can be constructed is perfect with the usage of totally nonsingular matrices which ensure both correctness and perfect security.

Finally we posed the analogous generalized problem, which implies the construction of specific structures with polynomials through the multivariate Birkhoff interpolation problem.

References

1. Ballico, E., Boato, G., Fontanari, C., Granelli, F.: Hierarchical secret sharing in ad hoc networks through Birkhoff interpolation. In: Elleithy, K., Sobh, T., Mahmood, A., Iskander, M., Karim, M. (eds.) Advances in Computer, Information, and Systems Sciences, and Engineering, pp. 157–164. Springer, Dordrecht (2006)
2. Beimel, A.: Secret-sharing schemes: a survey. Lect. Notes Comput. Sci. **6639**, 11–46 (2011)
3. Benaloh, J.C., Leichter, J.: Generalized secret sharing and monotone functions. Lect. Notes Comput. Sci. **403**, 27–36 (1990)
4. Birkhoff, G.D.: General mean value and remainder theorems with applications to mechanical differentiation and quadrature. Trans. Am. Math. Soc. **7**, 107–136 (1906)
5. Blakley, G.R.: Safeguarding cryptographic keys. In: Proceedings of the 1979 AFIPS National Computer Conference, vol. 48, pp. 313–317. AFIPS Press, Montvale, NJ (1979)
6. Brickel, E.F.: Some ideal secret sharing schemes. J. Combin. Math. Combin. Comput. **6**, 105–113 (1989)
7. Carnicer, J.M.: Interpolation shape control and shape properties. In: Peña, J.M. (ed.) Shape Preserving Representations in Computer-Aided Geometric Design, Chapter 2, pp. 15–43. Nova Science Publishers, Commack, NY (1999)
8. Crainic, M., Crainic, N.: Pólya conditions for multivariate Birkhoff interpolation: from general to rectangular sets of nodes. Acta Math. Univ. Comenianae **79**(1), 9–18 (2010)
9. Demmel, J., Koev, P.: The accurate and efficient solution of a totally positive generalized Vandermonde linear system. SIAM J. Matrix Anal. Appl. **27**(1), 142–152 (2005)
10. Desmedt, Y., Frankel, Y.: Shared generation of authenticators and signatures. Lect. Notes Comput. Sci. **576**, 457–469 (1992)
11. Farràs, O., Padró, C.: Ideal hierarchical secret sharing schemes. Lect. Notes Comput. Sci. **5978**, 219–236 (2010)
12. Farràs, O., Padró, C.: Ideal hierarchical secret sharing schemes. IEEE Trans. Inform. Theory **58**(5), 3273–3286 (2012)
13. Hei, X.-L., Du, X.-J., Song, B.-H.: Two matrices for Blakley's secret sharing scheme. In: Proceedings of the 2012 IEEE International Conference on Communications (ICC), pp. 810–814. IEEE (2012)
14. Lorentz, R.A.: Multivariate Birkhoff Interpolation. Lecture Notes in Mathematics Series, vol. 1516. Springer, Berlin/Heidelberg (1992)
15. Lorentz, G.G., Jetter, K., Riemenschneider, S.D.: Birkhoff Interpolation. Encyclopedia of Mathematics and Its Applications, vol. 19. Addison-Wesley, Reading (1982)
16. Peña, J.M.: Stability and error analysis of shape preserving representations. In: Peña, J.M. (ed.) Shape Preserving Representations in Computer-aided Geometric Design, Chapter 5, pp. 85–97. Nova Science Publishers, Commack, NY (1999)
17. Pólya, G.: Bemerkung zur Interpolation und zur Naherungstheorie der Balkenbiegung. Z. Angew. Math. Mech. **11**, 445–449 (1931)
18. Rouillier, F., El Din, M.S., Schost, E.: Solving the Birkhoff interpolation problem via the critical point method. Lect. Notes Artif. Intell. **2061**, 26–40 (2001)
19. Schoenberg, I.J.: On Hermite-Birkhoff interpolation. J. Math. Anal. Appl. **16**, 538–543 (1966)
20. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
21. Stinson, D.R.: An explication of secret sharing schemes. Designs Codes Cryptogr. **2**(4), 357–390 (1992)
22. Tassa, T.: Hierarchical threshold secret sharing. J. Cryptol. **20**(2), 237–264 (2007)

Advanced Truncated Differential Attacks Against GOST Block Cipher and Its Variants

Theodosis Mourouzis and Nicolas Courtois

Abstract GOST block cipher, defined in the GOST 28147-89 standard, is a well-known 256-bit symmetric cipher that operates on 64-bit blocks. The 256-bit level security can be even more increased by keeping the specifications of the S-boxes secret. GOST is implemented in many standard libraries such as OpenSSL and it has extremely low implementation cost and as a result of this it could be considered as a plausible alternative for AES-256 and 3-DES. Furthermore, nothing seemed to threaten its high 256-bit security [CHES 2010] and in 2010 it was submitted to ISO 18033-3 to become a worldwide industrial standard. During the period of submission many new attacks of different types were presented by the cryptographic communities against full 32-rounds of GOST. We have algebraic complexity reduction attacks, advanced differential attacks, attacks using reflection property, and many others. However, all of these attacks were against the version of GOST which uses the standard set of S-boxes. In this paper, we study the security of many variants of GOST against advanced forms of differential attacks which are based on truncated differentials techniques. In particular we present an attack against full GOST for the variant of GOST which is supposed to be the strongest one and uses the set of S-boxes proposed in ISO 18033-3. Our attack is of Depth-First key search style constructed by solving several underlying optimization problems and has time complexity $2^{245.4}$ and 2^{64} memory and data complexity. It is very interesting to note that this attack is unoptimized with respect to several aspects and can be immediately improved by discovering more efficient ad-hoc heuristics which could eventually lead to the discovery of better truncated differential properties.

1 Introduction

GOST 28147-89 encryption algorithm is the state standard of the Russian Federation and it is expected to be widely used in Russia and elsewhere [28]. It was standardized in 1989 as an official standard for the protection of confidential

T. Mourouzis (✉) • N. Courtois
University College London, London WC1E 6BT, UK
e-mail: tmourouz@cs.ucl.ac.uk; n.courtois@ucl.ac.uk

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_15

information. However, the specification of the cipher was kept confidential until 1994 when it was declassified, published [35], and translated to English [24]. According to the Russian standard, GOST is safe to be used for encrypting classified and secret information without any limitation [24]. Until 2010 most researchers would agree that despite considerable cryptanalytic efforts spent in the past 20 years, GOST is still not broken. Moreover, its large military-grade key size of 256 bits and its amazingly low implementation cost made it a plausible alternative to absolutely all standard encryption algorithms such as 3-DES or AES [28]. It appears that never in history of industrial standardization, we had such a competitive algorithm in terms of cost vs. claimed security level.

Accordingly in 2010 GOST was submitted to ISO 18033-3 to become a worldwide industrial standard. The submission has stimulated intense research and lead to the development of many interesting new cryptanalytic attacks. There are two main categories of attacks on GOST; attacks with complexity reduction which reduce the attack to an attack on a smaller number of rounds which can be solved by algebraic or software techniques at the final step [5, 6, 15], and advanced differential attacks which reduce the attack to the problem of distinguishing a certain number of rounds of GOST from a random permutation [7, 9, 12, 13].

In this paper, we present fundamental methodology for constructing general families of distinguishers on reduced-round GOST which can be eventually translated to attacks against the full 32 rounds of GOST and can be applied to all variants. By variants we mean the usage of different set of S-boxes. The design of the distinguisher is a highly nontrivial optimization problem which needs to be solved in order to be able to find a working differential attack against the complete full round cipher. Unhappily the number of potential attacks with sets of differential is very large and there is no hope to explore it systematically. In order to tackle the astronomical complexity of this task we introduce the new notion of general open setsİ, which allows us to consider similar differentials together. It is a compromise between the study of individual differentials (infeasible) and truncated differentials [21] which are already too large. Our new notion is a major refinement of truncated differential cryptanalysis of practical importance which allows for efficient discovery of better advanced differential distinguisher attacks on GOST.

The rest of this paper consists of four chapters. In the first chapter we introduce the reader to the GOST block cipher and we describe the low level design specifications of GOST. In third chapter we discuss some existing attacks on full GOST block cipher and make an introduction to the technique of differential cryptanalysis and in particular of truncated differential cryptanalysis. In fourth chapter we describe our methodology of computing transitional probabilities between truncated differentials which is similar to a Poisson process. Finally, in fifth chapter we present attacks against full 32 rounds of GOST cipher for three major variants.

2 GOST Block Cipher

GOST is a 256-bit symmetric-key block cipher that operates on 64-bit blocks and it was designed by the former Soviet Union [35]. It is an acronym for "Gosudarstvennyi Standard" or Government Standard, as translated in English [24]. This standard was given the number 28147-89 by the Government Committee for Standards of the USSR [14, 16].

GOST was developed in the 1970s and was classified as "Top Secret." In 1989, it was standardized for being used as an official standard for the protection of confidential information, but its specification remained confidential [35]. In 1990, it was downgraded to "Secret" and finally it was declassified and published in 1994, a short period after the dissolution of the USSR. Then, the standard was published and translated to English [24, 35].

According to the Russian standard, GOST is safe to be used for the encryption of secret and classified information, without any security limitation. At the beginning of the standard it states that "*GOST satisfies all cryptographic requirements and does not limit the grade of security of information to be protected*."

According to Schneier, there is no evidence that GOST was used for classified traffic, like classified military communications or if it was just used for civilian encryption. However, there are some claims which state that it was initially used for high-grade communication, including military communications [31].

It seems that GOST was considered by the Soviets as an alternative to DES but also a replacement of the rotor encryption machine FIALKA which was successfully cryptanalyzed by the Americans [31]. At the end of this chapter, we make an extensive comparative study between GOST and DES and we list all major differences. Schneier stated that designers of GOST tried to achieve a balance between efficiency and security and thus they modified the existing US DES to design an algorithm, which has a better software implementation. The same source states that the designers were not so sure of their algorithm's security and they have tried to ensure high-level security by using a large key, keeping the set of S-boxes secret and doubling the number of rounds from 16 to 32. However, it is not true that GOST was just a Soviet alternative to DES since DES is a commercial algorithm used for short-term security for encrypting unclassified documents, while GOST has a very long 256-bit key which offers military-grade security. According to Moore's law computing power doubles every 18–24 months, thus a 256-bit key cipher will remain secure for many years if no other shortcut attacks could be found assuming computing power allows to recover approximately 80-bit keys at the moment. Additionally, GOST has been shown to have a very efficient hardware implementation and this makes it a plausible alternative for AES-256 and triple DES.

A comparison among several versions of GOST and other industrial ciphers in terms of Gate Equivalence (GE) (cf. Definition 1) is presented in [28] and in Table 1.

Definition 1 ([Informal], More Details in [28]). One Gate Equivalent (GE) is equivalent to the silicon area of a 2-input NAND gate.

Table 1 The GE required for the implementation of different block ciphers

Set name	Gate equivalent
GOST-PS	651
GOST-FB	800
DES	4000
AES-128	3400
PRESENT-128	1900

As we observe from Table 1, a variant of GOST called GOST-PS, which is a fully Russian standard compliant variant (where the S-boxes of PRESENT are used) requires only 651 GE. The Russian Central Bank version called GOST-FB needs 800 GE. On the other hand, AES-128 and DES require 3400 and 4000 GE, respectively. Thus, it is not surprise the fact that it is implemented in many standard crypto libraries such as OpenSSL, Crypto++, RSA security products, and in many recent Internet Standards [16, 27].

GOST was studied by many cryptographers such as Schneier, Biham, Biryukov, Dunkelman, Wagner, and ISO cryptography experts [15, 28, 31]. All researchers always seemed to agree that it could be or should be secure, since no better way to break it except brute force was discovered. As a result of consensus among the cryptographic community, in 2010 GOST was submitted to ISO 18033 to become an international standard. Until 2010, all researchers in the cryptographic community claimed that " *Despite considerable cryptanalytic efforts spent in the past 20 years, GOST is still not broken*" [28].

Shortly after the submission, two attacks were published. One *single-key* attack against the full GOST block cipher was presented by Takanori Isobe at FSE 2011 [18]. Then, Courtois suggested a new general paradigm for effective symmetric cryptanalysis called *Algebraic Complexity Reduction* [6] and using this methodology, he constructed many more efficient attacks against GOST.

2.1 Structure of GOST

The GOST block cipher is a 32-round Feistel structure of 256-bit level security. It uses its 256-bit key to encrypt 64-bit blocks (cf. Fig. 1).

A given 64-bit block P is split into its left and right halves P_L, P_R, respectively. Given the key k_i for round i, the plaintext P is mapped to

$$(P_L, P_R) \rightarrow (P_R, P_L \oplus F_i(P_R)), \tag{1}$$

where F_i is the GOST round function. Given the round key k_i, the round function consists of the following sub-functions (cf. Fig. 2).

Firstly, the 32-bit right half is added with k_i (modulo 2^{32}). Then, the result is divided into eight 4-bit consecutive blocks and each block is given as input to a

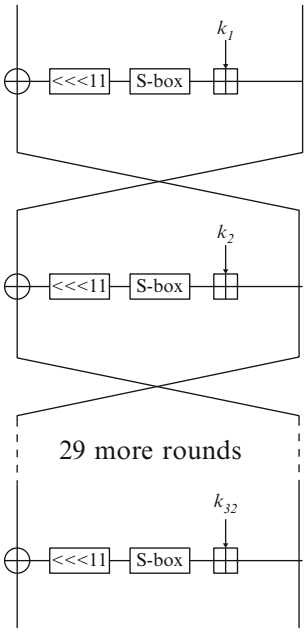


Fig. 1 The Feistel-like structure of GOST. It consists of 32 iterations of a round function which involves several bit level operations. The box S-box denotes concatenation of eight 4-bit to 4-bit S-boxes

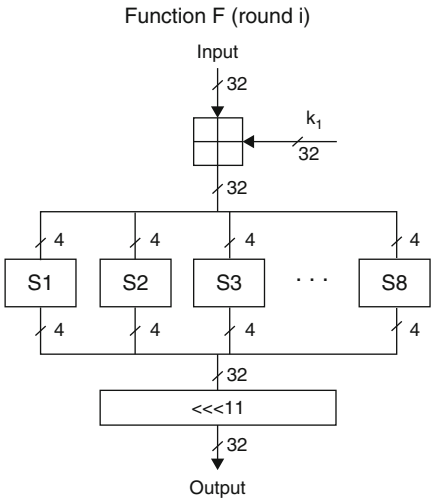


Fig. 2 Detailed description of the round function used in GOST. The initial input is initially added with the key bits modulo 2^{32}. Then, we have eight applications of 4-bit to 4-bit S-boxes and finally the 32-bit output undergoes a left rotation by 11 positions

different S-box. The first 4 bits go into the first S-box S_1, bits 5–8 go into S_2 and so on. Then, the 32-bit output undergoes a 11-bit left circular shift and finally the result is xored to the left 32-bit half of the data.

Table 2 Key Schedule Algorithm in GOST. A 256-bit word is split into 8 32-bit words

Rounds 1–8	Rounds 9–16
$k_0, k_1, k_2, k_3, k_4, k_5, k_6, k_7$	$k_0, k_1, k_2, k_3, k_4, k_5, k_6, k_7$
Rounds 17–24	Rounds 25–32
$k_0, k_1, k_2, k_3, k_4, k_5, k_6, k_7$	$k_7, k_6, k_5, k_4, k_3, k_2, k_1, k_0$

2.2 Key Schedule Algorithm

GOST has a relatively simple key schedule and this is exploited in several cryptanalytic attacks like in [5]. Its 256-bit key K is divided into eight consecutive 32-bit words $k_0, k_1, .., k_7$. These subkeys are used in this order for the first 24 rounds, while for the rounds 24–32 they are used in the reverse order (Table 2). Note that decryption is the same as encryption but with keys k_i used in the reverse order.

2.3 Addition Modulo 2^{32}

In addition to S-boxes, the GOST cipher uses addition modulo 2^{32} for key insertion. Modular addition is another source of introducing non-linearity in the cipher. There are ciphers which do not have S-boxes and the only non-linearity is via modular additions, like ARX ciphers [19]. The modular addition of two n-bit words x, y is algebraically described as follows:

$$(x, y) \rightarrow z = (x + y) \mathrm{mod} 2^n \tag{2}$$

The resulting n-bit word $(z_{n-1}, .., z_0)$ is given by,

$$\begin{cases} z_0 = x_0 + y_0 \\ z_1 = x_1 + y_1 + c_1 \\ z_2 = x_2 + y_2 + c_2 \\ . \\ . \\ z_i = x_i + y_i + c_i \\ . \\ . \\ z_{n-1} = x_{n-1} + y_{n-1} + c_{n-1} \end{cases}$$

where,

$$\begin{cases} c_1 = x_0.y_0 \\ c_2 = x_1.y_1 + c_1.(x_1 + y_1) \\ . \\ . \\ c_i = x_{i-1}.y_{i-1} + c_{i-1}(x_{i-1} + y_{i-1}) \\ . \\ . \\ c_{n-1} = x_{n-2}.y_{n-2} + c_{n-2}(x_{n-2} + y_{n-2}) \end{cases}$$

As we will explain in a later section, Multiplicative Complexity (MC), or equivalently the required number of multiplications, can be seen as a measure for the non-linearity of the cipher. The importance of MC is also discussed in [3]. The MC of the addition modulo 2^{32} is computed in Theorem 1.

Theorem 1. *The modular 2^n addition can be computed using precisely $n - 1$ multiplications. In other words its Multiplicative Complexity is $n - 1$.*

Proof. In characteristic 2 we have that

$$xy + (x + y)c = (x + c)(y + c) + c$$

Thus, we can compute the variables c_i, $1 \leq i \leq n$ using 1 multiplication for each, so $n - 1$ in total.

On the other hand, each c_i contains a multiplication of two new variables so at least one multiplication is needed per c_i.

Thus, the multiplicative complexity of this operation is exactly $n - 1$.

The existence of modular addition 2^{32} makes the study of the cipher with respect to known forms of cryptanalytic attacks such as LC and DC much more complex. We refer explicitly to DC in a later chapter.

2.4 S-Boxes and Variants of GOST

The Russian standard GOST 28147-89 does not give any recommendation regarding the generation of the S-boxes [16]. On the one hand, the fact that the S-boxes can be kept secret adds an extra security layer with approximately 354 extra bits of security (cf. *Lemma* 2). On the other hand, some problems might arise if the set of S-boxes is kept secret. For example, the generation and implementation of a set of S-boxes which is not cryptographically good would make the cipher less secure. Additionally, different algorithm implementations can use different set of S-boxes and thus can be incompatible with each other.

Even though the set of S-boxes can be kept secret, there are techniques to extract them from a chip very efficiently. We can reveal the values of the secret S-boxes by a

simple black-box chosen-key attack with approximately 2^{32} encryptions [17, 30]. In all of the attacks we describe, we assume that the S-boxes are known to the attacker.

Theorem 2. *Suppose that the 8 4-bit to 4-bit S-boxes in GOST block cipher are kept secret. Then, the effective key size becomes 610 bits.*

Proof. Each S-box is a bijective Boolean function S of the form

$$S : \mathbb{F}_2^4 \rightarrow \mathbb{F}_2^4. \tag{3}$$

Thus, each function S is a permutation on the set $\{0, 1, 2, 3, \ldots, 15\}$.

There are in total 16! such permutations.

If all 8 S-boxes are kept secret, this is equivalent of $\log_2(2^8.16!) = 354$ bits of secret information. Thus, the effective key size is increased to 610 bits from 256.

One set of S-boxes called "*id-GostR3411-94-CryptoProParamSet*," was published in 1994, as part of the Russian standard hash function specification GOST R 34.11-94. Schneier claims that this set of S-boxes is used by the Central Bank of the Russian Federation [31]. At least two sets of S-boxes have been identified as being used by two major Russian banks and institutions [31].

We are aware of the following sets of S-boxes,

1. *Gost-R-3411-94-TestParamSet*: (Table 3) This set is used by the Central Bank of the Russian Federation [31].
2. *Gost28147-TestParamSet*: This set is used when GOST is used to process large amounts of data, e.g. in CBC Mode [27].
3. *GostR3411-94-SberbankHashParamset*: This set was used by a large bank, as part of the Russian standard hash function specification GOST R 34.11-94.
4. *GostR3411-94-CryptoProParamSet*: As appearing in RFC4357, this set was published in 1994 as a part of the Russian standard hash function specification GOST R 34.11-94 [16]. It has another four versions: A, B, C, D.
5. *GOST ISO 18033-3*: This set is specified in 1WD ISO/IEC 18033-3/Amd1 and was submitted for standardization [29]. This is claimed by Russian cryptologists

Table 3 Gost-R-3411-94-TestParamSet

S-boxes	GostR3411-94-TestParamSet
S1	4,10,9,2,13,8,0,14,6,11,1,12,7,15,5,3
S2	14,11,4,12,6,13,15,10,2,3,8,1,0,7,5,9
S3	5,8,1,13,10,3,4,2,14,15,12,7,6,0,9,11
S4	7,13,10,1,0,8,9,15,14,4,6,12,11,2,5,3
S5	6,12,7,1,5,15,13,8,4,10,9,14,0,3,11,2
S6	4,11,10,0,7,2,1,13,3,6,8,5,9,12,15,14
S7	13,11,4,1,3,15,5,9,0,10,14,7,6,8,2,12
S8	1,15,13,0,5,7,10,4,9,2,3,14,6,11,8,12

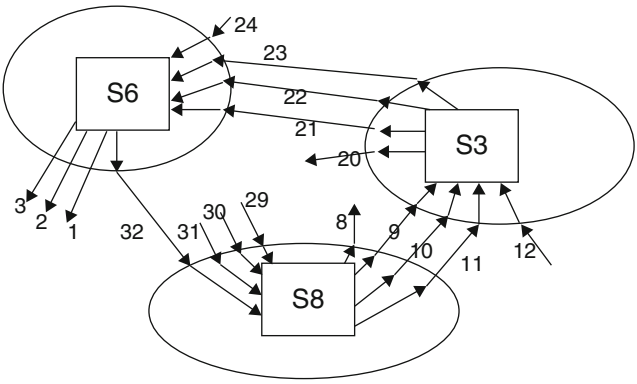


Fig. 3 Connections between GOST for 1 round

to be the most secure version to use [29]. However, in the last chapter we show that an attack faster than brute-force can be applied also to this version and there is no evidence that it is more secure.

2.5 Internal Connections in GOST

Figure 3 describes the flow of certain differences of a particular type inside GOST for one round. This structure is inherited to the left 11-bit rotation and we particularly refer to this in a later chapter.

3 Cryptanalysis of GOST

3.1 Brute-Force Attack on 256-Bit GOST Keys

Brute-force attack is a non-trivial attack in cases where the length of the key exceeds the size of the block since many false positives are expected when trying to recover the key. For example in GOST, given one (P, C) pair, we expect that $2^{256-64} = 2^{192}$ keys (out of the total 2^{256}) will satisfy $E_k(P) = C$. We can apply brute-force attack in GOST using the Depth-First search approach as follows.

Given a pair (P_1, C_1), we start testing keys $k \in K$ if they satisfy $E_k(P_1) = C_1$. During this stage, we discard a key k if it does not satisfy this relation and try a different key, otherwise we keep testing the same key k by requesting a new pair (P_2, C_2). Given this new pair, we check again if k satisfies $E_k(P_2) = C_2$. If the answer is positive, we request another distinct pair, otherwise we discard it and go again to the first stage of the attack. The first pair will reduce the space to 2^{192} possible key candidates, the second one to 2^{128} and the third one to 2^{64}. Using the

first pair, we expect that we will go through 2^{255} encryptions on average, then with the second pair we will go through 2^{127} encryptions on average and finally with the third one we will go through 2^{63} on average. Finally, a fourth pair is used to determine the key.

The expected (average) for the total time complexity in terms of GOST encryptions is given by $2^{255} + 2^{191} + 2^{127} + 2^{63} + 1 \simeq 2^{255}$.

3.2 Existing Attacks on Full GOST

Gabidulin et al. were the first who conducted a basic assessment of the security of GOST against linear and differential cryptanalysis [33, 34]. As they claim, five rounds are sufficient to secure GOST against LC at the security level of 2^{256}, while only six are enough even if the S-boxes are replaced by the Identity map. Additionally, they claim that seven rounds are sufficient for a 128-bit level security against naive DC.

Before the submission to ISO, no attack which was disputing the 256-bit level security was known. In the same year of submission, many attacks faster than brute force were developed; we have reflection attacks, attacks based on double reflections, related-key attacks, and advanced differential attacks [5–7, 10, 15, 32].

Ten years earlier, the Japanese researchers Seki and Kaneko developed an attack on 13 rounds of GOST using 2^{51} chosen plaintexts based on truncated differentials [32]. The notion of truncated differentials (partitioning type) allows us to reduce the influence of the round keys on the transitional probabilities and thus simplifies a lot the analysis. In the same paper, they have proved that naive DC always fails in GOST. This is because propagation of single differences for one round occurs with very low probability for the majority of the keys and as the number of rounds increases we expect this probability to vanish for most keys.

Isobe presented at FSE 2011 the first *single-key attack* against the full 32 rounds by developing a new attack framework called *Reflection-Meet-in-the-Middle* (RMITM) attack [18]. His method combines techniques of the reflection and the Meet-in-the-Middle attack in an optimized way. This attack has time complexity 2^{225} GOST encryptions and requires 2^{32} known plaintexts.

In parallel many other attacks based on different frameworks were developed. Courtois presented attacks based on the notion of *algebraic complexity reduction*, which allows one to reduce the problem of attacking the full cipher to a problem of attacking a reduced version of the same cipher [5]. This reduction takes into account many algebraic and structural properties of the cipher, such as the weak key schedule and the poor diffusion for limited number of rounds and makes use of software such as SAT solvers at the final solving stage for solving for the key a system that describes a reduced version (with less rounds) of the cipher [5, 6, 9].

In addition, *advanced differential attacks* were developed and successfully applied against the full block cipher. The first differential attack against full

Table 4 State-of-the-art in cryptanalysis of GOST

Author	Type	Time	Data	Scenario
Isobe [18]	RMITM	2^{224}	2^{64}	Single key
Dinur et al. [15]	2DMITM, fixed points	2^{192}	2^{64}	Single key
Courtois [5]	2DMITM, fixed points	2^{191}	2^{64}	Single key
Courtois [7]	Differential	2^{179}	2^{64}	Single key
Courtois [5]	Algebraic-differential	2^{101}	2^{32} per key	Multiple key

32-rounds of GOST was developed by Courtois and Misztal. The most complex task involved in this attack is the construction of a 20-round distinguisher. By the same year, an improved differential attack with complexity 2^{179} GOST encryptions was presented by Courtois [7].

Furthermore, Courtois studied the *multiple-key* scenario, where (P, C) pairs from randomly selected keys are available. This scenario is very realistic, as in real-life applications we expect encryptions with random keys rather than a fixed key. He proved that one such key can be revealed in approximately 2^{101} encryptions, provided that approximately 2^{32} pairs are available for each key.

Table 4 summarizes the state-of-the-art regarding cryptanalysis of full GOST for both single and multiple key scenarios. The reference point for the time complexity is the number of required GOST encryptions.

All the attacks presented so far are based on the most popular implementation of GOST, which uses the set of S-boxes *GostR3411-94-TestParamSet*. There was no attempt so far to find an attack against any other variant of GOST and provide a general methodology which would be able to work in all cases. The first who introduced such a method are Courtois and Mourouzis [11], who introduced the fundamental notion of *general open sets*, which are special forms of sets of differentials dictated by the structure of GOST and allows one to explore efficiently this space and obtain surprisingly good truncated differential properties which can be used to in some cases mount differential attacks against the full cipher. We introduce this notion in the next section.

3.3 Differential Cryptanalysis and GOST

Differential Cryptanalysis (DC) is a general form of probabilistic or statistical cryptanalytic technique that is primarily applicable to block ciphers but also to stream ciphers and cryptographic hash functions. It belongs to the category of chosen-plaintext attacks and its discovery was attributed to Eli Biham and Adi Shamir in the later 1980s, who were the first to publish a differential attack against DES [1, 2].

However, around 1994, Don Coppersmith as a member of the original IBM DES team confirmed that the technique of DC was known to IBM, as early as

1974. In addition, he said that one of the security criteria used to design DES was the resistance against this particular type of attack and this attack was known as "T-attack" or "Tickle attack" [4]. However, IBM after discussion with NSA decided to keep confidential the technique of DC as such a publication could be used against many other ciphers and cryptographic primitives that are widely used by the industry and possibly the government.

In this type of attack, the main task is to study the propagation of differences (cf. Definition 2) of inputs from round to round inside the cipher, and discover specific differences that propagate with comparatively higher probability as the probability expected assuming a uniform distribution. In this way, an attacker discovers where the cipher exhibits non-random behavior and by exploiting these properties further can recover parts of the secret key or the full key with time complexity lower than an exhaustive search on the key length which is the reference time complexity in case of block ciphers.

Definition 2 (Difference).
Let $(G, \otimes)$ be a finite abelian group with respect to the operator $\otimes$ and $x_1, x_2 \in G$ be two elements of the group. The difference between x_1, x_2 w.r.t operator $\otimes$ is defined as $\Delta x = \Delta(x_1, x_2) = x_1 \otimes x_2^1$, where x_2^1 is the inverse of x_2 with respect to $\otimes$.

In the majority of the cases in cryptanalysis, we use as operator $\otimes$ the exclusive-or operator $\oplus$ since in the greater majority of block ciphers the application of the key in the round function is a simple XOR operation. Note that any element is self-inverse with respect to XOR operation. The fact that in the greater majority the key is inserted via the XOR operation, that implies the key addition preserves the differences and this simplifies a lot our analysis. In particular, for two elements x_1 and x_2, we have that $(x_1 \oplus k) \oplus (x_2 \oplus k) = x_1 \oplus x_2$. However, in ciphers where the key is inserted via other operations, such as modular addition, as in the case of GOST which is the major subject of our study in this thesis, the application of differential cryptanalysis is not straightforward at all and many other tricks need to be employed to overcome the complexity of key insertion via other operations other than XOR.

In the rest of this section we analyze how DC can be used to obtain key bits for iterated block ciphers. Given an iterated cipher, we study the propagation of differences though different number of rounds. Then, individual differences are joined to form a differential characteristic for a larger number of rounds (cf. Definition 3). Constructing the best possible differential characteristic by combining several one-round characteristics is a non-trivial optimization task.

Definition 3 (Differential Characteristic, Fig. 4).
An s-round characteristic is an $(s + 1)$-tuple of differences $(\alpha_0, \ldots, \alpha_s)$, where α_i is the anticipated difference Δc^i after i rounds of encryption. The initial input difference $\Delta m = \Delta c^0$ is denoted by α_0.

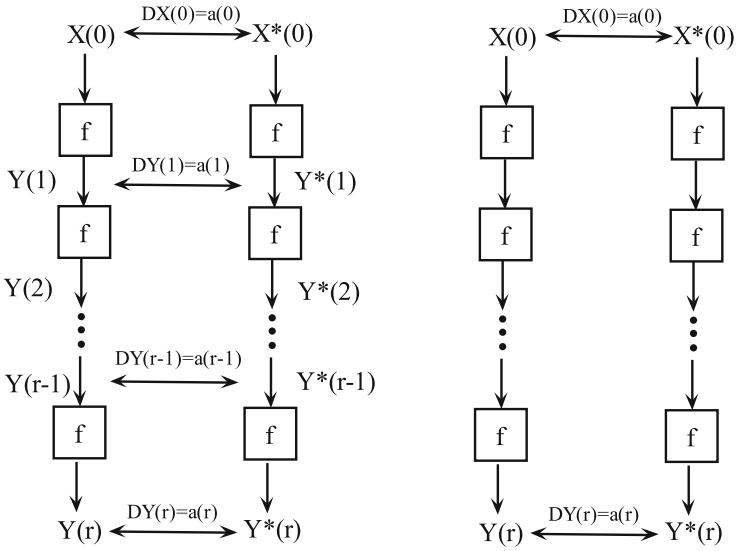


Fig. 4 A differential characteristic and a differential over r rounds

3.4 Computing the Probability of a Differential Characteristic

In differential attacks, the first task is to find series of input and output differences over several rounds, which appear with relatively high probability. For each pair of input–output difference, we need to determine the probability of propagation for each round individually. For the linear components, we can predict the propagation of the difference with probability one. However, in non-linear components, such as S-boxes, a probabilistic analysis is needed. This is a very similar task as in LC.

We call an S-box *active* if its input difference is non-zero, while we call it *inactive* or *passive* if the input difference is zero. Clearly, a zero input difference gives a zero output difference for an inactive S-box with probability 1. In the substitution layer of a cipher, S-boxes are applied in parallel to different chunks of data and thus they are independent and hence corresponding probabilities are multiplied. Another non-trivial optimization task for the attacker is to carefully select which S-boxes are taken as active in each round such that the overall differential characteristic has a relatively good probability of propagation. Many ad-hoc heuristics can be discovered by studying the structure of the rounds function of a cipher which might suggest how to select which differences are interesting to study.

In the rest of this section, we study how we can compute the probability of a differential characteristic for an iterated block cipher. Given an $(s+1)$-round characteristic $(\alpha_0 \ldots \alpha_s)$, the probability of propagation over all keys and plaintexts is given by,

$$P_{\mathscr{K},\mathscr{P}}(\Delta c^s = \alpha_s, \Delta c^{s-1} = \alpha_{s-1}, \ldots, \Delta c^1 = \alpha_1 | \Delta c^0 = \alpha_0)$$

Thus, we need to compute it on average over all key and plaintext space. This is difficult to determine for a certain class of ciphers since the model of computation does depend on the cipher. For example, in some ciphers it may be infeasible to compute it since it may depend on the key and the plaintext in a very complex way, while on some other ciphers this dependency may not be so complex and thus we may be able to enumerate all possible differential attacks. However, for the class of Markov ciphers (cf. Definition 4), this can be computed by simply computing transitional probabilities for each round and then multiplying them. The notion of a Markov cipher simplifies a lot the model of computation.

Definition 4 (Markov Cipher, [23]).
An iterated cipher round function $Y = f(X, Z)$ is a Markov cipher, if there is a group operation $\otimes$ for defining differences such that, for all choices of α ($\alpha \neq e$) and ($\beta \neq e$),

$$P(\Delta Y = \beta | \Delta X = \alpha, X = \gamma),$$

is independent of γ when the subkey Z is uniformly random.

For a Markov cipher, the probability of a one-round characteristic taken over all the key and plaintext space is independent of the plaintext space and thus it can be computed over the key space only.

Moreover, for an iterated r-round Markov cipher with r independent round keys chosen uniformly at random, the sequence of differences $\Delta c^0, .., \Delta c^r$ forms a homogeneous Markov chain (Definition 5).

Definition 5 (Markov Chain, [23]). A sequence of r random variables X_0, $X_1, .., X_r$, is called a Markov chain if

$$P(X_{i+1} = \beta_{i+1} | X_i = \beta_i, .., X_0 = \beta_0) = P(X_{i+1} = \beta_{i+1} | X_i = \beta_i),$$

for all $0 \leq i \leq r$.

Such a Markov chain is *homogeneous* if

$$P(X_{i+1} = \beta | X_i = \alpha) = P(X_i = \beta | X_{i-1} = \alpha)$$

Thus, the probability of a s-round characteristic for a Markov cipher with independent round keys can be computed as follows [22].

$$P(\Delta c^s = \alpha_s, \ldots, \Delta c^1 = \alpha_1 | \Delta c^0 = \alpha_0) = \prod_{1 \leq i \leq s} P(\Delta c^i = \alpha_i, \Delta c^{i-1} = \alpha_{i-1}) \tag{4}$$

3.5 Differentials vs. Differential Characteristics

An adversary does not have so much freedom to determine if the input difference follows a given differential characteristic in each step. However, he can choose the input difference and may be able to check the corresponding output difference after s rounds. An s-round characteristic is constructed by concatenating s one-round differentials.

In practice, it is very time consuming to find a really good differential characteristic over a sufficient number of rounds. The collection of all s-round characteristics with input α_0 and output difference α_s is called a differential (Definition 6).

Definition 6 (Differential, [22]).
An s-round differential is a pair of differences (α_0, α_s), also denoted as $\alpha_0 \rightarrow \alpha_s$, where α_0 is the chosen input difference and α_s the expected output difference Δc^s

Given an s-round differential (α_0, α_s), the probability of such differential on average over key space and all plaintexts is given by,

$$P_{K,\mathscr{P}}(\Delta c^s = \alpha_s | \Delta c^0 = \alpha_0) = \Sigma_{\alpha_1} \ldots \Sigma_{a_{s-1}} P_{K,\mathscr{P}}(\Delta c^s = \alpha_s, \ldots, \Delta c^1 = \alpha_1 | \Delta c^0 = \alpha_0)$$

In an attack the key is fixed and only the plaintext can be variable. Thus, in practice we may need to compute it over a fixed key which is not known to the attacker. Computing the following probability is enough to mount many cryptographic attacks,

$$P_{\mathscr{P}}(\Delta c^i = \alpha_i | \Delta c^{i-1} = \alpha_0, K = k).$$

However, the key is unknown, and thus we cannot compute this probability unless we consider the assumption of stochastic equivalence (Definition 7).

Definition 7 (Hypothesis of Stochastic Equivalence).
Consider an r-round iterated cipher, then for all highly probable differentials, $s \leq r$, (α, β),

$$P_{\mathscr{P}}(\Delta c^s = \beta | \Delta c^0 = \alpha, K = k) = P_{\mathscr{P},\mathscr{K}}(\Delta c^s = \beta | \Delta c^0 = \alpha),$$

holds for a substantial fraction of the key space $\mathscr{K}$.

3.6 Key Recovery Attacks

In this section, we describe how to derive some key bits using differential attacks. Consider a differential (α, β) over $r-1$ rounds, which holds with probability p for a r-round iterated block cipher. By partial decryption of the last round, we can recover some bits of the last key faster than brute-force.

Firstly, we encrypt N pairs of plaintexts (P, P'), such that $\Delta P = \alpha$ and get the corresponding ciphertext pairs (C, C'). Given these pairs, we guess some bits of the last round key and we partially decrypt the last round. Then, we check if the difference after $r-1$ rounds is obtained. If this difference is obtained we say that (P, P') suggests a candidate k_G. We expect approximately $p.N$ pairs to result in pairs with difference β in round before the last round. Such pairs are called right pairs (cf. Definition 8).

Definition 8 (Right Pair).
A pair (P, P') with $\Delta P = \alpha$ and associated ciphertexts (C, C') is called a *right pair* with respect to the $(r-1)$-round differential (α, β) if $\Delta c^{r-1} = \beta$. Otherwise, it is called a *wrong pair*.

In order to launch a successful differential attack, we need at least one right pair. First, in an attack we want to identify a right pair. However, we have also wrong pairs that do not follow our constructed characteristic and this is referred to as noise, while right pairs is the signal. Thus, wrong pairs should be filtered in a very early stage of our attack if it is possible. Often wrong pairs can be eliminated by considering the associated ciphertexts. This process is called *filtering*. Note that there are no general rules how to perform the filtering step and it depends on the cipher.

Algorithm below describes an attack on an r-round iterated block cipher using an $r-1$ differential characteristic. This attack can be used to obtain some bits of the last round key using a differential characteristic of the form $(\alpha_0, \ldots, \alpha_{r-1})$, which holds with probability p.

1. Let T_j a counter for (parts of) possible last round key guesses k_j
2. For $i = 1, ..N$ do
   (a) Choose P_i at random and compute $P'_i = P_i \oplus \alpha_0$. Obtain the corresponding ciphertexts (C_i, C'_i).
   (b) Use filtering. If (P_i, P'_i) is a wrong pair, discard it and continue with the next iteration. Otherwise do the following.
   (c) For each key guess k_j, partly decrypt the last round and get $(c_i^{(r-1)}, c_i^{'(r-1)})$
      i. Increase T_j, if $c_i^{(r-1)} \oplus c_i^{'(r-1)} = \alpha_{r-1}$
3. Find l, such that $T_l = \max_i(T_i)$
4. Return k_l as the guess for the correct key

In most cases after applying the method of DC, one pair might suggest several key candidates $\{k_1, k_2, \ldots, k_l\}$. On the contrary, a wrong pair is expected to suggest a set of candidates which do not include the correct key. The attack is successful, if the correct key value is suggested significantly more often than the other candidates. This is expected for a differential of probability p if approximately $\frac{c}{p}$ plaintexts are selected uniformly at random, where c a small constant depending on the cipher [22]. In the rest of this section, we discuss some advanced forms of DC, such as truncated and impossible differentials.

3.7 Truncated Differentials and GOST

Truncated Differential Cryptanalysis is a generalization of differential cryptanalysis developed by Lars Knudsen [21]. Usually, in DC we study the propagation of single differences between two plaintexts, while in truncated DC we consider differences that are partially determined (i.e. we are interested only in some parts of the difference). This technique has been successfully applied to many block ciphers such as SAFER, IDEA, Skipjack, Twofish, and many others. We define the truncation $TRUNC(a)$ of an n-bit string a as in Definition 9.

Definition 9 (Truncation, [21]).
Let $a = a_0a_1 \dots a_{n-1}$ be an n-bit string, then its truncation is the n-bit string b given by $b_0b_1..b_{n-1} = TRUNC(a_0a_1..a_{n-1})$, where either $b_i = a_i$ or $b_i = *$, for all $0 \leq i \leq n-1$ and $*$ is an unknown value

The notion of truncated differentials (cf. Definition 10) extends naturally to differences.

Definition 10 (Truncated Differentials, [21]).
Let (α, β) be an i-round differential, then if α' and β' are subsequences of α and β, respectively, then (α', β') is an i-round truncated differential.

For example, the truncated differential on 8 bytes of the form $0000000000 * 00000$, where $* = x_1x_2x_3x_4$, is the set of differences of size $16-1$ (excluding zero difference).

Given an s-round characteristic $\Delta_0 \to \Delta_1 \to \dots \to \Delta_s$, then $\Delta'_0 \to \Delta'_1 \to \dots \to \Delta'_s$ is a truncated characteristic, if $\Delta'_i = TRUNC(\Delta_i)$ for $0 \leq i \leq s$. A truncated characteristic predicts only part of the difference in a pair of texts after each round of encryption. A truncated differential is a collection of truncated characteristics. Truncated differentials proved to be a very useful cryptanalytic tool against many block ciphers which at first glance seem secure against basic differential cryptanalysis.

3.7.1 General Open Sets

As we have already mentioned, truncated differential cryptanalysis is a generalization of differential cryptanalysis against block ciphers developed by Lars Knudsen in 1984 [20]. The basic idea is to consider propagation of sets of differences instead of single differences and thus the attack works on making predictions of only some of the difference bits instead of a full block. Intuitively, in this way we hope to succeed in finding sets of differences which propagate with a comparatively higher probability than in the case of a random permutation by "gluing" differences together.

Even though truncated differential cryptanalysis was successfully applied in ciphers where naive differential cryptanalysis failed, like SAFER, IDEA, Skipjack,

E2, Twofish, Camellia, CRYPTON and even the stream cipher Salsa20, the exploration of the space of truncated differentials is computationally a very big overhead since the space is exponentially large. In order to speed up the process of discovery of interesting propagations in the space of truncated differentials we need to discover some ad-hoc heuristics suggested by the structure of the particular cipher we study which capture this structure and lead to the discovery of propagations we high bias.

In the case of GOST block cipher, we observe that the rotation by 11 bits to the left of the output bits from the S-boxes is enough to describe the connections between round to round inside the cipher.

In this section, we introduce a new type of sets of differences, which we name *general open sets* and are dictated by the structure of the GOST cipher. They can be seen as a refinement of Knudsen's truncated differentials [21]. The main difficulty in attacks using truncated differentials is the exploration of the exponentially large space of possible sets of differences and how to discover interesting truncated differential properties.

However, if we consider special sets which are dictated by the structure of the encryption algorithm that we study we may be able to explore this subspace and discover interesting properties. We follow this idea in case of GOST and we consider some special sets, which we name *general open sets* (cf. Definition 11) and these sets are constructed based on the connections between the S-boxes from round to round.

Definition 11 (General Open Sets, [11]).

We define a General Open Set X as a set of differences on 64 bits with additional constraints as follows. A General Open Set is represented by a string Q of 16 characters on the alphabet $\{0, 7, 8, F\}$ in the following way:

1. differences in X are "under" Q, by which we mean that for all $x \in X$ $\mathrm{Sup}(x) \subseteq \mathrm{Sup}(Q)$, where $\mathrm{Sup}(x)$ is the set of bits at 1 in x, $\mathrm{Sup}(x) \subset \{0, 1, \ldots., 63\}$.
2. AND in each of the up to 16 non-zero characters in string Q which may be any of $7, 8, F$, there is at least one "active" bit at 1 in x for all $x \in X$.
3. In the case of F the most significant bit is always active for each $x \in X$ and for each position in Q which is at F.

The main reason why we have this very special alphabet $\{0, 7, 8, F\}$ is the internal connections of the GOST cipher and in particular the 11-bit rotation to the left after the substitution layer. Informally, we can say that we group together bits which are likely to be flipped together. F is used to make the sets disjoint such that each difference x belongs to only one general open set.

Given a general open set represented by the string Q, we define the closure of this set as in Definition 12.

Definition 12 (Closure of Differential Sets).

The closure of a differential set Q is denoted by $[Q]$ and it is the set that contains all the differences that are under the string Q with the only rule to exclude the zero difference on the 64 bits. A set P such that $P = [Q]$ will be called a closed set.

Consider the general open set represented by 8070070080700700. Then, this set contains in total $(2^3 - 1)^4$. The closure of Q, denoted by $[Q]$, contains $2^{14} - 1$ elements. This is because due to Definitions 11 and 12, in general open set 8 can be only 1000 and 7 any difference in the set $\{0111, 0100, 0010, 0001, 0110, 0101, 0011\}$, while in case of a closed set 8 can be any element in the set $\{0000, 1000\}$ and the character 7 can be any element in the set $\{0000, 0111, 0100, 0010, 0001, 0110, 0101, 0011\}$ provided that the zero difference on 64-bits is excluded.

For example, in Fig. 3, we illustrate the connections for the study of truncated differential or closed set [8070070080700700]. We observe that the 3 least significant bits from S_3 are entering S_6 and this is denoted by 7 in the differential, while the most significant bit from S_6 is entering S_8 and this is denoted by 8.

It is a non-trivial task to define such sets in general since some heuristics suggested by the structure of the algorithm need to be discovered. Note that the same idea can be applied to any cipher. In the next section, we study the diffusion inside GOST aiming to illustrate that in particular for the first eight rounds the diffusion is really poor.

4 Propagation of Differentials in GOST

In order to compute the probability of a transition we use a simple Algorithm which just runs a large number of events and counts the events of our interest until the probability of transition converges up to some desired precision. We assume that the distribution of the number of events of our interest follows (approximately) a Poisson distribution. We use this distribution as we have experimentally observed that for all cases we have tried,

- We have a discrete distribution of small integers
- In all cases we have tried and are included in this thesis the variance is relatively close to the mean. The Poisson distribution has a variance equal to the mean.

Thus, for a sample of size N if x denotes the number of events that were observed (approximated by Poisson with parameter Np where p is the true mean), then the approximated Standard Deviation (SD) of the variable $\frac{x}{N}$, where N is assumed to be constant and p' the observed mean, is given by $\frac{\sqrt{Np'}}{N} = \sqrt{\frac{p'}{N}}$. This is because the variance equals to the mean in case of a Poisson distribution.

Denote by p' the approximated mean, then $SD' = \sqrt{\frac{p'}{N}}$. Then, for example with about 99 % confidence interval, the true mean is within $\pm 3.\sqrt{\frac{p'}{N}}$ of the observed mean.

Let I_1 be the interval $[p' - t\sqrt{\frac{p'}{N}}, p' + t\sqrt{\frac{p'}{N}}]$. In our simulations we would like this interval I_1 to be contained in the interval $I_2 = [p'.2^{-a}, p'.2^a]$, where a is an error we allow in the exponent of the mean as a power of 2.

Table 5 Time taken to compute the mean of a Poisson process up to some precision

Probability	Number of rounds	Time taken
$2^{-13.8}$	4	2 s
$2^{-16.5}$	6	15 s
$2^{-24.0}$	8	2.3 h
$2^{-24.0}$	10	2.8 h
$2^{-25.0}$	8	2.3 h
$2^{-25.0}$	10	2.8 h

We assume that the true mean that we are aiming to approximate by simulations is bigger than some probability value p_0 (for example $2^{-26.0}$) in order to ensure that the algorithm terminates in reasonable time. The inclusion of sets $I_1 \subseteq I_2$ implies that we need to run about $N > N_0$ simulations, where N_0 is given by

$$N_0 = (\frac{(2^a)t}{(2^a - 1)})^2 . \frac{1}{p_0} \tag{5}$$

in order to compute an approximated mean with the desired precision. For smaller values of probabilities we need to use different values for parameters a, t such that it is computationally feasible to run beyond this bound. Most of the later results which are less than approximately $2^{-26.0}$ are inexact results and were taken by setting $a = 0.3$ and $t = 5$ in most cases.

In an Intel i7 1.73 GHz PC with 4.00 GB RAM computer, we can run around 2^{22} full GOST encryptions per second per CPU. For probabilities above $2^{-26.0}$ we set $t = 3$ and $a = 0.1$, while for smaller probabilities we allow a to be around 0.3 or even higher and thus the results are inexact. In Table 5 we present the time taken to compute some probabilities that are presented in this thesis with some precision $a = 0.1$ and $t = 3$.

4.1 Statistical Distinguishers

In cryptanalysis, we very often study the problem of distinguishing distributions, one distribution that describes the variable of the number of certain events that occur at random and another distribution that describes the same variable but due to propagation inside the cipher. Thus, we would like to design a clever distinguisher which would be able to distinguish a given a cipher from a random permutation by capturing as much as possible of its mathematical structure. Such a distinguishing attack might reveal information which can be used to reduce the space of the key candidates and thus lead to an attack faster than exhaustive search. In cryptographic literature, there are several examples of successful attacks against either the full block cipher or some reduced-round version or more frequently against stream ciphers such as in [12, 13, 25, 26] where distinguishing attacks against block cipher GOST and stream cipher RC4 are described.

Thus, this can be seen as a hypothesis testing problem of distinguishing the two distributions as shown in Fig. 5. Suppose that a source is used to generate independent random samples in some given finite set with some distribution $\mathscr{P}$, which is either $\mathscr{P} = \mathscr{P}_0$ or $\mathscr{P} = \mathscr{P}_1$. A distinguisher is a construction used to determine which one is the most likely the one which was used to generate the sample. Hence, the overall attack based on the distinguishers considers the following underlying statistical hypothesis testing problem, where we have either a null hypothesis $H_0 : \mathscr{P} = \mathscr{P}_0$ or an alternative hypothesis $H_1 : \mathscr{P} = \mathscr{P}_1$.

Our scope is to study this hypothesis problem applied to differential cryptanalysis and its variants. The variable of our interest is the number of plaintext pairs whose output difference after r rounds lies in a particular truncated differential set ΔY given that their difference lies in another truncated differential set ΔX. We aim to use particular sets of differences which capture the mathematical structure of the cipher and these are known as general open sets and we described them in the previous section.

Assuming that we have two random variables $\mathscr{W}$ and $\mathscr{R}$ which are described by Gaussian distributions with parameters $(E(W), V(W))$, and $(E(R), V(R))$, respectively. Our task is given a measurement of the variable of our interest to determine from which distribution this sample is more likely to be taken. Thus, we have the following hypothesis testing problem, $H_0 : P = \mathscr{W}$ and $H_1 : P = \mathscr{R}$. For cryptanalytic purposes, we assume that distribution $\mathscr{W}$ corresponds to a wrong key, while $\mathscr{R}$ corresponds to the right key. In case of a Gaussian distribution, the probability density function of distribution $\mathscr{W}$ is given by the following equation,

$$f_{\mathscr{W}}(x) = \frac{1}{\sqrt{2\pi V(W)}} \exp^{-\frac{1}{2V(W)}(x-E(W))^2}. \tag{6}$$

Assume that we were given a sample P from which we can observe x events of our interest, in the particular case of differential cryptanalysis is the number of pairs which follow the differential $\alpha \to \beta$ after r rounds. Then, from Fig. 5 we observe that if x is greater than $E(R)$ then we can assume that this observation corresponds to the right key with probability set to $\frac{1}{2}$. On the other hand, the probability of a false positive, for example accepting the key as correct while it is wrong, which is also known as Type I error, is represented by the red-shaded region in Fig. 5 and given by the following formulae,

$$P(\mathscr{W} > \mathscr{R}) = \int_{E(R)}^{\infty} f_{\mathscr{W}}(x)dx = \frac{1}{2}(1 - \text{erf}(\frac{E(R) - E(W)}{\sqrt{2V(W)}}) \tag{7}$$

where $\text{erf}(x)$ is the Gaussian error function given by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2} dt. \tag{8}$$

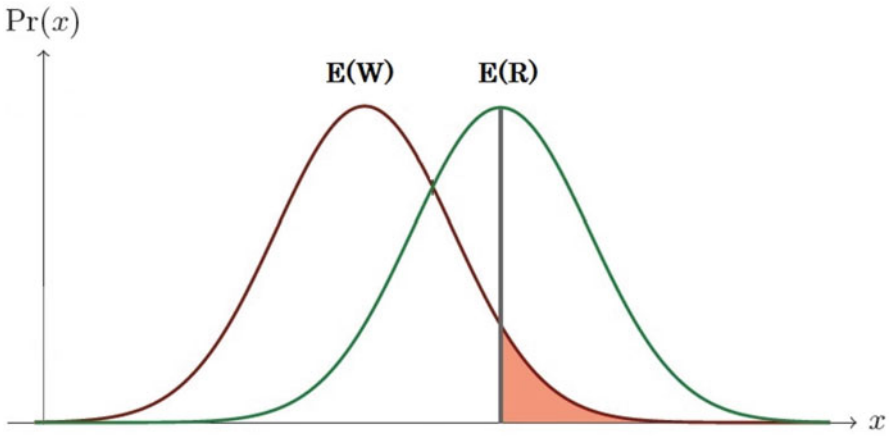


Fig. 5 Computation of advantage represented by the *red-shaded area* which represents the probability of Type I error for distinguishing the distributions $\mathscr{W}$ and $\mathscr{R}$

In the next section, following precisely this idea we construct distinguishers which allow us to distinguish 20 rounds of GOST from a random permutation and use this construction to launch attacks against full 32 rounds of the cipher. The basic idea behind our constructions is to distinguish the following two scenarios:

1. Propagation: 20 rounds of GOST with particular intermediate differences
2. Natural Occurrence: It could be 20 rounds of GOST, or more rounds of GOST or a random permutation

4.2 20-Round Distinguishers in GOST

Following the idea of the previous section we construct 20 round distinguishers for three variants of GOST.

Theorem 3 (20-Round Distinguisher on GOSTR3411-94-TestParamSet).

$$8780070780707000$$
$$\downarrow (10R)$$
$$[8070070080700700]$$
$$\downarrow (10R)$$
$$80707000087800707$$

is a 20 rounds distinguisher where [8070070080700700] *is a closed set, and satisfies the following properties,*

1. *If the 20 rounds are replaced by a random permutation, then out of the total of* 2^{77} *pairs of plaintexts* (P_i, P_j) *such that* $P_i \oplus P_j \in$ 8780070780707000*, we expect on average* $2^{27.1}$ *to satisfy also the output difference at the end of the 20 rounds.*
2. *Among all input pairs with input difference in the set* 8780070780707000*, we expect on average* $2^{18.1} + 2^{27.1}$ *after 20 rounds to follow the differential characteristic 10+10 shown above.*
3. *The advantage of the distinguisher is 25.8 standard deviations*

Proof. For a typical permutation on 64 bits (which does not have to be a random permutation, it can be GOST with more rounds) out of total 2^{77} plaintext pairs (P_i, P_j) which satisfy the specified input difference, we expect on average $2^{27.1}$ such pairs to satisfy also the output difference after 20 rounds. The distribution of the expected number of pairs which satisfy both input and output difference is approximated by a Normal distribution $\mathcal{N}(2^{27.1}, 2^{13.55})$.

For 20 rounds of GOST and for a given random key, we expect such pairs to occur both by accident (naturally occurring as in a random permutation) and due to propagation in GOST. This is because the length of the key is larger than the size of the block.

Let $\mathscr{X}$ denote the distribution of expected number of pairs occurring naturally and $\mathscr{Z}$ the distribution of expected number of pairs occurring due to propagation. By computer simulations, we have obtained the probability of the following transition

$$8780070780707000 \rightarrow [8070070080700700]$$

after 10 rounds of GOST and was found approximately equal to $2^{-29.4}$ (this result is inexact). That implies that the mean of the distribution $\mathscr{Z}$ is $2^{18.2}$.

In case of a random permutation, the expected number of pairs which have this additional middle difference is $2^{27.1-29.4-29.4} = 2^{-31.7}$ (no pairs in practice). Thus, this middle difference property can be seen as an artificial assumption which separates the two sets.

Hence, the distribution $\mathscr{Y} = \mathscr{X} + \mathscr{Z}$ has mean approximately $2^{27.1} + 2^{18.2}$. Thus, the advantage of the distinguisher is given by $\frac{2^{18.2}}{2^{13.55}}$, which is approximately 25.8 standard deviations.

If we set $2^{27.1} + 2^{18.2}$ as a threshold to accept the key as correct, then the guess is correct with probability set at $\frac{1}{2}$. The probability of a false positive (Type I error) is given by,

$$P(\mathscr{Y} > 2^{27.1} + 2^{18.1}) = \frac{1}{2}(1 - \mathrm{erf}(\frac{25.8}{\sqrt{2}})) \simeq 2^{-485}$$

Theorem 4 (20-Round Distinguisher on GOST28147-CryptoProParamSetA).

$$0770070077777770$$
$$\downarrow (10R)$$
$$[7007070070070700]$$
$$\downarrow (10R)$$
$$7777777007700700$$

is a 20 rounds distinguisher where $[7007070070070700]$ *is a closed set, and satisfies the following properties,*

1. *If 20 rounds are replaced by a random permutation, we expect on average* $2^{55.1}$ *to satisfy both input-output differences after 20 rounds.*
2. *Among all input pairs with input difference in the set* 0770070077777770*, we expect on average* $2^{33.0} + 2^{55.1}$ *after 20 rounds to follow the differential characteristic*
3. *The advantage of the distinguisher is 42.2 standard deviations*

Proof. For a typical permutation on 64 bits out of the total plaintext pairs (P_i, P_j) with this input difference we expect $2^{55.1}$ such pairs to satisfy also the desired output difference. The distribution of the expected number of pairs is approximated by a Normal distribution of the form $\mathcal{N}(2^{55.1}, 2^{27.55})$.

We have computed the probability of transition

$$[7007070070070700] \rightarrow 7777777007700700$$

and found to be approximately equal to $2^{-24.01}$ after ten rounds.

Again the two sets are entirely disjoint for same reasons explained in the previous theorem.

Thus, the distribution $\mathcal{Y} = \mathcal{X} + \mathcal{Z}$ has mean $2^{55.1} + 2^{33.0}$. The advantage of the distinguisher is approximately 42.24 standard deviations and corresponds to Type I error 2^{-1290}.

Theorem 5 (20-Round Distinguisher on GOST ISO 18033-3).

$$8000070770700000$$
$$\downarrow (6R)$$
$$[7078000070000700]$$
$$\downarrow (8R)$$
$$[7000070070780000]$$
$$\downarrow (6R)$$

7070000080000707

is a 20 rounds distinguisher where [7000070070780000] *is a closed set, and satisfies the following properties,*

1. *If 20 rounds are replaced by a random permutation, we expect on average* $2^{21.5}$ *to satisfy both input-output differences after 20 rounds.*
2. *Among all input pairs with input difference in the set* 8000070770700000*, we expect on average* $2^{15.9} + 2^{21.5}$ *after 20 rounds to follow the differential characteristic*
3. *The advantage of the distinguisher is 35.09 standard deviations*

Proof. For a typical permutation on 64 bits, we have the distribution $\mathcal{N}(2^{21.5}, 2^{10.75})$. By computer simulations we have obtained the following transitional probabilities after six and eight rounds, respectively,

$$P([7078000070000700] \rightarrow 8000070770700000) = 2^{-16.47}$$

$$P([7007070070070700] \rightarrow [7000070070780000]) = 2^{-27.20}$$

Hence, out of the total 2^{77} pairs with the input difference as specified in the 20-round construction, we expect approximately $2^{77-17.47-27.20-16.47} = 2^{15.9}$ (The size of the set 8000070770700000 is half the size of the set [7078000070000700] and thus we can assume that the probability is halved in the reverse direction). Thus, the mean of the distribution $\mathscr{Z}$ (due to propagation) is $2^{15.9}$.

In case of a random permutation, the expected number of pairs which have in addition this specific middle difference is $2^{21.5-17.47-16.47} = 2^{-12.44}$ (no pairs in practice).

Thus, the distribution $\mathscr{Y} = \mathscr{X} + \mathscr{Z}$ has mean $2^{15.9} + 2^{21.5}$. The advantage of the distinguisher is given by $\frac{2^{15.9}}{2^{10.75}}$, which is approximately 35.09 standard deviations and this corresponds to Type I error 2^{-894}.

5 Parametric Attacks against Full GOST

In this chapter, we present attacks against full 32 rounds of GOST by using the 20-round distinguisher constructions described in the previous chapter. Our attack in order to succeed takes into account several optimizations related to low-level structure of GOST. Theorem 6 summarizes our results.

Theorem 6 (20-R Distinguisher $X_i \rightarrow X_j$, Transitions $X'_i \rightarrow X_i$ and $X_j \rightarrow X'_j$ for $6 - x$ Rounds).

1. *For each guess of the k key bits for the first x ($x \leq 5$) rounds, do the following steps.*
2. *For all* 2^{64} *pairs* (P_l, C_l) *(full 32-R):*

Compute $P'_l = G_{x,k}(P_l)$ *and* $C'_l = G_{x,k}(C_l)$*, where* $G_{x,k}$ *is the encryption for the first* x *rounds using key* k *(which is the same for the last* x *rounds due to the Key Schedule). At this step we have computed all the* (P'_l, C'_l) *for the middle* $32 - 2x$ *rounds.*
Store a list of $(32 - 2x)$*-round* (P'_l, C'_l) *pairs in a hash table, sorted by their* $128 - \log_2(|X'_i|) - \log_2(|X'_j|)$ *inactive bits. While we are computing a* (P', C') *pair for the middle* $(32 - 2x)$ *rounds, we check if for a new pair computed we have a collision on the inactive bits. If such a collision is found, this corresponds to pair of plaintexts* (P'_l, P'_m) *such that* $P'_l \oplus P'_m \in X'_i$ *and* $C'_l \oplus C'_m \in X'_j$ *after* $(32-2x)$ *rounds (Because we do it for fixed number of rounds which is 20 rounds we assume the complexity of hash table construction is constant).*
This list requires memory of about $2^{64} \times 64 = 2^{70}$ *bits.*
The time complexity of this step in terms of GOST encryptions is

$$T_1(x) = 2^{32x} \times 2^{64} \times \frac{2x}{32} \simeq 2^{60+32x+\log_2(x)} \tag{9}$$

and it returns about $\frac{|X'_i| \times |X'_j|}{2}$ *triples* $(k, (P'_i, C'_i), (P'_j, C_j))$.

3. *For the total of* $\frac{|X'_i| \times |X'_j|}{2}$ *collisions of the form* $((P'_m, C'_m), (P'_n, C'_n))$ *which have been computed in the previous step, we want to count the number of pairs, which satisfy both input and output difference as specified by the middle 20-R distinguisher. Let* T *be the number of such pairs which satisfy the required constrained imposed by the distinguisher.*
We compute T *by guessing the remaining* $192 - 32x$ *bits for the remaining* $6 - x$ *rounds and each time the new pair* $((P''_m, C''_m), (P''_n, C''_n))$ *for the middle 20 rounds satisfy the required property we increase the counter by 1. This has time complexity in terms of GOST encryptions given by,*

$$T_2(x) = 2^{32x} \times 2^{32(6-x)} \times \frac{|X'_i| \times |X'_j|}{2} \times \frac{12 - 2x}{32} \simeq 2^{187+\log_2(6-x)+\log_2|X'_i|+\log_2|X'_j|} \tag{10}$$

If the counter $T > c$*, then we accept the 192-bit key assumption as correct, otherwise we reject it.*

4. *If the Type I error equals to* 2^{-y}*, this implies that we are left with approximately* 2^{192-y} *possible key candidates on the 192 bits of the key. The remaining* $256 - 192 = 64$ *can be found using additional pairs for the full 32-rounds.*
The complexity of this step is given by,

$$T_3(y) = 2^{192-y+64} = 2^{256-y} \tag{11}$$

The overall time complexity C_T *(in terms of GOST encryptions) is given by,*

$$C_T = 2 \times (T_1(x) + T_2(x) + T_3(y)) \tag{12}$$

since the Type II error is set to $\frac{1}{2}$.

Table 6 Best 1-round transitions in absolute value between general open sets for the three variants of GOST of our interest

Set	X_j	X'_j	$p(X_j \to X'_j)$	ADV_{filter}
TestParamSet	8070700087800707	8780070780787777	$2^{-5.34}$	0.6
CryptoParamSet	7777777007700700	07700700F77F7777	$2^{-3.73}$	3.2
ISO	7070000080000707	8000070770787780	$2^{-3.27}$	3.6

Table 7 T_1, T_2, T_3, C_T values in terms of GOST encryptions

Set	x	T_1	T_2	T_3	C_T
TestParamSet	5	$2^{222.3}$	$2^{231.9}$	$2^{255.1}$	$2^{256.1}$
CryptoParamSet	5	$2^{222.3}$	$2^{248.8}$	$2^{252.2}$	$2^{253.2}$
ISO	5	$2^{222.3}$	$2^{220.7}$	$2^{244.4}$	$2^{245.4}$

In the rest of this section we study the three variants of GOST of our interest. Using computer simulations we have computed some sufficiently good propagations which can be used in the filtering step for extending the 20-round distinguisher to a 22-round filter. Filtering which will allow us to gain four rounds was not achieved so far by our methodology. Table 6 presents our best results found so far by our heuristic discovery method.

Based on these transitions we have computed the associated Type I error for each of the three cases and they are found to be $2^{-0.9}$, $2^{-9.51}$, and $2^{-11.62}$, respectively. Table 7 presents the complexity for each step of our attack and the complexity of the overall attack for each variant of GOST.

As we observe from Table 7, the attack is not good against the GOST variant which uses the set of S-boxes TestParamSet, since its complexity exceeds brute-force. However, there are already plenty of attacks on this variant [5, 7, 8]. Using our technique, we can break the other two variants of GOST which use the sets CryptoParamSet and ISO in time complexity approximately $2^{253.2}$ (slightly faster than brute-force but not significantly) and $2^{245.4}$ GOST encryptions, respectively. The ISO version was supposed to be the strongest one and was proposed for standardization.

5.1 Conclusions and Further Research

GOST is an important government and industrial block cipher with a 256-bit key which is widely used implemented in standard crypto libraries such as OpenSSL and Crypto++ [28]. Several attacks on GOST have been found since 2010, the best of which are advanced differential attacks in which the main problem for the attacker is the design of an effective distinguisher for some 20 Rounds of GOST. In this paper we have proposed a methodology which allows for efficient discovery of İgoodİ attacks of this type. In order to achieve this we have introduced a fundamental notion of general open sets, which are special sets which are dictated by

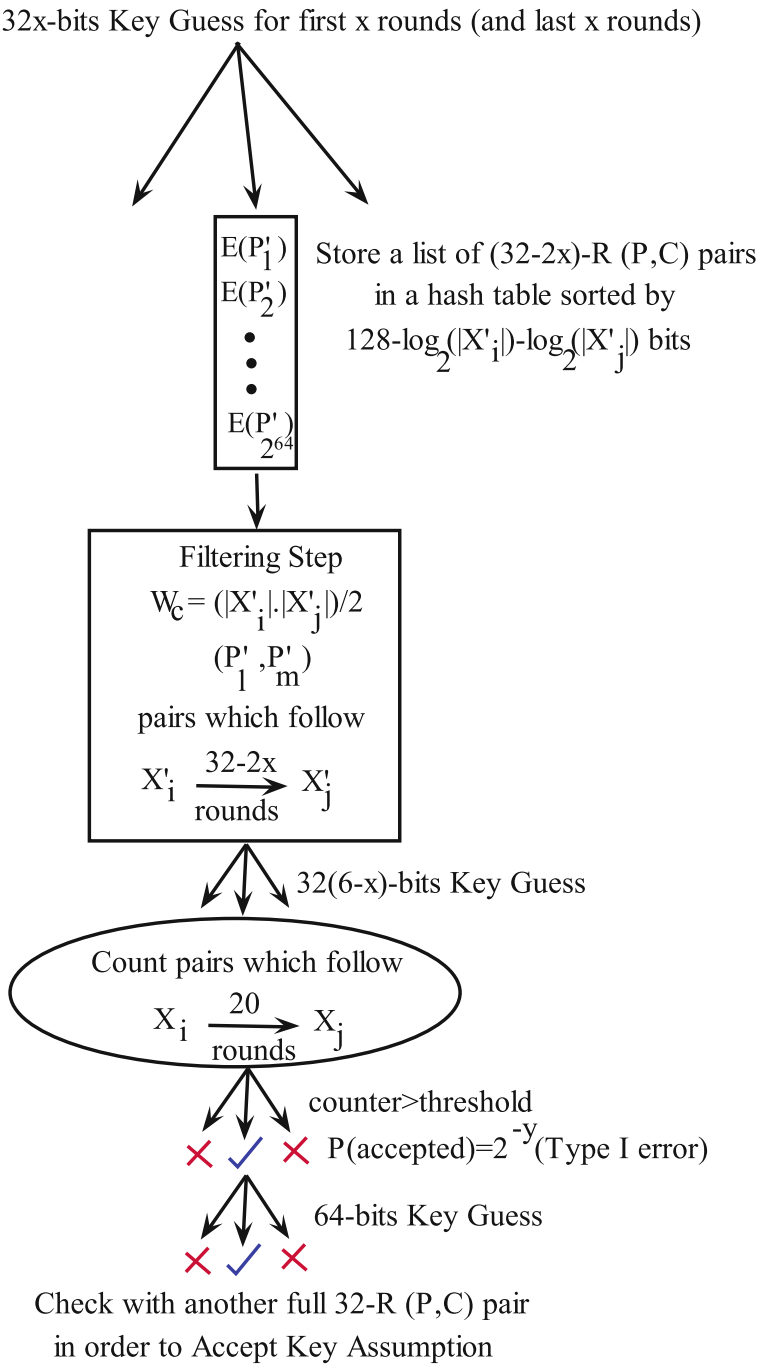


Fig. 6 Parametric Attack on Full GOST. A parametric attack against full GOST which is essentially a Depth-first search approach combined with an additional filtering step

the structure of GOST and by specific patterns which dominated earlier attacks and which allow to refine them. Then, we have developed a method to construct complex differential distinguishers for more rounds as a combination of disjoint paths. Our methodology is validated by the construction of very good distinguishers for 20 rounds for two variants of GOST; GostR3411-94-TestParamSet, and Gost28147-CryptoProParamSet which are more powerful than expected. Then, we convert these 20-round constructions to full attacks against full 32 rounds of GOST.

References

1. Biham, E., Shamir, A.: Differential cryptanalysis of the full 16-round des. In: Brickel, E.F. (ed.) CRYPTO 1992. Lecture Notes in Computer Science, vol. 740, pp. 487–496. Springer, Heidelberg (1992)
2. Biham, E., Shamir, A.: Differential Cryptanalysis of the Data Encryption Standard. Springer, Heidelberg (1993). ISBN: 0-387-97930-1, 3-540-97930-1
3. Boyar, J., Find, M., Peralta, R.: Four measures of nonlinearity. In: Algorithms and Complexity, pp. 61–72. Springer, Berlin Heidelberg (2013)
4. Coppersmith, D.: The data encryption standard (des) and its strength against attacks. IBM J. Res. Dev. **38**(3), 243 (1994). doi:10.1147/rd.383.0243
5. Courtois, N.: Algebraic Complexity Reduction and Cryptanalysis of GOST. IACR Cryptology ePrint Archive (2011)
6. Courtois, N.: Security evaluation of GOST 28147-89. In: View Of International Standardisation. IACR Cryptology ePrint Archive (2011)
7. Courtois, N.: An Improved Differential Attack on full GOST. IACR Cryptology ePrint Archive (2012)
8. Courtois, N.: Low complexity key recovery attacks on GOST block cipher. Cryptologia **37**(1), 1–10 (2013)
9. Courtois, N., Misztal, M.: First Differential cryptanalysis of full round 32- round GOST. In: ICICS'11, Beijing. LNCS, vol. 7043, pp. 216–227. Springer, Heidelberg (2011)
10. Courtois, N., Misztal, M.: Aggregated Differentials and Cryptanalysis of PP-1 and GOST. Period. Math. Hung. **65**(2), 177–192 (2012)
11. Mourouzis,T: Optimizations in Algebraic and Differential Cryptanalysis. PhD Thesis, UCL (2015)
12. Courtois, N., Mourouzis, T.: Enhanced truncated differential cryptanalysis of GOST. In: SECRYPT 2013, 10th International Conference on Security and Cryptography, Reykjavik, 29–31 July 2013
13. Courtois, N., Mourouzis, T., Grocholewska-Czurylo, A., Quisquater, J.: On Optimal Size in Truncated Differential Attacks, Budapest, 21–23 May 2014
14. Dolmatov, V.: GOST 28147-89: Encryption, Decryption, and Message Authentication Code (MAC) Algorithms. IETF, Anaheim (2010). ISSN: 2070-1721
15. Dinur, I., Dunkelman, O., Shamir, A.: Improved attacks on full GOST. In: Fast Software Encryption, pp. 9–28. Springer, Berlin Heidelberg (2011)
16. Dolmatov, V.: RFC 5830: GOST 28147-89 Encryption, Decryption and MAC algorithms (2010)
17. Furuya, S.: Slide attacks with a known-plaintext cryptanalysis. In Information Security and Cryptology—ICISC 2001, pp. 214–225. Springer, Berlin Heidelberg (2002)
18. Isobe ,T.: A single-key attack on the full GOST block cipher. In: Fast Software Encryption, pp. 290–305. Springer, Berlin Heidelberg (2011)

19. Khovratovich, D., Ivica Nikolic, I.: Rotational cryptanalysis of ARX. In: Fast Software Encryption, pp. 333–346. Springer, Berlin Heidelberg (2013)
20. Knudsen, L.: Truncated and higher order differentials. In: 2nd International Workshop on Fast Software Encryption, pp. 196–211. Springer, Heidelberg (1994)
21. Knudsen, L.: Truncated and higher order differentials. In: Fast Software Encryption, pp. 196–211. Springer, Berlin Heidelberg (1995)
22. Knudsen, L., Robshaw, M.: The Block Cipher Companion. Springer, Berlin Heidelberg (2011)
23. Lai, X., Massey, J.: Markov ciphers and differential cryptanalysis. In: Davies, D.W. (ed.) Advances in Cryptology. Springer, Heidelberg (1991)
24. Malchik, A.: An English Translation of GOST Standard by Aleksandr Malchik with an English Preface Co-written with Whitfield Diffie (1994)
25. Mantin, I., Shamir, A.: A practical attack on broadcast RC4. In: Fast Software Encryption, pp. 152–164. Springer, Heidelberg (2001)
26. Meier, W., Kunzli, S.: Distinguishing Attack on MAG. ENCRYPT Stream Cipher Project. eSTREAM (2013)
27. Popov, K., Leontiev, S.: Additional Cryptographic Algorithms for Use with GOST 28147-89, GOST R 34.10-94, GOST R 34.10-2001, and GOST R 34.11-94 Algorithms (2006)
28. Poschmann, A., Ling, S., Wang, H.: 256 bit standardized crypto for 650 GE GOST revisited. In: CHES 2010, LNCS, vol. 6225, pp. 219–233. Springer, Heidelberg (2010)
29. Rudskoy, V., Chmora, A.: Working draft for ISO/IEC 1st WD of AMD1/18033-3. In: Russian Block Cipher GOST, ISO/IEC JTC 1/SC 27 N9423, 2011-01-14 (2011)
30. Saarinen, M.: A Chosen Key Attack Against the Secret S-Boxes of GOST (1998)
31. Schneier, B.: Applied Cryptography, 2nd edn. Wiley, New York (1996)
32. Seki, H., Kaneko, T.: Differential cryptanalysis of reduced rounds of GOST. In: Selected Areas in Cryptography, pp. 315–323. Springer, Berlin Heidelberg (2001)
33. Shorin, V., Jelezniakov, V., Gabidulin, E.: Linear and differential cryptanalysis of Russian GOST. Electron. Notes Discret Math. **6**, 538–547 (2001)
34. Shorin, V., Jelezniakov, V., Gabidulin, E.: Security of algorithm GOST 28147-89. In: Abstracts of XLIII MIPT Science Conference (2000)
35. Zabotin. I., Glazkov, G., Isaeva, V.: Cryptographic Protection for Information Processing Systems, Government Standard of the USSR, GOST 28147-89. Government Committee of the USSR for Standards (1989)

A Supply Chain Game Theory Framework for Cybersecurity Investments Under Network Vulnerability

Anna Nagurney, Ladimer S. Nagurney, and Shivani Shukla

Abstract In this paper, we develop a supply chain game theory framework consisting of retailers and consumers who engage in electronic transactions via the Internet and, hence, may be susceptible to cyberattacks. The retailers compete noncooperatively in order to maximize their expected profits by determining their optimal product transactions as well as cybersecurity investments in the presence of network vulnerability. The consumers reveal their preferences via the demand price functions, which depend on the product demands and on the average level of security in the supply chain network. We prove that the governing Nash equilibrium conditions of this model can be formulated as a variational inequality problem, provide qualitative properties of the equilibrium product transaction and security investment pattern, and propose an algorithm with nice features for implementation. The algorithm is then applied to two sets of numerical examples that reveal the impacts on the equilibrium product transactions, the security levels, the product prices, the expected profits, and the retailer vulnerability as well as the supply chain network vulnerability, of such issues as: increased competition, changes in the demand price functions, and changes in the security investment cost functions.

Keywords: Supply chains • Cybersecurity • Investments • Game theory • Nash equilibrium • Variational inequalities • Network vulnerability

A. Nagurney (✉) • S. Shukla
Department of Operations and Information Management, Isenberg School of Management, University of Massachusetts, Amherst, MA 01003, USA
e-mail: nagurney@isenberg.umass.edu; sshukla@som.umass.edu

L.S. Nagurney
Department of Electrical and Computer Engineering, University of Hartford, West Hartford, CT 06117, USA
e-mail: nagurney@hartford.edu

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_16

1 Introduction

As supply chains have become increasingly globalized and complex, there are new risks and vulnerabilities associated with their IT infrastructure due to a spectrum of cyberattacks with greater exposure for both firms and consumers. Coupled with cyberattacks are associated costs, in the form of financial damages incurred by the supply chain firms, the loss of their reputations, as well as opportunity costs, etc. Consumers may also be affected financially by cyberattacks and suffer from the associated disruptions. Cyberattacks can affect numerous different industrial sectors from financial services, energy providers, high tech firms, and retailers to the healthcare sector as well as governments. As noted in [16], the Center for Strategic and International Studies [3] reports that the estimated annual cost to the global economy from cybercrime is more than $400 billion with a conservative estimate being $375 billion in losses, more than the national income of most countries.

For example, the 2013 breach of the major US-based retailer, Target, was accomplished when the cyberattacker entered a vulnerable supply chain link by exploiting the vulnerability in the remote diagnostics of the HVAC system supplier connected to the Target's IT system. In the attack, an estimated 40 million payment cards were stolen between November 27 and December 15, 2013 and upwards of 70 million other personal records compromised (cf. [10]). Target suffered not only financial damages but also reputational costs. Other cyber data breaches have occurred at the luxury retailer Neiman Marcus, the restaurant chain P.F. Chang's, and the media giant Sony (cf. [17]). The Ponemon Institute [22] calculates that the average annualized cost of cybercrime for 60 organizations in their study is $11.6 million per year, with a range of $1.3 million to $58 million. According to The Security Ledger [25], cyber supply chain risk escapes notice at many firms. Mandiant [11] reports that 229 was the median number of days in 2013 that threat groups were present on a victim's network before detection.

Given the impact of cybercrime on the economy and society, there is great interest in evaluating cybersecurity investments. Each year $15 billion is spent by organizations in the United States to provide security for communications and information systems (see [8, 13]). Nevertheless, breaches due to cyberattacks continue to make huge negative economic impacts on businesses and society at-large. There is, hence, growing interest in the development of rigorous scientific tools that can help decision-makers assess the impacts of cybersecurity investments. What is essential to note, however, is that in many industries, including retail, investments by one decision-maker may affect the decisions of others and the overall supply chain network security (or vulnerability). Hence, a holistic approach is needed and some are even calling for a new discipline of cyber supply chain risk management [2].

In this paper, we develop a supply chain game theory model consisting of two tiers: the retailers and the consumers. The retailers select the product transactions and their security levels so as to maximize their expected profits. The probability of a successful attack on a retailer depends not only on that retailer's investment

in security but also on the security investments of the other retailers. Hence, the retailers and consumers are connected. In our previous work (see [17]), we assumed that the probability of a successful attack on a seller depended only on his own security investments. We know that in retail, which we consider in a broad sense here from consumer goods to even financial services, including retail banks, decision-makers interact and may share common suppliers, IT providers, etc. Hence, it is imperative to capture the network effects associated with security investments and the associated impacts.

In our model, retailers seek to maximize their expected profits with the prices that the consumers are willing to pay for the product being a function not only of the demand but also of the average security in the supply chain which we refer to as the cybersecurity or network security. The retailers compete noncooperatively until a Nash equilibrium is achieved, whereby no retailer can improve upon his expected profit by making a unilateral decision in changing his product transactions and security level. Our approach is inspired, in part, by the work of Shetty et al. [24], but it is significantly more general since the retailers, that is, the firms, are not identical and we explicitly also capture the demand side of the supply chain network. Moreover, the retailers may be faced with distinct security investment cost functions, given their existing IT infrastructure and business scope and size, and they can also be spatially separated. Our framework can handle both online retailers and brick and mortar ones. In addition, the retailers are faced with, possibly, different financial damages in the case of a cyberattack. For simplicity of exposition and clarity, we focus on a single type of attack. For a survey of game theory, as applied to network security and privacy, we refer the reader to Manshaei et al. [12]. For highlight of optimization models for cybersecurity investments, see [9].

The supply chain game theory model is developed in Sect. 2. The behavior of the retailers is captured, the Nash equilibrium defined and the variational inequality formulation derived. We also provide some qualitative properties of the equilibrium product transaction and security level pattern. In Sect. 3, we outline the algorithm that we then utilize in Sect. 4 to compute solutions to our numerical examples. In two sets of numerical supply chain network examples, we illustrate the impacts of a variety of changes on the equilibrium solution, and on the retailer and supply chain network vulnerability. In Sect. 5, we summarize our results and present the conclusions along with suggestions for future research.

2 The Supply Chain Game Theory Model of Cybersecurity Investments Under Network Vulnerability

In the model, we consider m retailers that are spatially separated and that sell a product to n consumers. The retailers may be online retailers, engaging with consumers through electronic commerce, and/or brick and mortar retailers. Since our focus is on cybersecurity, that is, network security, we assume that the transactions in terms of payments for the product occur electronically through

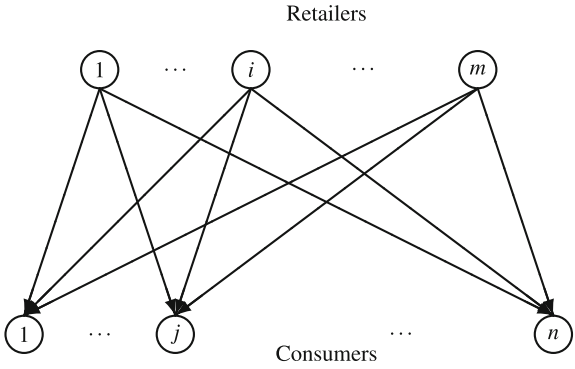


Fig. 1 The network structure of the supply chain game theory model

credit cards and/or debit cards. Consumers may also conduct searches to obtain information through cyberspace. We emphasize that here we consider retailers in a broad sense, and they may include consumer goods retailers, pharmacies, high technology product outlets, and even financial service firms as well as retail banks. The network topology of the supply chain model, which consists of a tier of retailers and a tier of consumers, is depicted in Fig. 1.

Since the Internet is needed for the transactions between retailers and consumers to take place, network security is relevant. Each retailer in our model is susceptible to a cyberattack through the supply chain network since retailers may interact with one another as well as with common suppliers and also share consumers. The retailers may suffer from financial damage as a consequence of a successful cyberattack, losses due to identity theft, opportunity costs, as well as a loss in reputation, etc. Similarly, consumers are sensitive as to how secure their transactions are with the retailers.

We denote a typical retailer by i and a typical consumer by j. Let Q_{ij} denote the nonnegative volume of the product transacted between retailer i and consumer j. Here s_i denotes the network security level, or, simply, the security of retailer i. The strategic variables of retailer i consist of his product transactions $\{Q_{i1}, \ldots, Q_{in}\}$ and his security level s_i. We group the product transactions of all retailers into the vector $Q \in R_+^{mn}$ and the security levels of all retailers into the vector $s \in R_+^m$. All vectors here are assumed to be column vectors, except where noted.

We have $s_i \in [0, 1]$, with a value of 0 meaning no network security and a value of 1 representing perfect security. Therefore,

$$0 \leq s_i \leq 1, \quad i = 1, \ldots, m. \tag{1}$$

The network security level of the retail-consumer supply chain is denoted by $\bar{s}$ and is defined as the average network security where

$$\bar{s} = \frac{1}{m} \sum_{i=1}^{m} s_i. \tag{2}$$

Let p_i denote the probability of a successful cyberattack on retailer i in the supply chain network. Associated with the successful attack is the incurred financial damage D_i. Distinct retailers may suffer different amounts of financial damage as a consequence of a cyberattack due to their size and their existing infrastructure including cyber infrastructure. As discussed in [23] and [24], but for an oligopoly model with identical firms and no demand side represented in the network, p_i depends on the chosen security level s_i and on the network security level $\bar{s}$ as in (2). Using similar arguments as therein, we also define the probability p_i of a successful cyberattack on retailer i as

$$p_i = (1 - s_i)(1 - \bar{s}), \quad i = 1, \ldots, m, \tag{3}$$

where the term $(1-\bar{s})$ represents the probability of a cyberattack in the supply chain network and the term $(1 - s_i)$ represents the probability of success of such an attack on retailer i. The network vulnerability level $\bar{v} = 1-\bar{s}$ with retailer i's vulnerability level v_i being $1 - s_i$; $i = 1, \ldots, m$.

In terms of cybersecurity investment, each retailer i, in order to acquire security s_i, encumbers an investment cost $h_i(s_i)$ with the function assumed to be continuously differentiable and convex. Note that distinct retailers, because of their size and existing cyber infrastructure (both hardware and software), may be faced with different investment cost functions. We assume that, for a given retailer i, $h_i(0) = 0$ denotes an entirely insecure retailer and $h_i(1) = \infty$ is the investment cost associated with complete security for the retailer (see [23, 24]). An example of a suitable $h_i(s_i)$ function is

$$h_i(s_i) = \alpha_i\left(\frac{1}{\sqrt{(1 - s_i)}} - 1\right) \text{ with } \alpha_i > 0. \tag{4}$$

The term α_i allows for different retailers to have distinct investment cost functions based on their size and needs.

The demand for the product by consumer j is denoted by d_j and it must satisfy the following conservation of flow equation:

$$d_j = \sum_{i=1}^{m} Q_{ij}, \quad j = 1, \ldots, n, \tag{5}$$

where

$$Q_{ij} \geq 0, \quad i = 1, \ldots, m; j = 1, \ldots, n, \tag{6}$$

that is, the demand for each consumer is satisfied by the sum of the product transactions between all the retailers with the consumer. We group the demands for the product for all buyers into the vector $d \in R^n_+$.

The consumers reveal their preferences for the product through their demand price functions, with the demand price function for consumer j, ρ_j, being:

$$\rho_j = \rho_j(d, \bar{s}), \quad j = 1, \ldots, n. \tag{7}$$

Observe that the demand price depends, in general, on the quantities transacted between the retailers and the consumers and the network security level. The consumers are only aware of the *average* network security level of the supply chain. This is reasonable since consumers may have information about a retail industry in terms of its cyber investments and security but it is unlikely that individual consumers would have information on individual retailers' security levels. Hence, as in the model of Nagurney and Nagurney [17], there is information asymmetry (cf. [1]).

In view of (2) and (5), we can define $\hat{\rho}_j(Q, s) \equiv \rho_j(d, \bar{s})$, $\forall j$. These demand price functions are assumed to be continuous, continuously differentiable, decreasing with respect to the respective consumer's own demand and increasing with respect to the network security level.

The revenue of retailer i; $i = 1, \ldots, m$, (in the absence of a cyberattack) is:

$$\sum_{j=1}^{n} \hat{\rho}_j(Q, s) Q_{ij}. \tag{8}$$

Each retailer i; $i = 1, \ldots, m$, is faced with a cost c_i associated with the processing and the handling of the product and transaction costs $c_{ij}(Q_{ij})$; $j = 1 \ldots, m$, in dealing with the consumers. His total cost, hence, is given by:

$$c_i \sum_{j=1}^{n} Q_{ij} + \sum_{j=1}^{n} c_{ij}(Q_{ij}). \tag{9}$$

The transaction costs, in the case of electronic commerce, can include the costs of transporting/shipping the product to the consumers. The transaction costs can also include the cost of using the network services, taxes, etc. We assume that the transaction cost functions are convex and continuously differentiable.

The profit f_i of retailer i; $i = 1, \ldots, m$ (in the absence of a cyberattack and security investment) is the difference between the revenue and his costs, that is,

$$f_i(Q, s) = \sum_{j=1}^{n} \hat{\rho}_j(Q, s) Q_{ij} - c_i \sum_{j=1}^{n} Q_{ij} - \sum_{j=1}^{n} c_{ij}(Q_{ij}). \tag{10}$$

If there is a successful cyberattack, a retailer i; $i = 1, \ldots, m$, incurs an expected financial damage given by

$$D_i p_i, \tag{11}$$

where D_i takes on a positive value.

Using expressions (3), (10), and (11), the expected utility, $E(U_i)$, of retailer i; $i = 1, \ldots, m$, which corresponds to his expected profit, is:

$$E(U_i) = (1 - p_i) f_i(Q, s) + p_i(f_i(Q, s) - D_i) - h_i(s_i). \tag{12}$$

We group the expected utilities of all the retailers into the m-dimensional vector $E(U)$ with components: $\{E(U_1), \ldots, E(U_m)\}$.

Let K^i denote the feasible set corresponding to retailer i, where $K^i \equiv \{(Q_i, s_i) | Q_i \geq 0, \text{ and } 0 \leq s_i \leq 1\}$ and define $K \equiv \prod_{i=1}^{m} K^i$.

The m retailers compete noncooperatively in supplying the product and invest in cybersecurity, each one trying to maximize his own expected profit. We seek to determine a nonnegative product transaction and security level pattern (Q^*, s^*) for which the m retailers will be in a state of equilibrium as defined below. Nash [19, 20] generalized Cournot's concept (see [4]) of an equilibrium for a model of several players, that is, decision-makers, each of which acts in his/her own self-interest, in what has been come to be called a noncooperative game.

Definition 1 (A Supply Chain Nash Equilibrium in Product Transactions and Security Levels). A product transaction and security level pattern $(Q^*, s^*) \in K$ is said to constitute a supply chain Nash equilibrium if for each retailer i; $i = 1, \ldots, m$,

$$E(U_i(Q_i^*, s_i^*, \widehat{Q_i^*}, \widehat{s_i^*})) \geq E(U_i(Q_i, s_i, \widehat{Q_i^*}, \widehat{s_i^*})), \quad \forall (Q_i, s_i) \in K^i, \tag{13}$$

where

$$\widehat{Q_i^*} \equiv (Q_1^*, \ldots, Q_{i-1}^*, Q_{i+1}^*, \ldots, Q_m^*); \quad \text{and} \quad \widehat{s_i^*} \equiv (s_1^*, \ldots, s_{i-1}^*, s_{i+1}^*, \ldots, s_m^*). \tag{14}$$

According to (13), an equilibrium is established if no retailer can unilaterally improve upon his expected profits by selecting an alternative vector of product transactions and security levels.

2.1 Variational Inequality Formulations

We now present alternative variational inequality formulations of the above supply chain Nash equilibrium in product transactions and security levels.

Theorem 1. *Assume that, for each retailer i; $i = 1, \ldots, m$, the expected profit function $E(U_i(Q, s))$ is concave with respect to the variables $\{Q_{i1}, \ldots, Q_{in}\}$, and s_i, and is continuous and continuously differentiable. Then $(Q^*, s^*) \in K$ is a supply chain Nash equilibrium according to Definition 1 if and only if it satisfies the variational inequality*

$$-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial E(U_i(Q^*,s^*))}{\partial Q_{ij}}\times(Q_{ij}-Q_{ij}^*)-\sum_{i=1}^{m}\frac{\partial E(U_i(Q^*,s^*))}{\partial s_i}\times(s_i-s_i^*)\geq 0,$$

$$\forall (Q,s)\in K, \tag{15}$$

or, equivalently, $(Q^*, s^*) \in K$ *is a supply chain Nash equilibrium product transaction and security level pattern if and only if it satisfies the variational inequality*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[c_i+\frac{\partial c_{ij}(Q_{ij}^*)}{\partial Q_{ij}}-\hat{\rho}_j(Q^*,s^*)-\sum_{k=1}^{n}\frac{\partial\hat{\rho}_k(Q^*,s^*)}{\partial Q_{ij}}\times Q_{ik}^*\right]\times(Q_{ij}-Q_{ij}^*)$$

$$+\sum_{i=1}^{m}\left[\frac{\partial h_i(s_i^*)}{\partial s_i}-(1-\sum_{j=1}^{m}\frac{s_j^*}{m}+\frac{1-s_i^*}{m})D_i-\sum_{k=1}^{n}\frac{\partial\hat{\rho}_k(Q^*,s^*)}{\partial s_i}\times Q_{ik}^*\right]$$

$$\times(s_i-s_i^*)\geq 0,$$

$$\forall (Q,s)\in K. \tag{16}$$

Proof. (15) follows directly from Gabay and Moulin [7] and Dafermos and Nagurney [5].

In order to obtain variational inequality (16) from variational inequality (15), we note that, at the equilibrium:

$$-\frac{\partial E(U_i)}{\partial Q_{ij}}=c_i+\frac{\partial c_{ij}(Q_{ij}^*)}{\partial Q_{ij}}-\hat{\rho}_j(Q^*,s^*)-\sum_{k=1}^{n}\frac{\partial\hat{\rho}_k(Q^*,s^*)}{\partial Q_{ij}}\times Q_{ik}^*;\quad \forall i,\forall j, \tag{17}$$

and

$$-\frac{\partial E(U_i)}{\partial s_i}=\frac{\partial h_i(s_i^*)}{\partial s_i}-(1-\sum_{j=1}^{m}\frac{s_j^*}{m}+\frac{1-s_i^*}{m})D_i-\sum_{k=1}^{n}\frac{\partial\hat{\rho}_k(Q^*,s^*)}{\partial s_i}\times Q_{ik}^*;\quad \forall i. \tag{18}$$

Making the respective substitutions using (17) and (18) in variational inequality (15) yields variational inequality (16) □

We now put the above Nash equilibrium problem into standard variational inequality form, that is: determine $X^* \in \mathcal{K} \subset R^N$, such that

$$\langle F(X^*), X-X^*\rangle \geq 0,\quad \forall X\in\mathcal{K}, \tag{19}$$

where F is a given continuous function from $\mathcal{K}$ to R^N and $\mathcal{K}$ is a closed and convex set.

We define the $(mn + m)$-dimensional vector $X \equiv (Q, s)$ and the $(mn + m)$-dimensional vector $F(X) = (F^1(X), F^2(X))$ with the (i, j)th component, F^1_{ij}, of $F^1(X)$ given by

$$F^1_{ij}(X) \equiv -\frac{\partial E(U_i(Q, s))}{\partial Q_{ij}}, \tag{20}$$

the ith component, F^2_i, of $F^2(X)$ given by

$$F^2_i(X) \equiv -\frac{\partial E(U_i(Q, s))}{\partial s_i}, \tag{21}$$

and with the feasible set $\mathscr{K} \equiv K$. Then, clearly, variational inequality (15) can be put into standard form (19).

In a similar way, one can prove that variational inequality (16) can also be put into standard variational inequality form (19). □

Additional background on the variational inequality problem can be found in the books by Nagurney [14] and Nagurney et al. [19].

2.2 Qualitative Properties

It is reasonable to expect that the expected utility of any seller i, $E(U_i(Q, s))$, would decrease whenever his product volume has become sufficiently large, that is, when $E(U_i)$ is differentiable, $\frac{\partial E(U_i(Q,s))}{\partial Q_{ij}}$ is negative for sufficiently large Q_{ij}. Hence, the following assumption is not unreasonable:

Assumption 1. *Suppose that in our supply chain game theory model there exists a sufficiently large M, such that for any (i, j),*

$$\frac{\partial E(U_i(Q, s))}{\partial Q_{ij}} < 0, \tag{22}$$

for all product transaction patterns Q with $Q_{ij} \geq M$.

We now give an existence result.

Proposition 1. *Any supply chain Nash equilibrium problem in product transactions and security levels, as modeled above, that satisfies Assumption 1 possesses at least one equilibrium product transaction and security level pattern.*

Proof. The proof follows from Proposition 1 in Zhang and Nagurney [26]. □

We now present the uniqueness result, the proof of which follows from the basic theory of variational inequalities (cf. [14]).

Proposition 2. *Suppose that F is strictly monotone at any equilibrium point of the variational inequality problem defined in (19). Then it has at most one equilibrium point.*

3 The Algorithm

For computational purposes, we will utilize the Euler method, which is induced by the general iterative scheme of Dupuis and Nagurney [6]. Specifically, iteration τ of the Euler method (see also [14]) is given by:

$$X^{\tau+1} = P_{\mathscr{K}}(X^{\tau} - a_{\tau}F(X^{\tau})), \tag{23}$$

where $P_{\mathscr{K}}$ is the projection on the feasible set $\mathscr{K}$ and F is the function that enters the variational inequality problem (19).

As proven in [6], for convergence of the general iterative scheme, which induces the Euler method, the sequence $\{a_{\tau}\}$ must satisfy: $\sum_{\tau=0}^{\infty} a_{\tau} = \infty$, $a_{\tau} > 0$, $a_{\tau} \to 0$, as $\tau \to \infty$. Specific conditions for convergence of this scheme as well as various applications to the solutions of other network-based game theory models can be found in [15, 16], and the references therein.

3.1 Explicit Formulae for the Euler Method Applied to the Supply Chain Game Theory Model

The elegance of this procedure for the computation of solutions to our model is apparent from the following explicit formulae. In particular, we have the following closed form expression for the product transactions $i = 1, \ldots, m; j = 1, \ldots, n$:

$$Q_{ij}^{\tau+1} = \max\{0, Q_{ij}^{\tau} + a_{\tau}(\hat{\rho}_j(Q^{\tau}, s^{\tau}) + \sum_{k=1}^{n} \frac{\partial \hat{\rho}_k(Q^{\tau}, s^{\tau})}{\partial Q_{ij}} Q_{ik}^{\tau} - c_i - \frac{\partial c_{ij}(Q_{ij}^{\tau})}{\partial Q_{ij}})\}, \tag{24}$$

and the following closed form expression for the security levels $i = 1, \ldots, m$:

$$s_i^{\tau+1} =$$

$$\max\{0, \min\{1, s_i^{\tau} + a_{\tau}(\sum_{k=1}^{n} \frac{\partial \hat{\rho}_k(Q^{\tau}, s^{\tau})}{\partial s_i} Q_{ik}^{\tau} - \frac{\partial h_i(s_i^{\tau})}{\partial s_i} + (1 - \sum_{j=1}^{m} \frac{s_j}{m} + \frac{1 - s_i}{m}) D_i)\}\}. \tag{25}$$

We now provide the convergence result. The proof is direct from Theorem 5.8 in [18].

Theorem 2. *In the supply chain game theory model developed above let* $F(X)=-\nabla E(U(Q,s))$ *be strictly monotone at any equilibrium pattern and assume that Assumption 1 is satisfied. Also, assume that F is uniformly Lipschitz continuous. Then there exists a unique equilibrium product transaction and security level pattern* $(Q^*, s^*) \in K$ *and any sequence generated by the Euler method as given by (23), with* $\{a_\tau\}$ *satisfies* $\sum_{\tau=0}^{\infty} a_\tau = \infty$, $a_\tau > 0$, $a_\tau \to 0$, *as* $\tau \to \infty$ *converges to* (Q^*, s^*).

In the next section, we apply the Euler method to compute solutions to numerical game theory problems.

4 Numerical Examples

We implemented the Euler method, as discussed in Sect. 3, using FORTRAN on a Linux system at the University of Massachusetts Amherst. The convergence criterion was $\epsilon = 10^{-4}$. Hence, the Euler method was considered to have converged if, at a given iteration, the absolute value of the difference of each product transaction and each security level differed from its respective value at the preceding iteration by no more than ϵ.

The sequence $\{a_\tau\}$ was: $0.1(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \ldots)$. We initialized the Euler method by setting each product transaction $Q_{ij} = 1.00$, $\forall i, j$, and the security level of each retailer $s_i = 0.00$, $\forall i$.

We present two sets of numerical examples. Each set of examples consists of an example with four variants.

Example Set 1. The first set of examples consists of two retailers and two consumers as depicted in Fig. 2. This set of examples begins with the baseline Example 1, followed by four variants. The equilibrium solutions are reported in Table 1.

The cost function data for Example 1 are:

$$c_1 = 5, \quad c_2 = 10,$$

$$c_{11}(Q_{11}) = .5Q_{11}^2 + Q_{11}, \quad c_{12}(Q_{12}) = .25Q_{12}^2 + Q_{12},$$

$$c_{21}(Q_{21}) = .5Q_{21}^2 + 2, \quad c_{22}(Q_{22}) = .25Q_{22}^2 + Q_{22}.$$

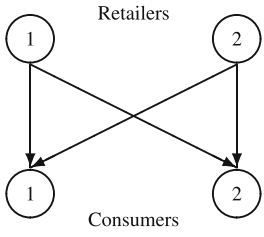


Fig. 2 Network topology for Example Set 1

Table 1 Equilibrium solutions for Examples in Set 1

Solution	Ex. 1	Var. 1.1	Var. 1.2	Var. 1.3	Var. 1.4
Q_{11}^*	24.27	49.27	49.27	24.27	24.26
Q_{12}^*	98.30	98.30	8.30	98.32	98.30
Q_{21}^*	21.27	46.27	46.27	21.27	21.26
Q_{22}^*	93.36	93.36	3.38	93.32	93.30
d_1^*	45.55	95.55	95.55	45.53	45.52
d_2^*	191.66	191.66	11.68	191.64	191.59
s_1^*	0.91	0.91	0.88	0.66	0.73
s_2^*	0.91	0.92	0.89	0.72	0.18
$\bar{s}^*$	0.91	0.915	0.885	0.69	0.46
$\rho_1(d_1^*, \bar{s}^*)$	54.55	104.55	104.54	54.54	54.52
$\rho_2(d_2^*, \bar{s}^*)$	104.35	104.35	14.34	104.32	104.30
$E(U_1)$	8136.45	10894.49	3693.56	8121.93	8103.09
$E(U_2)$	7215.10	9748.17	3219.94	7194.13	6991.11

The demand price functions are:

$$\rho_1(d, \bar{s}) = -d_1 + 0.1(\frac{s_1 + s_2}{2}) + 100, \quad \rho_2(d_2, \bar{s}) = -0.5d_2 + 0.2(\frac{s_1 + s_2}{2}) + 200.$$

The damage parameters are: $D_1 = 50$ and $D_2 = 70$ with the investment functions taking the form:

$$h_1(s_1) = \frac{1}{\sqrt{(1 - s_1)}} - 1, \quad h_2(s_2) = \frac{1}{\sqrt{(1 - s_2)}} - 1.$$

As can be seen from the results in Table 1 for Example 1, the equilibrium demand for Consumer 2 is over four times greater than that for Consumer 1. The price that Consumer 1 pays is about one half of that of Consumer 2. Both retailers invest in security and achieve equilibrium security levels of 0.91. Hence, in Example 1 the vulnerability of Retailer 1 is 0.09 and that of Retailer 2 is also 0.09, with the network vulnerability being 0.09.

In the first variant of Example 1, Variant 1.1, we change the demand price function of Consumer 1 to reflect an enhanced willingness to pay more for the product. The new demand price function for Consumer 1 is:

$$\rho_1(d, \bar{s}) = -d_1 + 0.1(\frac{s_1 + s_2}{2}) + 200.$$

The product transactions to Consumer 1 more than double from their corresponding values in Example 1, whereas those to Consumer 2 remain unchanged. The security level of Retailer 2 increases slightly whereas that of Retailer 1 remains unchanged. Both retailers benefit from increased expected profits. The vulnerability of Retailer 2 is decreased slightly to 0.08.

Variant 1.2 is constructed from Variant 1.1. Consumer 2 no longer values the product much so his demand price function is

$$\rho_2(d_2, \bar{s}) = -0.5d_2 + 0.2(\frac{s_1 + s_2}{2}) + 20,$$

with the remainder of the data as in Variant 1.1. The product transactions decrease by almost an order of magnitude to the second consumer and the retailers experience reduced expected profits by about 2/3 as compared to those in Variant 1.1. The vulnerability of Retailer 1 is now 0.12 and that of Retailer 2: 0.11 with the network vulnerability being: 0.115.

Variant 1.3 is constructed from Example 1 by increasing both security investment cost functions so that:

$$h_1(s_1) = 100(\frac{1}{\sqrt{(1 - s_1)}} - 1), \quad h_2(s_2) = 100(\frac{1}{\sqrt{(1 - s_2)}} - 1)$$

and having new damages: $D_1 = 500$ and $D_2 = 700$. With the increased costs associated with cybersecurity investments both retailers decrease their security levels to the lowest level of all the examples solved, thus far. The vulnerability of Retailer 1 is now 0.34 and that of Retailer 2: 0.28 with the network vulnerability =0.31.

Variant 1.4 has the same data as Variant 1.3, but we now further increase Retailer 2's investment cost function as follows:

$$h_2(s_2) = 1000(\frac{1}{\sqrt{(1 - s_2)}} - 1).$$

Retailer 2 now has an equilibrium security level that is one quarter of that in Variant 1.3. Not only do his expected profits decline but also those of Retailer 1 do.

The vulnerability of Retailer 1 is now: 0.27 and that of Retailer 2: 0.82. The network vulnerability for this example is: 0.54, the highest value in this set of examples. The cybersecurity investment cost associated with Retailer 2 is so high that he greatly reduces his security level. Moreover, the network security is approximately half of that obtained in Example 1.

Example Set 2. The second set of numerical examples consists of three retailers and two consumers as shown in Fig. 3.

In order to enable cross comparisons between the two example sets, we construct Example 2, which is the baseline example in this set, from Example 1 in Set 1. Therefore, the data for Example 2 is identical to that in Example 1 except for the new Retailer 3 data as given below:

$$c_3 = 3, \quad c_{31}(Q_{31}) = Q_{31}^2 + 3Q_{31}, \quad c_{32}(Q_{32}) = Q_{32}^2 + 4Q_{32},$$

$$h_3(s_3) = 3(\frac{1}{\sqrt{(1 - s_3)}} - 1), \quad D_3 = 80.$$

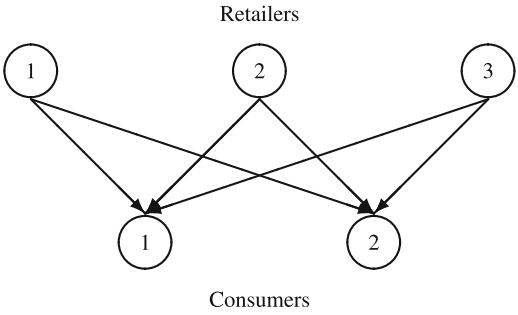


Fig. 3 Network topology for Example Set 2

Table 2 Equilibrium solutions for Examples in Set 2

Solution	Ex. 2	Var. 2.1	Var. 2.2	Var. 2.3	Var. 2.4
Q_{11}^*	20.80	20.98	20.98	11.64	12.67
Q_{12}^*	89.45	89.45	89.82	49.62	51.84
Q_{21}^*	17.81	17.98	17.98	9.64	10.67
Q_{22}^*	84.49	84.49	84.83	46.31	48.51
Q_{31}^*	13.87	13.98	13.98	8.73	9.50
Q_{32}^*	35.41	35.41	35.53	24.50	25.59
d_1^*	52.48	52.94	52.95	30.00	32.85
d_2^*	209.35	209.35	210.18	120.43	125.94
s_1^*	0.90	0.92	0.95	0.93	0.98
s_2^*	0.91	0.92	0.95	0.93	0.98
s_3^*	0.81	0.83	0.86	0.84	0.95
$\bar{s}^*$	0.87	0.89	0.917	0.90	0.97
$\rho_1(d_1^*, \bar{s}^*)$	47.61	47.95	47.96	40.91	44.01
$\rho_2(d_2^*, \bar{s}^*)$	95.50	95.50	95.83	80.47	83.77
$E(U_1)$	6654.73	6665.88	6712.29	3418.66	3761.75
$E(U_2)$	5830.06	5839.65	5882.27	2913.31	3226.90
$E(U_3)$	2264.39	2271.25	2285.93	1428.65	1582.62

The equilibrium solutions for examples in Set 2 are reported in Table 2. With the addition of Retailer 3, there is now increased competition. As a consequence, the demand prices for the product drop for both consumers and there is an increase in demand. Also, with the increased competition, the expected profits drop for the two original retailers. The demand increases for Consumer 1 and also for Consumer 2, both at upwards of 10 %.

The vulnerability of Retailer 1 is 0.10, that of Retailer 2: 0.09, and that of Retailer 3: 0.19 with a network vulnerability of: 0.13. The network vulnerability, with the addition of Retailer 3 is now higher, since Retailer 3 does not invest much in security due to the higher investment cost.

Variant 2.1 is constructed from Example 2 with the data as therein except for the new demand price function for Consumer 1, who now is more sensitive to the network security, where

$$\rho_1(d_1, \bar{s}) = -d_1 + (\frac{s_1 + s_2 + s_3}{3}) + 100.$$

The expected profit increases for all retailers since Consumer 1 is willing to pay a higher price for the product.

The vulnerability of Retailer 1 is now 0.08, that of Retailer 2: 0.08, and that of Retailer 3: 0.17 with a network vulnerability of: 0.11. Hence, all the vulnerabilities have decreased, since the retailers have higher equilibrium security levels.

Variant 2.2 is constructed from Variant 2.1. The only change is that now Consumer 2 is also more sensitive to average security with a new demand price function given by:

$$\rho_2(d_2, \bar{s}) = -0.5d_2 + (\frac{s_1 + s_2 + s_3}{3}) + 200.$$

As shown in Table 2, the expected profits are now even higher than for Variant 2.1. The vulnerability of Retailer 1 is now 0.05, which is the same for Retailer 2, and with Retailer 3 having the highest vulnerability at: 0.14. The network vulnerability is, hence, 0.08. Consumers' willingness to pay for increased network security reduces the retailers' vulnerability and that of the supply chain network.

Variants 2.1 and 2.2 demonstrate that consumers who care about security can also enhance the expected profits of retailers of a product through their willingness to pay for higher network security.

Variant 2.3 has the identical data to that in Variant 2.2 except that the demand price functions are now:

$$\rho_1(d_1, \bar{s}) = -2d_2 + (\frac{s_1 + s_2 + s_3}{3}) + 100, \quad \rho_2(d_2, \bar{s}) = -d_2 + (\frac{s_1 + s_2 + s_3}{3}) + 100.$$

As can be seen from Table 2, the product transactions have all decreased substantially, as compared to the respective values for Variant 2.2. Also, the demand prices associated with the two consumers have decreased substantially as have the expected profits for all the retailers.

The vulnerabilities of the retailers are, respectively: 0.07, 07, and 0.16 with the network vulnerability equal to 0.10.

Variant 2.4 is identical to Variant 2.3 except that now the demand price function sensitivity for the consumers has increased even more so that:

$$\rho_1(d_1, \bar{s}) = -2d_2 + 10(\frac{s_1 + s_2 + s_3}{3}) + 100, \quad \rho_2(d_2, \bar{s}) = -d_2 + 10(\frac{s_1 + s_2 + s_3}{3}) + 100.$$

All the equilibrium product transactions now increase. The demand prices have both increased as have the expected profits of all the retailers.

In this example, the vulnerabilities of the retailers are, respectively: 0.02, 0.02, and 0.05, yielding a network vulnerability of 0.03. This is the least vulnerable supply chain network in our numerical study.

5 Summary and Conclusions

Cybercrime is affecting companies as well as other organizations and establishments, including governments, and consumers. Recent notable data breaches have included major retailers in the United States, resulting in both financial damage and a loss in reputation. With companies, many of which are increasingly global and dependent on their supply chains, seeking to determine how much they should invest in cybersecurity, a general framework that can quantify the investments in cybersecurity in supply chain networks is needed. The framework should also be able to illuminate the impacts on profits as well as a firm's vulnerability and that of the supply chain network.

In this paper, we develop a supply chain network game theory model consisting of a tier of retailers and a tier of consumers. The retailers may be subject to a cyberattack and seek to maximize their expected profits by selecting their optimal product transactions and cybersecurity levels. The firms compete noncooperatively until a Nash equilibrium is achieved, whereby no retailer can improve upon his expected profits. The probability of a successful attack on a retailer, in our framework, depends not only on his security level, but also on that of the other retailers. Consumers reveal their preferences for the product through the demand price functions, which depend on the demand and on the network security level, which is the average security of the supply chain network.

We derive the variational inequality formulation of the governing equilibrium conditions, discuss qualitative properties, and demonstrate that the algorithm that we propose has nice features for computations. Specifically, it yields, at each iteration, closed form expressions for the product transactions between retailers and consumers and closed form expressions for the retailer security levels. The algorithm is then applied to compute solutions to two sets of numerical examples, with a total of ten examples. The examples illustrate the impacts of an increase in competition, changes in the demand price functions, changes in the damages incurred, and changes in the cybersecurity investment cost functions on the equilibrium solutions and on the incurred prices and the expected profits of the retailers. We also provide the vulnerability of each retailer in each example and the network vulnerability.

The approach of applying game theory and variational inequality theory with expected utilities of decision-makers to network security/cybersecurity that this paper adopts is original in itself. The results in this paper pave the way for a range of investigative questions and research avenues in this area. For instance, at present, the model considers retailers and consumers in the supply chain network. However, it can be extended to include additional tiers, namely, suppliers, as well as transport service providers, and so on. The complexity of the supply chain network would

then make it even more susceptible to cyberattacks, wherein a security lapse in one node can affect many others in succession. Moreover, to account for the fact that the exchange of data takes place through multiple forms, the model could be extended to include multiple modes of transactions.

While the solution equilibrium in the context of competition does moderate investments, the model can also be extended to explicitly include constraints on cybersecurity investments subject to expenditure budgets allocated to cybersecurity. The numerical examples section dealt with multiple retailer and consumer scenarios and their variants to validate the ease of adoption and practicality of the model. A case study and empirical analysis can further corroborate the cogency of the model and assist in the process of arriving at investment decisions related to cybersecurity. This could also provide insights as to how to strike a balance between effectiveness of service and security. We leave the above research directions for future work.

Acknowledgements This research of the first author was supported by the National Science Foundation (NSF) grant CISE #1111276, for the NeTS: Large: Collaborative Research: Network Innovation Through Choice project awarded to the University of Massachusetts Amherst as well as by the Advanced Cyber Security Center through the grant: Cybersecurity Risk Analysis for Enterprise Security. This support is gratefully acknowledged.

References

1. Akerlof, G.A.: The market for 'lemons': quality uncertainty and the market mechanism. Q. J. Econ. **84**(3), 488–500 (1970)
2. Boyson, S.: Cyber supply chain risk management: revolutionizing the strategic control of critical IT systems. Technovation **34**(7), 342–353 (2014)
3. Center for Strategic and International Studies: Net Losses: Estimating the Global Cost of Cybercrime, Santa Clara (2014)
4. Cournot, A.A.: Researches into the Mathematical Principles of the Theory of Wealth, English translation. MacMillan, London (1838)
5. Dafermos, S., Nagurney, A.: Oligopolistic and competitive behavior of spatially separated markets. Reg. Sci. Urban Econ. **17**, 245–254 (1987)
6. Dupuis, P., Nagurney, A.: Dynamical systems and variational inequalities. Ann. Oper. Res. **44**, 9–42 (1993)
7. Gabay, D., Moulin, H.: On the uniqueness and stability of Nash equilibria in noncooperative games. In: Bensoussan, A., Kleindorfer, P., Tapiero, C.S. (eds.) Applied Stochastic Control of Econometrics and Management Science. North-Holland, Amsterdam (1980)
8. Gartner: "Gartner reveals Top 10 Security Myths", by Ellen Messmer. NetworkWorld (11 June 2013)
9. Gordon, L.A., Loeb1, M.P., Lucyshyn, W., Zhou, L.: Externalities and the magnitude of cyber security underinvestment by private sector firms: a modification of the Gordon-Loeb model. J. Inf. Secur. **6**, 24–30 (2015)
10. Kirk, J.: Target Contractor Says It Was Victim of Cyberattack. PC World (6 February 2014)
11. Mandiant: M-trends: Beyond the Breach. 2014 Threat report. Alexandria, Virginia (2014)
12. Manshei, M.H., Alpcan, T., Basar, T., Hubaux, J.-P.: Game theory meets networks security and privacy. ACM Comput. Surv. **45**(3), Article No. 25 (2013)
13. Market Research: United States Information Technology Report Q2 2012 (24 April 2013)

14. Nagurney, A.. Network Economics: A Variational Inequality Approach, 2nd and revised edn. Kluwer Academic, Boston (1993)
15. Nagurney, A.: Supply Chain Network Economics: Dynamics of Prices, Flows, and Profits. Edward Elgar, Cheltenham (2006)
16. Nagurney, A.: A multiproduct network economic model of cybercrime in financial services. Service Science **7**(1), 70–81 (2015)
17. Nagurney, A., Nagurney, L.S.: A Game Theory Model of Cybersecurity Investments with Information Asymmetry. Netnomics, (2015). in press
18. Nagurney, A., Zhang, D.: Projected Dynamical Systems and Variational Inequalities with Applications. Kluwer Academic, Boston (1996)
19. Nagurney, A., Yu, M., Masoumi, A.H., Nagurney, L.S.: Networks Against Time: Supply Chain Analytics for Perishable Products. Springer, New York (2013)
20. Nash, J.F.: Equilibrium points in n-person games. Proc. Natl. Acad. Sci. USA **36**, 48–49 (1950)
21. Nash, J.F.: Noncooperative games. Ann. Math. **54**, 286–298 (1951)
22. Ponemon Institute: Second Annual Cost of Cyber Crime Study: Benchmark Study of U.S. Companies (2013)
23. Shetty, N.G.: Design of Network Architectures: Role of Game Theory and Economics. PhD dissertation, Technical Report No. UCB/EECS-2010-91, Electrical Engineering and Computer Sciences, University of California at Berkeley (4 June 2010)
24. Shetty, N., Schwartz, G., Felegehazy, M., Walrand, J.: Competitive cyber-insurance and Internet security. In: Proceedings of the Eighth Workshop on the Economics of Information Security (WEIS 2009). University College London, 24–25 June 2009
25. The Security Ledger: Supply Chain Risk Escapes Notice at Many Firms (6 November 2014)
26. Zhang, D., Nagurney, A.: On the stability of projected dynamical systems. J. Optim. Theory Appl. **85**, 97–124 (1995)

A Method for Creating Private and Anonymous Digital Territories Using Attribute-Based Credential Technologies

Panayotis E. Nastou, Dimitra Nastouli, Panos M. Pardalos, and Yannis C. Stamatiou

Abstract In this paper, the privacy aspect of the Digital Territory concept is considered within the general domain of Ambience Intelligence. Digital Territories (or DTs for short) are digital, artificial entities that are dynamically created by their owners as they move about in a physical space. In brief, a Digital Territory is defined as a subset of physical space which is created by some technological means. It has semipermeable boundaries and properties defined by its owners. An example of a Digital Territory is the range defined by a WiFi access point or the access range of a bluetooth device. Since Digital Territories are created in the open space, a major issue that arises during their creation and lifetime is their security and privacy, in terms of what entities can have access to them and with which access rights. In this work a generic privacy preserving architecture is proposed for DTs of any kind based on a new Privacy Enhancing Technology, the Privacy-ABCs.

Keywords: Digital Territories • Privacy Enhancing Technology • Ambience Intelligence • Attribute Based Credentials

P.E. Nastou (✉)
Center for Applied Optimization, University of Florida, Gainesville, FL, USA

Department of Mathematics, University of Aegean, Samos, Greece
e-mail: pnastou@aegean.gr

D. Nastouli
Department of Business Administration, University of Patras, Patra, Greece
e-mail: nastouli@upatras.gr

P.M. Pardalos
Center for Applied Optimization, University of Florida, Gainesville, FL, USA

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, USA
e-mail: pardalos@ise.ufl.edu

Y.C. Stamatiou
Department of Business Administration, University of Patras, Patra, Greece

Computer Technology Institute and Press ("Diophantus"), Patras 26504, Greece
e-mail: stamatiu@ceid.upatras.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_17

1 Introduction

A *Digital Territory* is an artificial entity that is dynamically created and destroyed, upon its creator's will, in order to fulfill certain goals. It has certain access properties and interacts with its environment through a well-defined interface. In this sense, it can be likened with a *magnet* that creates a territory/field around it which can interact with metal objects within a certain range depending on the strength of the magnet. Since its inception, the Digital Territory concept has attracted attention both in terms of defining its conceptual framework and its real life implementation. The concept of a *Digital Territory*, DT for short, seems to combine two other well-known concepts: *Artificial Life* and *Artificial Intelligence*. However, Digital Territories are of a different nature from these popular "artificial" concepts. Their study touches on mathematical techniques ranging from formal logic (when one needs to formally describe their properties and interrelationships) to random graphs (when large, interacting communities of Digital Territories are investigated) as well as on technological advances in the Information and Communication Technologies (ICTs) with respect to their realization.

The DT discipline involves mobile, interacting agents which co-exist in complex domains (e.g., physical space or in the Internet) which, also, exhibit intelligence while interacting. A very informative account Digital Territories can be found in [13].

As a DT is specifically created to exist in the open space and it can be accessible from its environment, it is natural that its existence and operation are beset with privacy and security threats, both for its creator and those who access it, as well as the DT itself. Moreover, recent advances in the Internet as well as the capabilities of portable devices have opened up DT creation and maintenance possibilities unforeseen a decade ago, when DT principles were being developed. We are on the verge of being surrounded by DTs wherever we happen to be, created from devices ranging from smart phones and environment sensors to smart houses and autonomous vehicles. All these ubiquitous networked devices with smart capabilities can give rise to DTs of widely varying properties. However, all privacy and security issues that beset these devices and the networking environment, most often a local network or the Internet, are carried over to the created DTs, within the communication range of the devices, themselves. In this paper we focus on the privacy aspect of DTs and attempt to delineate a general framework within which it can be properly handled. As discussed in [12], privacy is one of the main issues of a Digital Territory and its creator.

Almost all applications and services based on computer systems, and DTs also fall in this category, require some form of user authentication to establish trust relations or service access rights, either for only one endpoint of communication or for both. One widely used mechanism for this is password-based authentication. Given the weaknesses of such a simple authentication method, multiple alternate techniques have been developed to provide a higher degree of access control. Cryptographic certificates are one known example of this. Although such certificates

can offer sufficient security for many purposes, they do not typically handle privacy adequately because they reveal completely the identity of a person. Any usage of such a certificate exposes the identity of the certificate holder to the party (usually a service) requesting authentication. There are many scenarios where the use of such certificates reveals, unnecessarily, the identity of the holder. For example, this is the case for scenarios where a service platform only needs to verify the age of a user but not his/her actual identity. Revealing more information than necessary not only harms the privacy of the users but also increases the risk of abuse of information such as identify theft when information revealed falls in the wrong hands.

Over the past 10–15 years, a number of technologies have been developed to build Attribute Based Credential (ABC) systems in a way that they can be trusted, much like normal cryptographic certificates, while at the same time protecting the privacy of their holder (e.g., hiding the real holder's identity). Such certificates, called Attribute Based Credentials are issued just like ordinary credentials (e.g., the X.509 credentials commonly employed in Public Key Infrastructures) using a digital (secret) signature key. However, ABCs allow their holder to transform them into a new credential that contains only a subset of the attributes contained in the original credential. Still, these transformed credentials can be verified like ordinary cryptographic credentials (using the public verification key of the issuer) and offer the same strong security.

The rest of the paper is organized as follows. In Sect. 2 we briefly discuss the Digital Territory concept as well as its related concepts in order to define our target privacy domain. In Sect. 3, we discuss the main privacy threats for DTs and the risks they pose to individuals accessing DTs while in Sect. 4, we further discuss the threats in DTs but this time from the point of view of the emerging *Semantic Web* (or Web 3.0). We show that the Semantic Web, as useful as it will be in locating and semantically processing information and knowledge on the Web, it can nevertheless threaten individuals' privacy as their data and personal information will be more easily amenable to automatic processing and inferencing. In Sect. 5 we present our main privacy preserving tool, the *Privacy-ABC* technology, which we will employ in order to protect individuals and DT owners' privacy when accessing and creating DTs, respectively. In this section we, also, present our approach towards the deployment of the Privacy-ABC technology in the DT domain while in Sect. 6 we discuss our approach and provide thoughts for further investigation.

2 Digital Territories

For our purposes, a *Digital Territory* is a *transient*, in general, *Ambient Intelligent* space: it is created in space (ambience) for a specific purpose and integrates the intention of its creator (either a human being, most often, or a machine). *Ambient Intelligence*, or AmI for short, is also named pervasive computing, ubiquitous computing and embedded intelligence among other well-known synonymous terms and concepts. An AmI space is composed of available ICTs, network infrastructures,

services that cover any conceivable human activity domain ranging from a smart home to a car. Existing in an AmI environment entails a number of prerequisites that include freedom of action, access to data and information, protection of privacy, security of personal information, and trust towards the AmI environment. In other words, such an environment should be highly personalized and privacy respecting. One approach, which we advocate in this paper, towards achieving these goals is to employ special technologies, termed *Privacy Enhancing Technologies*, or *PETs* for short. Using these technologies one may build such environments establishing boundaries to what can be known about individuals or AmI spaces with an eye towards preserving privacy and establishing trust.

A Digital Territory, which is the realization of an AmI space, is a virtual entity created by a group of entities or an individual entity towards the realization of specific goals (e.g. a public service or data gathering application). A DT has a number of salient characteristics such as its infrastructure, its access properties, the offered services as well as auxiliary entities or objects. The Digital Territory, being a form of territory, is defined and enclosed by *borders*.

Boundaries are points where interactions occur between the interior and exterior of the DT. Boundaries are defined by negotiations for interaction between involved parties. Some examples of fundamental boundaries are the following (see the papers in [11]):

- disclosure boundary (between private/public)
- identity boundary (between self/others)
- time boundary (between past/future)

Borders, in turn, are the realization of the boundaries. The goal is to be visualized, externally, in a clear and well-understood way. Border access is controlled by the DT creator, who can impose access restrictions and control mechanisms of varying levels.

In order negotiations to take place, interaction must be possible. Multi-lateral interaction requires individual internal interaction and interpretation mechanisms, individual goals, a commonly understood protocol (concepts, interface and language) and a negotiable boundary.

A Digital Territory (DT) exists in both physical (e.g., a public WiFi access point) and digital spaces (e.g., the information concerning an individual or a service provider). It is a place wherein information processing and storage happens; it causes information communication across its borders; it perceives and affects its environment through the management of its boundaries. In essence, it is an information processing entity.

Markers are the means defining the borders and the points of negotiation and crossing between DTs and individuals. Thus, a marker can be defined as a set of landmarks with associated constraints, both of which denoted by symbols. *Markers* are the technical means of realizing the borders as intuitive interfaces. They can be expanded to include interfaces, authorization, access control, information visualization, affordances, semantics, functionality.

A *bubble* is a frequently used metaphor for the visualization of a DT. It has an owner (visualized at the DT center), a radius (that defines the DT's range), and duration (it is ephemeral). Its enclosing membrane can be set to different degrees of opacity.

3 Privacy Threats and Protection Strategies

As it happens with any visit of an entity to a publicly available digital space (a DT is such a space), digital traces are left during the interaction of the entity with the DT. These traces are, usually, not under the control or, even, knowledge of the visiting entity. Thus, most individuals visiting a DT are concerned about possible privacy violations, such as the preservation and sharing of personal information such as their preferences or beliefs as well as other personal or identifying information.

In general, the most important privacy related threats in DTs are the following:

- Traffic data and exchanged information may be disclosed to third parties during a transaction or even stored by the DT owner for further processing or distribution. The distribution of such data may result to spam communications as well as their exploitation for illegal actions by third parties.
- Personal life violation may occur.
- Location data of the DT visitors may be inferred by monitoring their transactions with a DT.
- Identity appropriation may result from disclosure of identity information and authentication credentials.
- Visitor profiling is possible through recording and analyzing transaction data. Such profiling may include choices, product preferences, reading habits, beliefs, etc.
- In general, these transaction traces may remain intact for an indefinite period of time and their association with an individual may result to never ending privacy violations whose source may be difficult to locate.

With respect to privacy protection, there are some generally agreed upon strategies that one can adopt to impose privacy protection during DT-visitor interactions. These strategies include the following:

- A privacy policy should be designed and enforced, addressing appropriately the handling of personal data. This policy should be compliant with local and international legislations so as to protect the rights of the individuals accessing DTs.
- Security measures should be adopted by DT owners for protecting personal data residing within the DT. This protection targets loss of data or unauthorized data access, malicious or accidental data modification, data deletion or disclosure and identity theft among others. Technical protection measures include authentication, role-based access and access control, accountability, cryptography,

anonymity, pseudonymity, and action unlinkability. Administrative protection measures include privacy strategic planning and development of suitable privacy policies.
- Personal communication data and pseudonymity are protected. Only under exceptional and clearly defined cases, data and identity can be uncovered.
- Personal data collection should be limited only to the absolutely necessary data for accessing a DT and accomplishing a transaction (minimal disclosure principle).
- The uses of collected data should be clearly defined in a policy accessible by DT visitors. Later use of data should strictly adhere to the principles stated in the policy.
- Individuals must be informed about the collection of their personal data when data is collected. Use of data should be performed only upon the data owner's consent.
- Individuals should be able to know what a DT knows about them as well as access this data upon suitable authentication.
- Processors of personal information should comply with legislation and best practices. Auditing should be in effect in order to ascertain that this is respected.
- The data controller should give individuals the possibility to report complaints and demand remedial actions, if data misuse is suspected.

Identity theft is a privacy related threat which may result in fraudulent actions against individuals. We may differentiate between real identities, i.e. information used for the real identification of an individual, and on-line (digital) identities, i.e. real identities or partial information or pseudonyms used by individual entities or their proxies in their interactions in different digital territories. Examples of partial identities are driver's license number, frequent flier number, home phone number, credit card number, health registration number, e-mail addresses, cookies etc. On the other hand, we may differentiate between on-line and off-line identities. Then, examples of on-line identities are usernames, pseudonyms, e-mail addresses, cookies, etc., and of off-line identities, a driver's license number, frequent flier number, home phone number, credit card number and health registration number. However, off-line identities may also be used as on-line identities.

Identity management is of crucial importance in deploying and operating DTs since their acceptance and usefulness will depend on building and maintaining trust relationships between all involved entities. There are at least the following three requirements stated in the literature, regarding identity management [1]:

- Reliability and dependability. A digital entity must protect users against forgery and related attacks while guaranteeing to other entities that users can meet transaction related obligations.
- Controlled information disclosure. Users must have control over which identity to use in specific circumstances, as well as over its secondary use and the possible replication of any identity information revealed in a transaction.

– Mobility support. Mobile computing infrastructure and components of DTs must be able to apply multiple and dependable digital identities, i.e. to remove technical limits that do not allow applying such identity management solutions.

Multiple and dependable digital identities could be based on a public key infrastructure and trusted third parties, much like Privacy ABCs, which we are going to describe in Sect. 5.

4 The Semantic Web and Its Implications on Individuals' Privacy

It is not hard to see that most of DTs are bound to rely on the Web for their creation and service delivery, much like other Web services. If not Web itself, at least on some local network or wireless connection protocol which, indirectly, may connect to the Web in order to enhance outreach and visibility. In this section, we will outline an important concern one should have in mind when creating a DT within the Web infrastructure with respect to privacy and personal information protection.

Since its inception, the Web was, initially, destined for use by human beings by storing information directly in a text format, searchable (indexed) by suitable algorithms. As it is organized today, the Web does not provide appropriate processing facilities for machines, that is facilities that allow computer algorithms to quickly combine existing information, derive new information thereof, reach conclusions and, in general, produce new knowledge. Since the vastness of information in the Web should be amenable to fast machine based processing in order to be possible to extract information. The goal, thus, should be to make a Web friendlier to machines and algorithms than it currently is.

One approach is to have a Web that stores information in a way that a machine can easily "understand" and process. This gave rise to the *Semantics Web* initiative that was, first, proposed in [1], a work coauthored by Tim Berners-Lee himself, the invertor of the Web. The Semantic Web initiative aims at extending the existing capabilities of the Web for human–machine interaction so as to include, also, machines, which will be able to "understand" and process Web information more efficiently. Thus, they will be able to derive more information than the information existing in the web in explicit (e.g., text or XML) form.

According to this initiative, all information residing in the Web (including, also, all types of personal information of individuals that is exposed either inadvertently or on purpose) is indexed and described based on the architecture which appears in Fig. 1. This architecture, called the *Semantic Web Stack*, was presented by Tim Berners-Lee's talk at the conference XML2000. It forms the basis of all enhancements that are currently being studied and evaluated by involved researchers, mainly the *World Wide Web Consortium* (W3C).

When this architecture is fully implemented (Web 3.0 to Web 4.0) searching machines will be able to index, locate, and "understand" this information, deriving

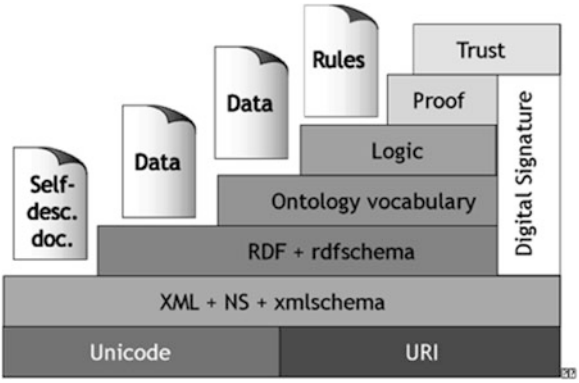


Fig. 1 The semantic web stack

more facts from the existing ones, not explicitly appearing before. These algorithms are foreseen to be offered to their users by all devices connected on the Web (in the unified"Internet of Things" or IoT) and will offer all Web entities to manipulate the information contained in the Semantic Web (including, also, stored personal information). Consequently, all entities will be able to describe and use information based on available semantic descriptions (languages) so as drawing conclusions or deriving new information will be a commonplace Web functionality. In conclusion, the Semantic Web initiative aspires to interconnect all information existing in the Web into a unified semantics-based description that can be used by machines in order to search, locate, and process information in such a way to enable the production of new knowledge, information, and facts which may, possibly, not be stored in explicit form.

A positive aspect of the Semantic Web initiative is that entities will be able to locate, accurately and fast, *what is known* about them, by the Web. This is strongly related to privacy since individuals will have a way to see if personal information is directly or indirectly stored in the Web or the visited services. Therefore, at least theoretically, one may employ (using suitable applications) the Semantic Web to, periodically or even on a daily basis, search for existing or derivable personal information on the Web.

There is a negative aspect of this, however. With the new Web, profiling and personal information processing procedures will be easier and more informative. More individuals will be able to build other people's profiles and link their actions, threatening their privacy. As the Web never forgets, as it is now, maybe it will be even harder to forget when it reaches the Semantic Web state.

Our proposal is to handle the privacy issues that will beset the Semantic Web, as well as any service (DTs included), using PETs, such as the one discussed in Sect. 5.

5 A New Privacy Preserving Authentication Technology

In this section, we will describe the employment of a new privacy preserving authentication technology, called *Privacy ABC*, involving a special kind of credentials. In what follows, we explain these credentials' functionality as well as applicability in the development of a privacy preserving DT environment. Initially, we will describe a generic ABC system.

5.1 A Generic ABC Architecture

Just like any other identity management and authentication service, ABC systems typically involve a number of mandatory actors as well as some additional optional actors, depending on the specific features an application requires. The *User* is one actor who can be any holder, customer, citizen, or participant of an ABC system. The *Verifier* (also Relying Party, Service Provider) and an *Issuer* (also: Identity Service Provider) are necessary. However, some additional actor roles are foreseen for ABC systems, which are optional. In particular, these are the *Revocation Service* and the *Inspector*.

The Issuer generates and provides credentials containing Attributes to the User. On request, the Issuer generates the credential during the issuance protocol and provides it to the User. Depending on the use case, the credential information may be provided either by the User herself or the Issuer, if he already holds the respective information in the attribute database. Ideally, the Issuer can provide the information he attests directly, being an authoritative source. In doing so, he should have the right to assign the relevant attribute to various entities. Examples would be assigning attributes of the User to the university for the student status, to the bar association for the attribute of being an advocate, or to the trade register for the company status. Finally, attributes may also be generated jointly at random, e.g. where this may be useful for specific uses or cryptographic processes.

The User issues the credentials while interacting with the Issuer enabling her to provide proof of certain attributes towards the Verifier. The User acts in different roles. She receives credentials from the Issuer and provides a proof for certain requested attributes towards the Verifier. In some cases, additional information needed for inspection are provided as well.

The Verifier receives a presentation token from the User allowing him to check that the User has certain attributes. The Verifier usually provides some kind of access restricted service to the User to which the User needs to authenticate and stipulates a policy for access. This will require the User to either reveal or to proof possession of certain attributes values.

The Inspector reveals the identity or other encrypted attribute values of a User (e.g., lifting anonymity) upon legitimate request. For this, the Inspector has to examine the legitimacy according to the previously declared inspection grounds. The Inspector is an optional entity in an ABC system.

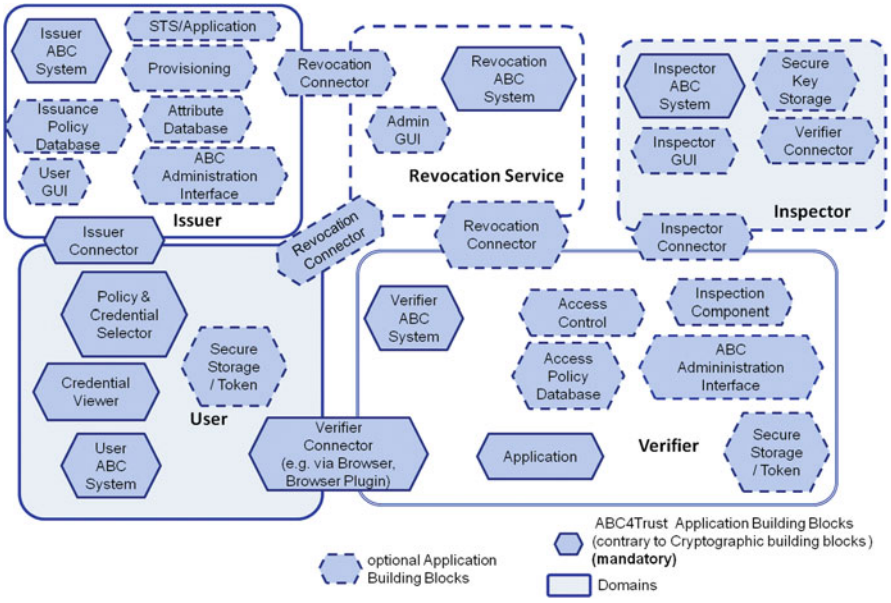


Fig. 2 Building blocks and domains

A Revocation Service is responsible for revoking issued credentials. After revocation, the credentials cannot produce valid presentation tokens (i.e., proofs about credentials). The Revocation Service is an optional component of an ABC system. Often, the entity offering the Revocation Service is the same as that offering the Issuer service, which can be assumed to have the most accurate information about users' attributes and credentials. In Fig. 2, we can see the architecture of a generic ABC system based on the services described above.

5.2 A Privacy Preserving ABC Architecture for DTs

In order to provide privacy to the entities involved in the creation and use of an ABC based DT, the main goal is to build, for all involved parties, electronic identities based on some type of *Privacy Preserving Technology* or, briefly, *PET*.

In general, the commonly used entity authentication and service registration methods (e.g., PKI-based) that are employed today for controlling access to Internet services most often fall short with regard to respecting users' privacy. This situation arises in services in which only a subset of a user's full identity profile is necessary to allow access to a service or the users can, simply, use the service using a pseudonym (i.e., full anonymity). Such services range from accessing online libraries, where there is no need to give full identity profile to access books but only a proof that

you are subscribed to the library, to online borrowing of movies, where you may have to prove that you are of appropriate age (e.g., older than 18) in order to watch particular films. In such types of applications there is, clearly, a need for a partial, and not complete, revelation of the user's identity.

Privacy Attribute Based Credentials or Privacy-ABCs, for short, is a technology that enables privacy preserving and partial authentication of users. Privacy-ABCs are issued just like normal electronic credentials (e.g., those based on currently employed PKIs) using a secret signature key owned by the credential issuer.

However, and this is a key feature of this technology, the user is in position to transform the credentials into a new form, called *presentation token*, that reveals only the information about him which is really necessary in order to access a service. This new token can be verified with the issuer's public key.

As we have already noted, a credential is a set of attributes issued, and certified, by the credential Issuer to a User. By issuing a credential, the Issuer certifies for the correctness of the attributes it contains. These attributes are identity elements and information about the User and a Service. After issuance, the User can use the credential in order to produce presentation tokens (i.e., proofs) that uncover to other entities, the Verifiers, partial information about the attributes contained in the credential. Although attributes can be of any type (e.g., integers, strings, etc.) they must eventually be mapped onto integers in order to be suitably encoded into a credential. This mapping, along with the list and type of encoded credentials, is defined in the credentials specification of the Issuer.

A User can provide certified information to Verifiers in order, for instance, to authenticate herself towards a service, using one or more of her credentials to produce a presentation token which is, then, sent to the Verifier. A presentation token can combine information from any subset of the credentials possessed by the User. Thus, the presentation token can: (i) reveal the values of a subset of the attributes contained in the credentials (e.g., IDcard.firstname = "John"), (ii) show that a credential value satisfies a predicate, such as an inequality (e.g., IDcard.birthdate $<$ 1993/01/01), and (iii) the values of two different credentials satisfy a predicate (e.g., IDcard.lastname = creditcard.lastname).

In addition to revealing information about attributes, a presentation token can, also, sign an application-specific message as well as a random nonce, if necessary, to guarantee freshness. Moreover, presentation tokens support a number of advanced features such as pseudonyms, device binding, inspection, and revocation that were described earlier.

A Verifier announces in its presentation policy which credentials from which Issuers it accepts and which information the presentation token must reveal from these credentials. The Verifier can cryptographically verify the authenticity of a received presentation token using the credential specifications and issuer parameters of all credentials involved in the token. The Verifier must obtain the credential specifications and issuer parameters in a trusted manner, e.g., by using a traditional PKI to authenticate them or retrieving them from a trusted location.

Presentation tokens based on privacy-ABCs are in principle cryptographically unlinkable and untraceable, meaning that Verifiers cannot tell whether two presentation tokens were derived from the same or from different credentials, and that Issuers cannot trace a presentation token back to the issuance of the underlying credentials.

Actually, the basic idea in a Privacy-ABC based DT is the Issuer to create credentials to potential DT users so that they can authenticate themselves towards the DT, revealing only the information they want or need to reveal that is prove only their eligibility to enter a DT and nothing else. Now, the Issuer will be a simple user registration service requiring, from participants, the relative information through a Web form channeled over secure SSL connection to the user device.

After registration, the issuer (which may be even be a DT owner) sends to the registered mobile device a registration token signed by the issuer, that the device will use later in order to communicate with the platform and send data. The Verifier, which can be any DT, receives data from the user along with the registration token, which proves that the device sending the data is a registered device. The verifier will check the registration token for validity and then store the corresponding data along with the demographic information associated with the user of the device (without any identifying information, however).

In general, the ultimate goal of using Privacy-ABC systems is to provide Users with the ability of acting fully anonymously while using services of different kinds. By deploying an Inspector entity, the purviews he is assigned to cannot be fully anonymous anymore. Therefore, a Privacy-ABC system involving an Inspector entity are to be considered as pseudonymous only, with the correlating consequences for legal data protection requirements. Therefore, the use of an Inspector building block should not be the default setting but based on a well-considered decision. Where the alternative is that data controllers collect the identifying information of all Users not as these are necessary for the normal service provision but just for the case that something goes wrong (unpaid bill, upload of illegal content, etc.) inspection may offer a more privacy-preserving solution. With this data controller can effectively hide the personal data from themselves. But as identification is possible this fact as well as the reasons that would allow user identification must be made clear to the users. Presently, we do not see any role for the Inspector and Revocation entities in the context of Digital Territories.

Research has led to different techniques of how to realize anonymous credentials [3, 8, 9] which are based on different number-theoretic problems and also differ somewhat in the functionality that they offer. There are two leading anonymous credentials systems: Idemix of IBM and U-prove of Microsoft. These two systems provide nearly the same functionality, using different cryptographic primitives. Idemix relies on the hardness of the strong RSA problem while U-prove relies on the difficulty of discrete logarithms. Also, credentials are represented in different formats within these two systems. The ABC4Trust project unified these two credential formats into one, with an eye towards interoperability and efficiency. Some of the outcomes of this project may be found in [2] (reference architecture and implementation) and [14] (pilot application).

These two credential types provide the necessary functionality for supporting user credentials with the following properties:

- Unforgeability (issuing).
- Selective disclosure with the user controlling the disclosed information set.
- Soundness (no false claims about the validity of a credential).
- Untraceability (showings, with respect to issuing).
- Unlinkability (between different showings).

Technical descriptions, for the interested reader, are given in [3] and [4] which have been incorporated in Microsoft's U-prove system as well as the other credential technologies described in [5–7] which have been implemented into IBM's Idemix ABCs system (see [10]).

6 Conclusions

In this paper we have made a first step towards integrating the DT and Ambience Intelligence framework with a Privacy framework, based on the Privacy-ABC technology. Today's technological advances are certainly beyond limits. Nearly every seemingly impossible idea involving mobile and sensor devices is easy to become a reality. The increase in computing speed and memory capacity of electronic components, along with their miniaturization, has made possible the creation of autonomous devices able to accomplish a variety of very demanding tasks.

On a more optimistic angle of view, we already have primitive DT examples which can show the way of implementing the more sophisticated DT concepts. We already have ad-hoc, sensor networks performing useful collective computations based on the signals they sense from their environment. These networks contain units which can also be mobile and move autonomously. On the other hand, we also have human carrying wireless devices with them all the time. We have smart homes containing wireless networks connected to the Internet. We have a very wide frequency spectrum (from 0 to 30 GHz) wisely divided into zones dedicated to specific uses, leaving much space for free communication (no licensing required to operate approved devices), the ISM (Industrial Scientific Medical) zone. We have a great variety of broadband services too 3G or WiMax networks and we will have much more in 4G. We also have an abundance of "hot spots" created by other people or even moving vehicles. As usual, the missing ingredient of exploiting the available technological wealth is people's awareness in privacy issues and political will to regulate, with a focus on the people, the DT concept and its implementation using the technology.

Acknowledgements Research was supported by US Air Force and DTRA grants as well as the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 257782 for the project Attribute-Based Credentials for Trust (ABC4Trust).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
2. Bjones, R., Krontiris, I., Paillier, P., Rannenberg, K.: Integrating anonymous credentials with eIDs for privacy-respecting online authentication. In: Proceedings of EU Annual Privacy Forum. Springer, Berlin (2012)
3. Brands, S.: Rethinking Public Key Infrastructures and Digital Certificates; Building in Privacy, 1st edn. MIT Press, New York (2000). ISBN 0-262-02491-8
4. Brands, S., Demuynck, L., De Decker, B.: A practical system for globally revoking the unlinkable pseudonyms of unknown users. In: Pieprzyk, J., Ghodosi, H., Dawson, Ed. (eds.) 12th Australasian Conference on Information Security and Privacy, ACISP, pp. 400–415. Springer, Berlin (2007)
5. Camenisch, J.: Protecting (anonymous) credentials with the trusted computing group's TPM V1.2. In: Security and Privacy in Dynamic Environments, SEC, 2006, pp. 135–147
6. Camenisch, J.: Thomas Groß: efficient attributes for anonymous credentials. In: ACM Conference on Computer and Communications Security, 2008, pp. 345–356
7. Camenisch, J., Groß, T.: Efficient attributes for anonymous credentials. ACM Trans. Inf. Syst. Secur. (TISSEC) **15**(1), pp. 4:1–4:30 (2012). Article No. 4
8. Camenisch, J., Lysyanskaya, A.: Efficient non-transferable anonymous multi-show credential system with optional anonymity revocation. In: Birgit, Pf. (ed.) Proceedings of EUROCRYPT, Lecture Notes in Computer Science, vol. 2045, pp. 93–118. Springer, Berlin (2001)
9. Camenisch, J., Lysyanskaya, A.: Signature schemes and anonymous credentials from bilinear maps. In: Advances in Cryptology- CRYPTO, 2004, pp. 56–72
10. Camenisch, J., Van Herreweghen, E.: Design and implementation of the idemix anonymous credential system. Research Report RZ 3419, IBM Research Division, June (2002). Also appeared in ACM Computer and Communication Security 2002
11. Jacko, J. (ed.): Ambient, ubiquitous and intelligent interaction. In: Proceedings, Part III, of the 13th International Conference on Human-Computer Interaction (HCI International 2009). Springer, Berlin (2009)
12. Lemos, A.: Mobile communication and new sense of places: a critique of spatialization in cyberculture. Galaxia **16**, 91–109 (2008)
13. Lemos, A.: Pervasive computer games and processes of spatialization: informational territories and mobile technologies. Can. J. Commun. **36**, 277–294 (2011)
14. Liagkou, V., Metakides, G., Pyrgelis, A., Raptopoulos, C., Spirakis, P., Stamatiou, Y.: Privacy preserving course evaluations in Greek higher education institutes: an e-Participation case study with the empowerment of Attribute Based Credentials. In: Proceedings of EU Annual Privacy Forum. Springer, Berlin (2012)

Quantum Analogues of Hermite–Hadamard Type Inequalities for Generalized Convexity

Muhammad Aslam Noor, Khalida Inayat Noor, and Muhammad Uzair Awan

Abstract In this chapter, we discuss quantum calculus and generalized convexity. We briefly discuss some basic concepts and results regarding quantum calculus. Some quantum analogues of derivatives and integrals on finite intervals are discussed. After this we move towards generalized convexity. Examples are given to illustrate the importance and significance of generalized convex sets and generalized convex functions. We establish some quantum Hermite–Hadamard inequalities for generalized convexity. Results proved in this paper may stimulate further research activities.

Keywords: Generalized convexity • Integral inequalities • Quantum inequality • Hermite-Hadmard inequalities • Simpson inequalities • Examples

1 Introduction

Quantum calculus or q-calculus is the study of calculus without limits. The study of quantum calculus had been started by Euler (1707–1783), which first introduced the q in tracks of Newton's infinite series. However it started formerly in early twentieth century with the work of F.H. Jackson. In quantum calculus, we establish q-analogues of mathematical objects which can be recaptured as $q \to 1$. There are two types of q-addition, the Nalli-Ward-Al-Salam q-addition (NWA) and the Jackson-Hahn-Cigler q-addition (JHC). The first one is commutative and associative, while the second one is neither. That is why sometimes more than one q-analogue exists. It has been noticed that quantum calculus is subfield of time scale calculus. Time scale calculus provides a unified framework for studying dynamic equations on the both discrete and continuous domains. In quantum calculus, we are concerned with a specific time scale, called the q-time scale. The quantum calculus can be treated as bridge between Mathematics and Physics. It has large applications in different mathematical areas such as number theory, combinatorics, orthogonal

M.A. Noor (✉) • K.I. Noor • M.U. Awan
COMSATS Institute of Information Technology, Park Road, Islamabad, Pakistan
e-mail: noormaslam@gmail.com; khalidanoor@hotmail.com; awan.uzair@gmail.com

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_18

polynomials, basic hypergeometric functions and in other sciences such as quantum theory, mechanics and in theory of relativity. Due its applications in Mathematics and Physics, this subject has received special attention by many researchers. As a result, quantum calculus has emerged as interdisciplinary subject. For some useful details on quantum calculus, interested readers are referred to [1–7, 14–18, 20–22, 31, 33, 37–40].

The modern analysis directly or indirectly involves the applications of convexity. In theory of convexity we basically study about convex sets and convex functions. In recent years several researchers extended and generalized the classical concepts of convex sets and convex functions in different directions, see [10]. Youness [41] introduced the g-convex sets and g-convex functions. Youness with the help of examples remarked that g-convex sets and g-convex functions are significantly different from convex sets and convex functions, respectively. Noor [25] has shown that optimality conditions of the differentiable g-convex functions can be characterized by a class of variational inequality, which he called as general variational inequality. Inspired by this ongoing research Noor [27] introduced generalized convex sets and generalized convex functions. He studied basic properties of these concepts. Noor has shown that these generalized convex sets and generalized convex functions are nonconvex sets and nonconvex functions, respectively, but enjoys some nice properties which the convex sets and convex functions have. He also introduced another class of variational inequalities that is general non-linear variational inequality. He also noticed that general non-linear variational inequality is quite different from the class of variational inequality introduced and studied by Noor [25]. Cristescu et al. [8, 9] explored some fascinating applications of generalized convexity and also shown that generalized convexity is quite different from g-convexity.

In [36] authors have remarked that, since the publications of the two papers in 1905 and 1906 by J.L.W.V. Jensen, the celebrated Danish engineer and mathematician, the theory of convex functions has experienced a rapid development. This can be attributed to several causes: first, a great many areas in modern analysis directly or indirectly involve the application of convex functions; secondly, convex functions are closely related to the theory of inequalities, and many important inequalities are consequences of the applications of convex functions. For example, the important AG inequality or the general inequality between means of orders r and s, such as Holder's and Minkowski's inequalities, are all consequences of Jensen's inequality for convex functions.

On November 22, 1881, Hermite (1822–1901) sent a letter to the journal Mathesis, which was published letter in Mathesis 3 (1883, p. 82). In this letter Hermite has given following inequalities:

$$(b-a)f\left(\frac{a+b}{2}\right) < \int_a^b f(x)\mathrm{d}x < (b-a)\frac{f(a)+f(b)}{2}; \tag{1}$$

$$(b-a)f\left(\frac{a+b}{2}\right) > \int_a^b f(x)\mathrm{d}x > (b-a)\frac{f(a)+f(b)}{2}. \tag{2}$$

It is interesting to note that this short note of Hermite is nowhere mentioned in mathematical literature, and that these important inequalities (of Hermite) are not widely known as Hermites result. The term convex also stems from this classical note of Hermite. In the booklet on Hermite by Jordan and Mansion (1901), Mansion published a bibliography of Hermite's writings, but this note in Mathesis was not included [23]. Prof. Beckenbach, a leading expert on the history and theory of complex functions, wrote that the inequality (1) was proved by Hadamard in 1893 and apparently was not aware of Hermite's result. Throughout this chapter, we acknowledge this inequality as Hermite–Hadamard's inequality. In recent years, Hermite–Hadamard's inequality has been extensively studied by many researchers. This inequality can be viewed as necessary and sufficient condition for function to be convex. For some recent extensions and generalizations of Hermite–Hadamard's inequality, see [11–13, 19, 24, 26, 28–36, 38, 40].

Motivated by this ongoing research, we in this chapter discuss quantum calculus, generalized convexity and Hermite–Hadamard type inequalities. We review some basic concepts and results of quantum calculus regarding q-derivative and q-antiderivatives. We introduce the concept of generalized convexity. We define generalized convex sets and generalized convex functions, which are mainly due to Noor [27]. We give the examples which are mainly due to Cristescu et al. [8, 9] to show the importance of generalized convexity. We derive some quantum Hermite–Hadamard inequalities for generalized convexity. It is expected that the readers may find this brief chapter useful in their future study and research.

2 Preliminaries of Quantum Calculus

In this section, we discuss some basic known concepts and results pertaining to quantum calculus. For further details of this section, readers may consult [17, 22].

2.1 q-Differentiation

Let us consider

$$\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0}=\frac{\mathrm{d}f}{\mathrm{d}x},$$

the above expression gives the derivative of a function $f(x)$ at $x = x_0$.

If $x = qx_0$ where $0 < q < 1$ is a fixed number and do not take limits, then we enter in the fascinating world of Quantum calculus. The q-derivative of x^n is $[n]x^{n-1}$, where

$$[n] = \frac{q^n - 1}{q - 1},$$

is the q-analogue of n in the sense that n is the limit of $[n]$ as $q \to 1$.

Now we give the formal definition of q-derivative of a function f.

Definition 1. The q-derivative is defined as

$$D_q f(x) = \frac{f(qx) - f(x)}{(q-1)x}. \tag{3}$$

Note that when $q \to 1$, then we have ordinary derivative.

q-derivative has also some nice properties as ordinary derivative:

1. $D_q(\alpha f(x) + \beta g(x)) = \alpha D_q f(x) + \beta D_q g(x)$;
2. $D_q(f(x)g(x)) = f(qx)D_q g(x) + g(x)D_q f(x)$,
   by symmetry, we can write
   $D_q(f(x)g(x)) = f(x)D_q g(x) + g(qx)D_q f(x)$,
3. $D_q\left(\frac{f(x)}{g(x)}\right) = \frac{g(qx)D_q f(x) - f(qx)D_q g(x)}{g(x)g(qx)}$.

Now one can ask a question that what is the q-analogue of chain rule for derivatives? The answer is the there doesn't exist a general chain rule for q-derivatives.

Let us elaborate the definition of q-derivative with the help of an example, that is, how to compute the derivative of $f(x) = x^n$ in quantum calculus? Where n is a positive integer.

$$D_q x^n = \frac{(qx)^n - x^n}{(q-1)x} x^{n-1} = [n]x^{n-1},$$

which resembles to the ordinary derivative of x^n as $q \to 1$,

$$[n] = q^{n-1} + \cdots + 1 = 1 + 1 + \cdots + 1 = n.$$

It is worth to mention here that $[n]$ plays the same role in quantum calculus as n plays the role in ordinary calculus.

2.2 q-Antiderivatives

Now we move to words q-antiderivatives of function f.

Definition 2. The function $F(x)$ is a q-antiderivative of $f(x)$ if $D_qF(x) = f(x)$. It is denoted by

$$\int f(x)\mathrm{d}_q x. \tag{4}$$

Remark 1. It is worth to mention here that we have written "a" q-antiderivative instead of "the" q-antiderivative, because, as in ordinary calculus, an antiderivative is not unique.

Proposition 1. *Let $0 < q < 1$. Then, up to adding a constant, any function $f(x)$ has at most one q-antiderivative that is continuous at $x = 0$.*

We now give the definition of Jackson integral.

Definition 3. The Jackson integral of $f(x)$ is defined as

$$\int f(x)\mathrm{d}_q x = (1-q)x\sum_{j=0}^{\infty} q^j f(q^j x). \tag{5}$$

It is evident from above definition that

$$\begin{aligned}\int f(x)D_q g(x)\mathrm{d}_q x &= (1-q)x\sum_{j=0}^{\infty} q^j f(q^j x)D_q g(q^j x)\\ &= (1-q)x\sum_{j=0}^{\infty} q^j f(q^j x)\frac{g(q^j x) - g(q^{j+1}x)}{(1-q)q^j x}.\end{aligned}$$

Theorem 1. *Suppose $0 < q < 1$. If $|f(x)x^\alpha|$ is bounded on the interval $(0, A]$ for some $0 \le \alpha < 1$, then the Jackson integral defined by (3) converges to a function $F(x)$ on $(0, A]$, which is a q-antiderivative of $f(x)$. Moreover, $F(x)$ is continuous at $x = 0$ with $F(0) = 0$.*

Proof. Let

$$|f(x)x^\alpha| < M \quad \text{on } A.$$

Then, for any $0 < x \le A$, $j \ge 0$,

$$|f(q^j x)| < M(q^j x)^{-\alpha}.$$

This implies for $0 < x \le A$, we have

$$|q^j f(q^j x)| < Mx^{-\alpha}(q^{1-\alpha})^j. \tag{6}$$

Since $1-\alpha > 0$ and $0 < q < 1$, thus our series is majorized by convergent geometric series. Hence the right-hand side of (5) converges pointwise to some function $F(x)$. From (5) it is clear that $F(0) = 0$. The fact that $F(x)$ is continuous at $x = 0$, that is, $F(x) \to 0$ as $x \to 0$, is clear if we consider using (4),

$$\left|(1-q)x\sum_{j=0}^{\infty} q^j f(q^j x)\right| < \frac{M(1-q)x^{1-\alpha}}{1-q^{1-\alpha}}, \quad 0 < x \le A.$$

Let us q-differentiate it:

$$\begin{aligned} D_q F(x) &= \frac{1}{(1-q)x}\left((1-q)x\sum_{j=0}^{\infty} q^j f(q^j x) - (1-q)qx\sum_{j=0}^{\infty} q^j f(q^{j+1}x)\right) \\ &= \sum_{j=0}^{\infty} q^j f(q^j x) - \sum_{j=0}^{\infty} q^{j+1} f(q^{j+1}x) \\ &= \sum_{j=0}^{\infty} q^j f(q^j x) - \sum_{j=1}^{\infty} q^j f(q^j x) = f(x). \end{aligned}$$

It is very much clear that if $x \in (0, A]$ and $0 < q < 1$, then $qx \in (0, A]$ and the q-differentiation is valid. This completes the proof. □

Now we give an example where Jackson integral fails.

Example 1. Consider $f(x) = \frac{1}{x}$, then

$$D_q \log x = \frac{\log(qx) - \log(x)}{(q-1)x} = \frac{\log q}{q-1}\frac{1}{x},$$

and

$$\int \frac{1}{x} \mathrm{d}_q x = \frac{q-1}{\log q} \log x.$$

However, the Jackson formula gives

$$\int \frac{1}{x} \mathrm{d}_q x = (1-q)\sum_{j=0}^{\infty} 1 = \infty.$$

We now define the definite q-integral.

Definition 4. Let $0 < a < b$. The definite q-integral is defined as

$$\int_0^b f(x)\mathrm{d}_q x = (1-q)b\sum_{j=0}^{\infty} q^j f(q^j b), \tag{7}$$

provided the sum converge absolutely.

A more general formula for definite integrals is given as

$$\int_0^b f(x)\mathrm{d}_q g(x) = \sum_{j=0}^{\infty} f(q^j b)(g(q^j b) - g(q^{j+1}b)).$$

Remark 2. From above definition of definite q-integral in a generic interval $[a, b]$ is given by

$$\int_a^b f(x)\mathrm{d}_q x = \int_0^b f(x)\mathrm{d}_q x - \int_0^a f(x)\mathrm{d}_q x.$$

Remark 3. Note that the above definition conforms the fact that the Jackson integral vanishes at $x = O$. Geometrically, the integral in (7) corresponds to the area of the union of an infinite number of rectangles.

Definition 5. The improper q-integral of $f(x)$ on $[0, +\infty)$ is defined to be

$$\int_0^{\infty} f(x)\mathrm{d}_q x = \sum_{j=-\infty}^{\infty} \int_{q^{j+1}}^{q^j} f(x)\mathrm{d}_q x, \tag{8}$$

where $0 < q < 1$.

Also

$$\int_0^{\infty} f(x)\mathrm{d}_q x = \sum_{j=-\infty}^{\infty} \int_{q^j}^{q^{j+1}} f(x)\mathrm{d}_q x, \tag{9}$$

when $q > 1$.

Proposition 2. *The improper q-integral defined above converges if $f(x)$ is bounded in a neighborhood of $x = 0$ with some $\alpha < 1$ and for sufficiently large x with some $\alpha > 1$.*

In ordinary calculus, a derivative is defined as the limit of a ratio, and a definite integral is defined as the limit of an infinite sum. Their subtle and surprising relation is given by the Newton–Leibniz formula, also called the fundamental theorem of calculus. Now we give the q-analogue of fundamental theorem of calculus.

Theorem 2. *If $F(x)$ is an antiderivative of $f(x)$ and $F(x)$ is continuous at $x = 0$, we have*

$$\int_a^b f(x)\mathrm{d}_q x = F(b) - F(a), \tag{10}$$

where $0 \leq a < b \leq \infty$.

Proof. Let $F(x)$ is continuous at $x = 0$, since $F(x)$ is given by the Jackson formula, up to adding a constant, that is,

$$F(x) = (1-q)x\sum_{j=0}^{\infty} q^j f(q^j x) + F(0).$$

This implies

$$\begin{aligned}\int_0^a f(x)\mathrm{d}_q x &= (1-q)a\sum_{j=0}^{\infty} q^j f(q^j a)\\ &= F(a) - F(0).\end{aligned} \tag{11}$$

Similarly

$$\int_0^b f(x)\mathrm{d}_q x = F(b) - F(0). \tag{12}$$

From (11) and (12), we have

$$\int_a^b f(x)\mathrm{d}_q x = F(b) - F(a).$$

This completes the proof. □

Remark 4. Putting $a = q^{j+1}$ (or q^j) and $b = q^j$ (or $qj + 1$), where $0 < q < 1$ (or $q > 1$), and considering the definition of improper q-integral, we notice that (10) is true for $b = \infty$ as well provided if $F(x)$ exists.

Corollary 1. *If $f'(x)$ exists in a neighborhood of $x = 0$ and is continuous at $x = 0$, where $f(x)$ denotes the ordinary derivative of $f(x)$, we have*

$$\int_a^b D_q f(x) \mathrm{d}_q x = f(b) - f(a). \tag{13}$$

Proof. Using L'Hospital's rule, we have

$$\begin{aligned} \lim_{x\to 0} D_q f(x) &= \lim_{x\to 0} \frac{f(qx) - f(x)}{(q-1)x} \\ &= \lim_{x\to 0} \frac{qf'(qx) - f'(x)}{q-1} = f'(0). \end{aligned}$$

Hence $D_q f(x)$ can be made continuous at $x = 0$ if we define $(D_q f)(0) = f'(0)$, and (13) follows from the fundamental theorem of calculus. □

Remark 5. An important difference between the definite q-integral and its ordinary counterpart is that even if we are integrating a function on an interval like $[1, 2]$, we have to care about its behavior at $x = 0$. This has to do with the definition of the definite q-integral and the condition for the convergence of the Jackson integral.

Remark 6. Now suppose $f(x)$ and $g(x)$ are two functions whose ordinary derivatives exist in a neighborhood of $x = 0$ and are continuous at $x = 0$. Using the product rule, we have

$$D_q(f(x)g(x)) = f(x)D_q g(x) + g(qx)D_q f(x).$$

Since the product of differentiable functions is also differentiable, so using Corollary 1, we have

$$\int_a^b f(x)\mathrm{d}_q g(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(qx)\mathrm{d}_q f(x).$$

This above formula is for q-integration by parts. Note that $b = \infty$ is allowed as well.

2.3 Riemann-Type q-Integral

Rajkovic et al. [37] defined Riemann type q-integrals as:

$$R_q(f; a, b) = (b-a)(1-q)\sum_{k=0}^{\infty} f(a + (b-a)q^k)q^k. \tag{14}$$

Taf et al. [38] extended the above definition by dividing the integral as:

$$\frac{2}{b-a}\int_a^b f(x)\mathrm{d}_q^R x$$
$$= (1-q)\sum_{k=0}^{\infty}\left(f\left(\frac{a+b}{2}+q^k\left(\frac{b-a}{2}\right)\right)+f\left(\frac{a+b}{2}-q^k\left(\frac{b-a}{2}\right)\right)\right)q^k. \tag{15}$$

Using q-Jackson integral, we have

$$\frac{2}{b-a}\int_a^b f(x)\mathrm{d}_q^R x$$
$$= \int_{-1}^{1} f\left(\frac{1-t}{2}a+\frac{1+t}{2}b\right)\mathrm{d}_q t+\int_{-1}^{1} f\left(\frac{1+t}{2}a+\frac{1-t}{2}b\right)\mathrm{d}_q t. \tag{16}$$

3 Quantum Calculus on Finite Intervals

Now we recall the basic definitions and results of quantum calculus on finite intervals. These results are mainly due to Tariboon et al. [39, 40].

Let $J=[a,b]\subseteq\mathbb{R}$ be an interval and $0<q<1$ be a constant. The q-derivative of a function $f:J\to\mathbb{R}$ at a point $x\in J$ on $[a,b]$ is defined as follows.

Definition 6. Let $f:J\to\mathbb{R}$ be a continuous function and let $x\in J$. Then q-derivative of f on J at x is defined as

$${}_aD_qf(x)=\frac{f(x)-f(qx+(1-q)a)}{(1-q)(x-a)},\quad x\neq a. \tag{17}$$

It is obvious that ${}_aD_qf(a)=\lim_{x\to a}{}_aD_qf(x)$.

A function f is q-differentiable on J if ${}_aD_qf(x)$ exists for all $x\in J$. Also if $a=0$ in (17), then ${}_0D_qf=D_qf$, where D_q is the q-derivative of the function f [17, 22] defined as

$$D_qf(x)=\frac{f(x)-f(qx)}{(1-q)x}.$$

Remark 7. Let $f : J \to \mathbb{R}$ is a continuous function. Let us define the second-order q-derivative on interval J, which is denoted by ${}_aDf_qf$, provided ${}_aD_qf$ is q-differentiable on J with ${}_aDf_q^2f = {}_a\ D_q({}_aD_qf) : J \to \mathbb{R}$. Similarly, one can define higher order q-derivative on J, ${}_aD_q^n : J_k \to \mathbb{R}$.

Let us elaborate above definitions with the help of an example.

Example 2. Let $x \in [a, b]$ and $0 < q < 1$. Then, for $x \neq a$, we have

$$\begin{aligned} {}_aD_qx^2 &= \frac{x^2 - (qx + (1-q)a)^2}{(1-q)(x-a)} \\ &= \frac{(1+q)x^2 - 2qax - (1-q)x^2}{x-a} \\ &= (1+q)x + (1-q)a. \end{aligned}$$

Note that when $x = a$, we have $\lim\limits_{x\to a}({}_aD_qx^2) = 2a$.

Definition 7. Let $f : J \to \mathbb{R}$ is a continuous function. A second-order q-derivative on J, which is denoted as ${}_aD_q^2f$, provided ${}_aD_qf$ is q-differentiable on J is defined as ${}_aD_q^2f = {}_a\ D_q({}_aD_qf) : J \to \mathbb{R}$. Similarly higher order q-derivative on J is defined by ${}_aD_q^nf =: J_k \to \mathbb{R}$.

Lemma 1. *Let $\alpha \in \mathbb{R}$, then*

$${}_aD_q(x-a)^\alpha = \Big(\frac{1-q^\alpha}{1-q}\Big)(x-a)^{\alpha-1}.$$

Tariboon et al. [39, 40] defined the q-integral as follows:

Definition 8. Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a continuous function. Then q-integral on I is defined as

$$\int_a^x f(t)\,{}_a\mathrm{d}_qt = (1-q)(x-a)\sum_{n=0}^{\infty} q^n f(q^nx + (1-q^n)a), \tag{18}$$

for $x \in J$.

If $a = 0$ in (18), then we have the classical q-integral, that is

$$\int_0^x f(t)\,{}_0\mathrm{d}_qt = (1-q)x\sum_{n=0}^{\infty} q^n f(q^nx), \quad x \in [0, \infty).$$

For more details, see [17, 22].

Moreover, if $c \in (a, x)$, then the definite q-integral on J is defined by

$$\begin{aligned}\int_c^x f(t)_a\mathrm{d}_q t &= \int_a^x f(t)_a\mathrm{d}_q t - \int_a^c f(t)_a\mathrm{d}_q t \\ &= (1-q)(x-a)\sum_{n=0}^{\infty} q^n f(q^n x + (1-q^n)a) \\ &= (1-q)(c-a)\sum_{n=0}^{\infty} q^n f(q^n c + (1-q^n)a).\end{aligned}$$

Example 3. Let a constant $c \in J$, then

$$\begin{aligned}\int_c^b (t-c)_a\mathrm{d}_q t &= \int_a^b (t-c)_a\mathrm{d}_q t - \int_a^c (t-c)_a\mathrm{d}_q t \\ &= \left[\frac{(t-a)(t+qa)}{1+q} - ct\right]_a^b - \left[\frac{(t-a)(t+qa)}{1+q} - ct\right]_a^c \\ &= \frac{b^2 - (1+q)bc + qc^2}{1+q} - \frac{a(1-q)(b-c)}{1+q}.\end{aligned}$$

Note that when $q \to 1$, then the above integral reduces to the classical integration

$$\int_c^b (t-c)\mathrm{d}t = \frac{(b-c)^2}{2}.$$

Theorem 3. *Let $f : I \to \mathbb{R}$ be a continuous function, then*

1. ${}_aD_q \int_a^x f(t)\,{}_a\mathrm{d}_q t = f(x)$
2. $\int_c^x {}_aD_q f(t)_a\mathrm{d}_q t = f(x) - f(c)$ *for* $x \in (c, x)$.

Theorem 4. *Let $f, g : I \to \mathbb{R}$ be a continuous functions, $\alpha \in \mathbb{R}$, then $x \in J$*

1. $\int_a^x [f(t) + g(t)]\,{}_a\mathrm{d}_q t = \int_a^x f(t)\,{}_a\mathrm{d}_q t + \int_a^x g(t)\,{}_a\mathrm{d}_q t$
2. $\int_a^x (\alpha f(t))(t)\,{}_a\mathrm{d}_q t = \alpha \int_a^x [f(t) + g(t)]\,{}_a\mathrm{d}_q t$
3. $\int_a^x f(t)\,{}_a\mathrm{d}_q t \int_a^x g(t)\,{}_a\mathrm{d}_q t = (fg)|_c^x - \int_c^x g(qt + (1-q)a)_a D_q f(t)_a\mathrm{d}_q t$ *for* $x \in (a, x)$.

Lemma 2. *Let $\alpha \in \mathbb{R} \setminus \{-1\}$, then*

$$\int_a^x (t-a)_a^{\alpha} \mathrm{d}_q t = \Big(\frac{1-q}{1-q^{\alpha+1}}\Big)(x-a)^{\alpha+1}.$$

Proof. Let $f(x) = (x-a)^{\alpha+1}$, $x \in J$ and $\alpha \in \mathbb{R} \setminus \{-1\}$, then by definition, we have

$$\begin{aligned} {}_aD_q f(x) &= \frac{(x-a)^{\alpha+1} - (qx + (1-q)a - a)^{\alpha+1}}{(1-q)(x-a)} \\ &= \frac{(x-a)^{\alpha+1} - q^{\alpha+1}(x-a)^{\alpha+1}}{(1-q)(x-a)} \\ &= \left(\frac{1-q^{\alpha+1}}{1-q}\right)(x-a)^{\alpha}. \end{aligned} \tag{19}$$

Applying q-integral on J for (19), we obtain the required result. □

Example 4. Let $f(x) = x$ for $x \in J$, then, we have

$$\begin{aligned} \int_a^x f(t)_a \mathrm{d}_q t = \int_a^x t_a \mathrm{d}_q t &= (1-q)(x-a)\sum_{n=0}^{\infty} q^n(q^n x + (1-q^n)a) \\ &= \frac{(x-a)(x+qa)}{1+q}. \end{aligned}$$

4 Basic Concepts and Results for Generalized Convexity

In this section, we recall the concept of generalized convex sets and generalized convex functions, respectively.

Definition 9 ([27]). Let K_φ be any set in H. The set K_φ is said to be generalized convex with respect to an arbitrary function $\varphi : H \to H$ such that

$$(1-t)u + t\varphi(v) \in K_\varphi, \quad \forall u, v \in H : u, \varphi(v) \in K_\varphi, t \in [0,1].$$

If $\varphi = I$, the identity function, then the definition of generalized convex set coincides with the definition of classical convex set.

Note that the generalized convex sets are distinctly different than that of Youness's generalized convex set [41].

Now we give some examples which are mainly due to Cristescu et al. [8, 9]. These examples show the significance of generalized convex sets.

Example 5 ([8]). One of the most important goals of the International Union of Railways (U.I.C.) is to enable the railway companies to measure the impact of their activity on the environment (see the U.I.C. guide). The environment indicators in the domain of railway transport defined under U.I.C. and presented into the above-mentioned guide include the level of noise, which should be, in normal conditions, in the interval $[0, 50]db(A)$. The actual noise level produced by wagons is $[125, 130]db(A)$. The noise level around the railway stations located in towns is represented by the set $[0, 50] \cup [125, 130]$. By relocating the railway transport system outside the towns, the resulted level of noise becomes $[0, 50]$. Let us denote by $\varphi : \mathbb{R} \to \mathbb{R}$ the function defined by

$$\varphi(x) = \begin{cases} x & \text{if } x \in [0, 50] \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

which is the function describing the efforts of kipping the normal level of sound, which works under this project. Then the set $[0, 50] \cup [125, 130]$ is generalized convex.

Other examples are easy to find in the domain of image processing, in which a transformation of the real plane $\mathbb{R}^2$ into a set of grid-points, $\mathbb{Z}^2$ for example, is necessary. In order to present this type of examples we need to choose a transformation of the space, which performs the space digitization. The general definition of this kind of transformations is

Definition 10. A function $f : \mathbb{R}^n \to \mathbb{Z}^n$ is said to be a method of digitization of $\mathbb{R}^n$ into $\mathbb{Z}^n$ if $f(x) = x$ whenever $x \in \mathbb{Z}^n$.

In what follows we assume that $n = 2$ and $\varphi = E = f$ is the digitization method used in black and white picture processing by Rosenfeld (1969) and in colored image processing by Chassery (1978). It is defined by $f : \mathbb{R}^2 \to \mathbb{Z}^2$, $f(x, y) = (i, j), i \in \mathbb{Z}, j \in \mathbb{Z}$ whenever $(x, y) \in [i-1/2, i+1/2) \times [j-1/2, j+1/2)$.

Example 6 ([8]). The set $A := B \cup < (i, j), (i+m, j) >$, whenever $i \in \mathbb{Z}, j \in \mathbb{Z}, m \in \mathbb{Z}$ and $B \subseteq [i-1/2, i+m+1/2) \times [j-1/2, j+1/2)$ is a union of triangles having one side $< (i, j), (i + m, j) >$ is generalized convex. Indeed, considering two points $x, y \in A$, there are two numbers $k \in \mathbb{Z}$ and $l \in \mathbb{Z}$ such that $x \in [i + k - 1/2, i + k + 1/2) \times [j - 1/2, j + 1/2)$ and $y \in [i + l - 1/2, i + l + 1/2) \times [j - 1/2, j + 1/2)$. Therefore $\varphi(y) = (i + l, j)$. Then for any $t \in [0, 1]$, there is the integer s between k and l such that $tx + (1-t)\varphi(y) \in [i + s - 1/2, i + s + 1/2) \times [j - 1/2, j + 1/2) \subseteq A$ since B is an union of triangles having one side $< (i, j), (i + m, j) >$. It means that A is generalized convex. In the same manner one can take vertical columns of pixels and obtain generalized convex sets.

Definition 11 ([27]). A function $f : K_\varphi \to H$ is said to be generalized convex, if there exists an arbitrary function $\varphi : H \to H$ such that

$$f((1-t)u + t\varphi(v)) \leq (1-t)f(u) + tf(\varphi(v)),$$
$$\forall u, v \in H : u, \varphi(v) \in K_\varphi, t \in [0, 1]. \tag{21}$$

Definition 12 ([27]). The function $f : K_\varphi \to H$ is said to be generalized quasi convex, if there exists an arbitrary function $\varphi : H \to H$ such that

$$f((1-t)u + t\varphi(v)) \leq \max\{f(u), f(\varphi(v))\},$$
$$\forall u, v \in H : u, \varphi(v) \in K_\varphi, t \in [0,1]. \tag{22}$$

Noor [26] extended the classical Hermite–Hadamard inequality for generalized convex functions.

Theorem 5 ([26]). *Let $f : [a, \varphi(b)] \to \mathbb{R}$ be a generalized convex function with respect to an arbitrary function $\varphi : H \to H$. Then, we have*

$$f\left(\frac{a+\varphi(b)}{2}\right) \leq \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\mathrm{d}\varphi(x) \leq \frac{f(a)+f(\varphi(b))}{2}.$$

Theorem 6 ([26]). *Let $f, w : [a, \varphi(b)] \to \mathbb{R}$ be generalized convex functions with respect to an arbitrary function $\varphi : H \to H$. Then for all $t \in [0,1]$, we have*

$$2f\left(\frac{a+\varphi(b)}{2}\right) w\left(\frac{a+\varphi(b)}{2}\right) - \left[\frac{1}{6}M(a,\varphi(b)) + \frac{1}{2}N(a,\varphi(b))\right]$$
$$\leq \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))w(\varphi(x))\mathrm{d}\varphi(x) \leq \frac{1}{3}M(a,\varphi(b)) + \frac{1}{6}N(a,\varphi(b)),$$

where

$$M(a,\varphi(b)) = f(a)w(a) + f(\varphi(b))w(\varphi(b)), \tag{23}$$

and

$$N(a,\varphi(b)) = f(a)w(\varphi(b)) + f(\varphi(b))w(a). \tag{24}$$

5 Some Quantum Estimates of Hermite–Hadamard Type Inequalities Via Generalized Convexity

In this section, we establish some quantum estimates of Hermite–Hadamard type inequalities via generalized convexity.

Theorem 7. *Let $f : [a, \varphi(b)] \to \mathbb{R}$ be generalized convex continuous function with respect to an arbitrary function $\varphi : H \to H$. Then*

$$f\left(\frac{a+\varphi(b)}{2}\right) \leq \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\mathrm{d}_q^R\varphi(x) \leq \frac{f(a)+f(\varphi(b))}{2}.$$

Proof. Since f is generalized convex function, then, we have

$$f\left(\frac{a+\varphi(b)}{2}\right) \leq \frac{1}{2}\left[f\left(\frac{1-t}{2}a+\frac{1+t}{2}\varphi(b)\right)+f\left(\frac{1+t}{2}a+\frac{1-t}{2}\varphi(b)\right)\right].$$

Riemann type q-integrating above inequality with respect to t on $[-1, 1]$, we have

$$f\left(\frac{a+\varphi(b)}{2}\right) \leq \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\mathrm{d}_q^R\varphi(x). \tag{25}$$

Also

$$f\left(\frac{1-t}{2}a+\frac{1+t}{2}\varphi(b)\right) \leq \left(\frac{1-t}{2}\right)f(a)+\left(\frac{1+t}{2}\right)f(\varphi(b)).$$

Riemann type q-integrating above inequality with respect to t on $[-1, 1]$, we have

$$\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\mathrm{d}_q^R\varphi(x) \leq \frac{f(a)+f(\varphi(b))}{2}. \tag{26}$$

Combining (25) and (26) completes the proof. □

Next we derive some quantum estimates of Hermite–Hadamard type inequalities via generalized convexity on finite intervals.

Theorem 8. *Let $f : J = [a, \varphi(b)] \to \mathbb{R}$ be generalized convex continuous function on J with respect to an arbitrary function $\varphi : H \to H$. Then for $0 < q < 1$, we have*

$$f\left(\frac{a+\varphi(b)}{2}\right) \leq \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(t)_a\mathrm{d}_q t \leq \frac{qf(a)+f(\varphi(b))}{1+q}. \tag{27}$$

Proof. Let f be a generalized convex function on $[a, \varphi(b)]$, then by taking q-integration with respect to t on $[0, 1]$, we have

$$
\begin{aligned}
f\left(\frac{a+\varphi(b)}{2}\right) &= \int_0^1 f\left(\frac{(1-t)a+t\varphi(b)+ta+(1-t)\varphi(b)}{2}\right) {}_0\mathrm{d}_q t \\
&\leq \frac{1}{2}\left[\int_0^1 f((1-t)a+t\varphi(b)){}_0\mathrm{d}_q t + \int_0^1 f(ta+(1-t)\varphi(b)){}_0\mathrm{d}_q t\right] \\
&= \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(t){}_a\mathrm{d}_q t \\
&= \int_0^1 f((1-t)a+t\varphi(b)){}_0\mathrm{d}_q t \\
&\leq f(a)\int_0^1 (1-t){}_0\mathrm{d}_q t + f(\varphi(b))\int_0^1 t\,{}_0\mathrm{d}_q t \\
&= \frac{qf(a)+f(\varphi(b))}{1+q},
\end{aligned}
$$

where by definition, we have

$$
\begin{aligned}
&\int_0^1 f((1-t)a+t\varphi(b)){}_0\mathrm{d}_q t \\
&\quad = (1-q)\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&\quad = \frac{(1-q)(\varphi(b)-a)}{\varphi(b)-a}\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&\quad = \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(t){}_a\mathrm{d}_q t,
\end{aligned}
$$

and

$$
\int_0^1 t{}_0\mathrm{d}_q t = \frac{1}{1+q}, \qquad \int_0^1 (1-t){}_0\mathrm{d}_q t = \frac{q}{1+q}.
$$

This completes the proof. □

Note that when $\varphi = I$, the identity function, our result coincides with Theorem 3.2 [40].

Theorem 9. *Let $f, w : I = [a, \varphi(b)] \to \mathbb{R}$ be generalized convex functions, then*

$$\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))w(\varphi(x))_a\mathrm{d}_q\varphi(x)$$
$$\leq \theta_1 f(a)w(a) + \theta_2 N(a,\varphi(b)) + \theta_3 f(\varphi(b))w(\varphi(b)),$$

where

$$\theta_1 = \frac{q(1+q^2)}{(1+q)(1+q+q^2)};$$
$$\theta_2 = \frac{q^2}{(1+q)(1+q+q^2)};$$
$$\theta_3 = \frac{1}{1+q+q^2},$$

and

$$N(a,\varphi(b)) = f(a)w(\varphi(b)) + f(\varphi(b))w(a).$$

Proof. Since f and w are generalized convex functions, then

$$f((1-t)a + tg(b)) \leq tf(a) + (1-t)f(\varphi(b)),$$
$$w((1-t)a + tg(b)) \leq tw(a) + (1-t)w(\varphi(b)).$$

Multiplying above inequalities, we have

$$f((1-t)a + tg(b))w((1-t)a + tg(b))$$
$$\leq (1-t)^2 f(a)w(a) + t(1-t)\{f(a)w(\varphi(b))$$
$$+ f(\varphi(b))w(a)\} + t^2 f(\varphi(b))w(\varphi(b)).$$

Taking q-integral of both sides of above inequality with respect to t on $[0,1]$, we have

$$\int_0^1 f((1-t)a + t\varphi(b))w((1-t)a + t\varphi(b))_0\mathrm{d}_q t$$
$$\leq f(a)w(a)\int_0^1 (1-t)^2_0\mathrm{d}_q t + \{f(a)w(\varphi(b)) + f(\varphi(b))w(a)\}\int_0^1 t(1-t)_0\mathrm{d}_q t$$
$$+ f(\varphi(b))w(\varphi(b))\int_0^1 t^2_0\mathrm{d}_q t.$$

This implies that

$$\begin{aligned}
&\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))w(\varphi(x))_a\mathrm{d}_q\varphi(x)\\
&\leq\left[\frac{q(1+q^2)}{(1+q)(1+q+q^2)}\right]f(a)w(a)\\
&\quad+\left[\frac{q^2}{(1+q)(1+q+q^2)}\right]\{f(a)w(\varphi(b))+f(\varphi(b))w(a)\}\\
&\quad+\left[\frac{1}{1+q+q^2}\right]f(\varphi(b))w(\varphi(b)).
\end{aligned}$$

This completes the proof. □

Theorem 10. *Since f and w are generalized convex functions, then*

$$\begin{aligned}
&2f\left(\frac{a+\varphi(b)}{2}\right)w\left(\frac{a+\varphi(b)}{2}\right)-\frac{2q^2M(a,\varphi(b))+(1+2q+q^3)N(a,\varphi(b))}{2(1+q)(1+q+q^2)}\\
&\leq\frac{1}{(\varphi(b)-a)}\int_a^{\varphi(b)} f(\varphi(x))w(\varphi(x))_a\mathrm{d}_qx,
\end{aligned}$$

where $M(a,\varphi(b))$ and $N(a,\varphi(b))$ are given by (23) and (24), respectively.

Proof. Since f and w are generalized convex function, then

$$\begin{aligned}
&f\left(\frac{a+\varphi(b)}{2}\right)w\left(\frac{a+\varphi(b)}{2}\right)\\
&\leq\frac{1}{4}[f((1-t)a+t\varphi(b))+f(ta+(1-t)\varphi(b))\\
&\quad+w((1-t)a+t\varphi(b))+w(ta+(1-t)\varphi(b))]\\
&\leq\frac{1}{4}[f((1-t)a+t\varphi(b))w((1-t)a+t\varphi(b))\\
&\quad+f(ta+(1-t)\varphi(b))w(ta+(1-t)\varphi(b))\\
&\quad+[f(a)w(a)+f(\varphi(b))w(\varphi(b))]\{2t(1-t)\}\\
&\quad+[f(a)w(\varphi(b))+f(\varphi(b))w(a)]\{t^2+(1-t)^2\}].
\end{aligned}$$

Applying q-integration with respect to t on $[0,1]$, we have

$$
\begin{aligned}
& f\left(\frac{a+\varphi(b)}{2}\right) w\left(\frac{a+\varphi(b)}{2}\right) \\
& \leq \frac{1}{4}\Bigg[\int_0^1 [f((1-t)a+t\varphi(b))w((1-t)a+t\varphi(b)) \\
& \quad + f(ta+(1-t)\varphi(b))w(ta+(1-t)\varphi(b))]\,{}_0\mathrm{d}_q t \\
& \quad + [f(a)w(a)+f(\varphi(b))w(\varphi(b))]\int_0^1 \{2t(1-t)\}_0\mathrm{d}_q t \\
& \quad + [f(a)w(\varphi(b))+f(\varphi(b))w(a)]\int_0^1 \{t^2+(1-t)^2\}_0\mathrm{d}_q t\Bigg] \\
& = \frac{1}{2(\varphi(b)-a)}\int_a^{\varphi(b)} f(\varphi(x))w(\varphi(x))_a\mathrm{d}_q x \\
& \quad + \frac{1}{4}\Bigg[\frac{2q^2\{f(a)w(a)+f(\varphi(b))w(\varphi(b))\}}{(1+q)(1+q+q^2)} \\
& \quad + \frac{(1+2q+q^3)[f(a)w(\varphi(b))+f(\varphi(b))w(a)]}{(1+q)(1+q+q^2)}\Bigg].
\end{aligned}
$$

This completes the proof. □

We now derive the following auxiliary result, which will be useful in proving our main results.

Lemma 3. *Let $f : I = [a,\varphi(b)] \subset \mathbb{R} \to \mathbb{R}$ be a q-differentiable function on I° (the interior of I) with ${}_aD_q$ be continuous and integrable on I where $0 < q < 1$, then*

$$
\begin{aligned}
& \frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))_a\mathrm{d}_q\varphi(x) - \frac{qf(a)+f(\varphi(b))}{1+q} \\
& = \frac{q(\varphi(b)-a)}{1+q}\int_0^1 (1-(1+q)t)_aD_q f((1-t)a+t(\varphi(b)))\,{}_0\mathrm{d}_q t.
\end{aligned}
$$

Proof. From Definitions 6 and 8, we have

$$\begin{aligned}
&\int_0^1 (1-(1+q)t)_aD_qf((1-t)a+t(\varphi(b)))_0\mathrm{d}_qt \\
&= \int_0^1 (1-(1+q)t)\left(\frac{f((1-t)a+t(\varphi(b)))-f((1-tq)a+qt(\varphi(b)))}{(1-q)(\varphi(b)-a)t}\right)_0\mathrm{d}_qt \\
&= \frac{1}{\varphi(b)-a}\left\{\sum_{n=0}^{\infty} f((1-q^n)a+q^n\varphi(b))-\sum_{n=0}^{\infty} f((1-q^{n+1})a+q^{n+1}\varphi(b))\right\} \\
&\quad -\frac{1+q}{\varphi(b)-a}\left\{\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b))\right. \\
&\quad \left.-\sum_{n=0}^{\infty} q^n f((1-q^{n+1})a+q^{n+1}\varphi(b))\right\} \\
&= \frac{f(\varphi(b))-f(a)}{\varphi(b)-a}-\frac{1+q}{\varphi(b)-a}\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&\quad -\frac{1+q}{q(\varphi(b)-a)}\sum_{n=1}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&= \frac{f(\varphi(b))-f(a)}{\varphi(b)-a}-\frac{1+q}{\varphi(b)-a}\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&\quad -\frac{1+q}{q(\varphi(b)-a)}f(\varphi(b))+\frac{1+q}{q(\varphi(b)-a)}\sum_{n=0}^{\infty} q^n f((1-q^n)a+q^n\varphi(b)) \\
&= \frac{-f(\varphi(b))-qf(a)}{q(\varphi(b)-a)}+\frac{1+q}{q(\varphi(b)-a)^2}\int_a^{\varphi(b)} f(\varphi(x))_a\mathrm{d}_q\varphi(x).
\end{aligned}$$

Multiplying both sides by $\frac{q(\varphi(b)-a)}{1+q}$ completes the proof. □

Theorem 11. *Let $f : I = [a, \varphi(b)] \subset \mathbb{R} \to \mathbb{R}$ be a q-differentiable function on I° (the interior of I) with ${}_aD_q$ be continuous and integrable on I where $0 < q < 1$. If $|{}_aD_qf|^r$, $r \geq 1$ is generalized convex function, then*

$$\left|\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))_a\mathrm{d}_q\varphi(x)-\frac{qf(a)+f(\varphi(b))}{1+q}\right|$$

$$\leq \frac{q(\varphi(b)-a)}{1+q}\left(\frac{q(2+q+q^3)}{(1+q)^3}\right)^{1-\frac{1}{r}} \times\left[\frac{q(1+4q+q^2)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(a)|^r + \frac{q(1+3q^2+2q^3)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(\varphi(b))|^r\right]^{\frac{1}{r}}.$$

Proof. Since $|{}_aD_qf|^r$ is generalized convex function, so from Lemma 3 and using power mean inequality, we have

$$\begin{aligned}
&\left|\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\,{}_a\mathrm{d}_q\varphi(x) - \frac{qf(a)+f(\varphi(b))}{1+q}\right| \\
&= \left|\frac{q(\varphi(b)-a)}{1+q}\int_0^1 (1-(1+q)t)\,{}_aD_qf((1-t)a+t(\varphi(b)))\,{}_0\mathrm{d}_qt\right| \\
&\leq \frac{q(\varphi(b)-a)}{1+q}\left(\int_0^1 |1-(1+q)t|\,{}_0\mathrm{d}_qt\right)^{1-\frac{1}{r}} \\
&\quad\times\left(\int_0^1 |1-(1+q)t|\,|{}_aD_qf((1-t)a+t(\varphi(b)))|^r\,{}_0\mathrm{d}_qt\right)^{\frac{1}{r}} \\
&\leq \frac{q(\varphi(b)-a)}{1+q}\left(\frac{q(2+q+q^3)}{(1+q)^3}\right)^{1-\frac{1}{r}} \\
&\quad\times\left(\int_0^1 |1-(1+q)t|[(1-t)|{}_aD_qf(a)|^r + t|{}_aD_qf(\varphi(b))|^r]\,{}_0\mathrm{d}_qt\right)^{\frac{1}{r}} \\
&= \frac{q(\varphi(b)-a)}{1+q}\left(\frac{q(2+q+q^3)}{(1+q)^3}\right)^{1-\frac{1}{r}} \\
&\quad\times\left[\frac{q(1+4q+q^2)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(a)|^r + \frac{q(1+3q^2+2q^3)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(\varphi(b))|^r\right]^{\frac{1}{r}}.
\end{aligned}$$

This completes the proof. □

Theorem 12. *Let $f : I = [a, \varphi(b)] \subset \mathbb{R} \to \mathbb{R}$ be a q-differentiable function on I° (the interior of I) with ${}_aD_q$ be continuous and integrable on I where $0 < q < 1$. If $|{}_aD_q f|^r$ is generalized convex function where $p, r > 1$, $\frac{1}{p} + \frac{1}{r} = 1$, then*

$$\left| \frac{qf(a) + f(\varphi(b))}{1+q} - \frac{1}{\varphi(b)-a} \int_a^{\varphi(b)} f(\varphi(x))_a \mathrm{d}_q \varphi(x) \right|$$

$$\leq \frac{q(\varphi(b)-a)}{1+q} \left(\frac{q(2+q+q^3)}{(1+q)^3} \right)^{\frac{1}{p}}$$

$$\times \left[\frac{q(1+4q+q^2)}{(1+q+q^2)(1+q)^3} |{}_aD_q f(a)|^r + \frac{q(1+3q^2+2q^3)}{(1+q+q^2)(1+q)^3} |{}_aD_q f(\varphi(b))|^r \right]^{\frac{1}{r}}.$$

Proof. Since $|{}_aD_q f|^r$ is generalized convex function, so from Lemma 3 and using Holder's inequality, we have

$$\left| \frac{qf(a) + f(\varphi(b))}{1+q} - \frac{1}{\varphi(b)-a} \int_a^{\varphi(b)} f(\varphi(x))_a \mathrm{d}_q \varphi(x) \right|$$

$$= \left| \frac{q(\varphi(b)-a)}{1+q} \int_0^1 (1-(1+q)t)\, {}_aD_q f((1-t)a + t(\varphi(b)))_0 \mathrm{d}_q t \right|$$

$$\leq \left| \frac{q(\varphi(b)-a)}{1+q} \int_0^1 (1-(1+q)t)^{1-\frac{1}{r}} \right.$$

$$\left. (1-(1+q)t)^{\frac{1}{r}}\, {}_aD_q f((1-t)a + t(\varphi(b)))_0 \mathrm{d}_q t \right|$$

$$\leq \frac{q(\varphi(b)-a)}{1+q} \left(\int_0^1 |1-(1+q)t|_0 \mathrm{d}_q t \right)^{\frac{1}{p}}$$

$$\times \left(\int_0^1 |1-(1+q)t|\, |{}_aD_q f((1-t)a + t(\varphi(b)))|^r {}_0 \mathrm{d}_q t \right)^{\frac{1}{r}}$$

$$= \frac{q(\varphi(b)-a)}{1+q} \left(\frac{q(2+q+q^3)}{(1+q)^3} \right)^{\frac{1}{p}}$$

$$\times\left[\frac{q(1+4q+q^2)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(a)|^r\right.$$

$$\left.+\frac{q(1+3q^2+2q^3)}{(1+q+q^2)(1+q)^3}|{}_aD_qf(\varphi(b))|^r\right]^{\frac{1}{r}}.$$

This completes the proof. □

Theorem 13. *Let $f : I = [a, \varphi(b)] \subset \mathbb{R} \to \mathbb{R}$ be a q-differentiable function on I° (the interior of I) with ${}_aD_q$ be continuous and integrable on I where $0 < q < 1$. If $|{}_aD_qf|^r$ is quasi generalized convex function where $p, r > 1$, $\frac{1}{p} + \frac{1}{r} = 1$, then*

$$\left|\frac{qf(a)+f(\varphi(b))}{1+q}-\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))_a\mathrm{d}_q\varphi(x)\right|$$

$$\leq\frac{q(\varphi(b)-a)}{1+q}\left(\frac{q(2+q+q^3)}{(1+q)^3}\right)^{\frac{1}{p}}$$

$$\times\left(\frac{q(2+q+q^3)}{(1+q)^3}\Big[\sup\{|{}_aD_qf(a)|,|{}_aD_qf(\varphi(b))|\}\Big]\right)^{\frac{1}{r}}.$$

Proof. Using Lemma 3, Holder's inequality and the fact that $|{}_aD_qf|^r$ is quasi-generalized convex function, we have

$$\left|\frac{qf(a)+f(\varphi(b))}{1+q}-\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))_a\mathrm{d}_q\varphi(x)\right|$$

$$=\left|\frac{q(\varphi(b)-a)}{1+q}\int_0^1(1-(1+q)t)_aD_qf((1-t)a+t(\varphi(b)))_0\mathrm{d}_qt\right|$$

$$\leq\left|\frac{q(\varphi(b)-a)}{1+q}\int_0^1(1-(1+q)t)^{1-\frac{1}{r}}\right.$$

$$\left.(1-(1+q)t)^{\frac{1}{r}}{}_aD_qf((1-t)a+t(\varphi(b)))_0\mathrm{d}_qt\right|$$

$$\leq\frac{q(\varphi(b)-a)}{1+q}\left(\int_0^1|1-(1+q)t|_0\mathrm{d}_qt\right)^{\frac{1}{p}}$$

$$\times\left(\int_0^1 |1-(1+q)t|\,{}_aD_qf((1-t)a+t(\varphi(b)))|^r\,{}_0\mathrm{d}_qt\right)^{\frac{1}{r}}$$

$$=\frac{q(\varphi(b)-a)}{1+q}\left(\frac{q(2+q+q^3)}{(1+q)^3}\right)^{\frac{1}{p}}$$

$$\times\left(\frac{q(2+q+q^3)}{(1+q)^3}\Big[\sup\{|{}_aD_qf(a)|,|{}_aD_qf(\varphi(b))|\}\Big]\right)^{\frac{1}{r}}.$$

This completes the proof. □

Theorem 14. *Let $f: I=[a,\varphi(b)]\subset\mathbb{R}\to\mathbb{R}$ be a q-differentiable function on I° (the interior of I) with ${}_aD_q$ be continuous and integrable on I where $0<q<1$. If $|{}_aD_qf|^r$ is quasi generalized convex function where $r>1$, then*

$$\left|\frac{qf(a)+f(\varphi(b))}{1+q}-\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\,{}_a\mathrm{d}_q\varphi(x)\right|$$

$$\le\frac{q^2(\varphi(b)-a)(2+q+q^3)}{(1+q)^4}\left(\sup\{|{}_aD_qf(a)|,|{}_aD_qf(\varphi(b))|\}\right)^{\frac{1}{r}}.$$

Proof. Using Lemma 3, power mean inequality and the fact that $|{}_aD_qf|^r$ is quasi-generalized convex function, we have

$$\left|\frac{qf(a)+f(\varphi(b))}{1+q}-\frac{1}{\varphi(b)-a}\int_a^{\varphi(b)} f(\varphi(x))\,{}_a\mathrm{d}_q\varphi(x)\right|$$

$$=\left|\frac{q(\varphi(b)-a)}{1+q}\int_0^1 (1-(1+q)t)\,{}_aD_qf((1-t)a+t(\varphi(b)))\,{}_0\mathrm{d}_qt\right|$$

$$\le\frac{q(\varphi(b)-a)}{1+q}\left(\int_0^1 |1-(1+q)t|\,{}_0\mathrm{d}_qt\right)^{1-\frac{1}{r}}$$

$$\times\left(\int_0^1 |1-(1+q)t|\,|{}_aD_qf((1-t)a+t(\varphi(b)))|^r\,{}_0\mathrm{d}_qt\right)^{\frac{1}{r}}$$

$$=\frac{q^2(\varphi(b)-a)(2+q+q^3)}{(1+q)^4}\left(\sup\{|{}_aD_qf(a)|,|{}_aD_qf(\varphi(b))|\}\right)^{\frac{1}{r}}.$$

This completes the proof. □

Acknowledgements Authors would like to express their gratitude to Prof. Dr. Themistocles M. Rassias for his kind invitation. The authors also would like to thank Dr. S.M. Junaid Zaidi, Rector, COMSATS Institute of Information Technology, Pakistan, for providing excellent research and academic environment. This research is supported by HEC NRPU project No: 20-1966/R&D/11-2553.

References

1. Al-Salam, W.A.: q-Bernoulli numbers and polynomials. Math. Nachr. **17**, 239260 (1959)
2. Al-Salam, W.A.: Operational representations for the Laguerre and other polynomials. Duke Math. J. **31**, 127142 (1964)
3. Al-Salam, W.A.: Saalschützian theorems for basic double series. J. Lond. Math. Soc. **40**, 455458 (1965)
4. Al-Salam, W.A.: q-Appell polynomials. Ann. Mat. Pura Appl. **77**(4), 3145 (1967)
5. Al-Salam, N.A.: On some q-operators with applications. Nederl. Akad. Wetensch. Indag. Math. **51**(1), 113 (1989)
6. Al-Salam, W.A., Carlitz, L.: Some orthogonal q-polynomials. Math. Nachr. **30**, 4761 (1965)
7. Alzer, H.: Sharp bounds for the ratio of q-gamma functions. Math. Nachr. **222**, 514 (2001)
8. Cristescu, G., Găianu, M.: Shape properties of Noor's convex sets. In: Proceedings of the 12th Symposium of Mathematics and Its Applications, Timioara, 5–7 Nov 2009, pp. 91–100
9. Cristescu, G., Gianu, M.: Detecting the non-convex sets with Youness and Noor types convexities. Bul. Stiinț. Univ. Politeh. Timiş. Ser. Mat.-Fiz. **55**(69), 1, 20–27 (2010)
10. Cristescu, G., Lupsa, L.: Non-connected Convexities and Applications. Kluwer Academic Publishers, Dordrecht/Holland (2002)
11. Cristescu, G., Noor, M.A., Awan, M.U.: Bounds of the second degree cumulative frontier gaps of functions with generalized convexity. Carpath. J. Math. **31**(2), 173–180 (2015)
12. Dragomir, S.S., Agarwal, R.P.: Two inequalities for differentiable mappings and applications to special means of real numbers and to trapezoidal formula. Appl. Math. Lett. **11**(5), 91–95 (1998)
13. Dragomir, S.S., Pearce, C.E.M.: Selected Topics on Hermite-Hadamard Inequalities and Applications. Victoria University, Melbourne (2000)
14. Ernst, T.: The history of q-calculus and a new method. U.U.D.M. Report 2000:16, ISSN 1101-3591. Department of Mathematics, Uppsala University (2000)
15. Ernst, T.: A new method for q-calculus. Uppsala Dissertations (2002)
16. Ernst, T.: A method for q-calculus. J. Nonlinear Math. Phys. **10**(4), 487–525 (2003)
17. Ernst, T.: A Comprehensive Treatment of q-Calculus. Springer, Basel/Heidelberg/New York/Dordrecht/London (2014)
18. Gauchman, H.: Integral inequalities in q-calculus. Comput. Math. Appl. **47**, 281–300 (2004)
19. Ion, D.A.: Some estimates on the Hermite-Hadamard inequality through quasi-convex functions. Ann. Univ. Craiova. Math. Comput. Sci. Ser. **34**, 82–87 (2007)
20. Jackson, F.H.: On q-functions and a certain difference operator. Trans. R. Soc. Edinb. **46**, 253281 (1908)
21. Jackson, F.H.: On a q-definite integrals. Q. J. Pure Appl. Math. **41**, 193–203 (1910)
22. Kac, V., Cheung, P.: Quantum Calculus. Springer, New York (2002)
23. Mitrinovic, D.S., Lackovic, I.: Hermite and convexity. Aequationes Math. **28**, 229232 (1985)
24. Niculescu, C.P., Persson, L.-E.: Convex Functions and Their Applications: A Contemporary Approach. CMS Books in Mathematics, vol. 23. Springer, New York (2006)
25. Noor, M.A.: New approximation schemes for general variational inequalities. J. Math. Anal. Appl. **251**, 217229 (2000)
26. Noor, M.A.: On some characterizations of nonconvex functions. Nonlinear Anal. Forum **12**(2), 193–201 (2007)

27. Noor, M.A.: Differentiable non-convex functions and general variational inequalities. Appl. Math. Comput. **199**, 623–630 (2008)
28. Noor, M.A., Awan, M.U., Noor, K.I.: On some inequalities for relative semi-convex functions. J. Inequal. Appl. **2013**, 332 (2013)
29. Noor, M.A., Noor, K.I., Awan, M.U.: Geometrically relative convex functions. Appl. Math. Inf. Sci. **8**(2), 607–616 (2014)
30. Noor, M.A., Noor, K.I., Awan, M.U.: Generalized convexity and integral inequalities. Appl. Math. Inf. Sci. **9**(1), 233–243 (2015)
31. Noor, M.A., Noor, K.I., Awan, M.U.: Some quantum estimates for Hermite-Hadamard inequalities. Appl. Math. Comput. **251**, 675–579 (2015)
32. Noor, M.A., Postolache, M., Noor, K.I., Awan, M.U.: Geometrically nonconvex functions and integral inequalities. Appl. Math. Inf. Sci. **9**(3), 1273–1282 (2015)
33. Ogunmez, H., Ozkan, U.M.: Fractional quantum integral inequalities. J. Inequal. Appl. **2011** (2011). Article ID 787939
34. Ozdemir, M.E.: On Iyengar-type inequalities via quasi-convexity and quasi-concavity. arXiv:1209.2574v1 [math.FA] (2012)
35. Pearce, C.E.M., Pecaric, J.E.: Inequalities for differentiable mappings with application to special means and quadrature formulae. Appl. Math. Lett. **13**, 51–55 (2000)
36. Pecaric, J.E., Prosch, F., Tong, Y.L.: Convex Functions, Partial Orderings, and Statistical Applications. Academic, New York (1992)
37. Rajkovic, P.M., Stankovic, M.S., Marinkovic, S.D.: The Zeros of Polynomials Orthoganal with Respect to q-Integral on Several Intervals in the Complex Plane, pp. 178–188. Softex, Sofia (2004)
38. Taf, S., Brahim, K., Riahi, L.: Some results for Hadamard-type inequalities in quantum calculus. LE Mat. **LXIX**, 243–258 (2014)
39. Tariboon, J., Ntouyas, S.K.: Quantum calculus on finite intervals and applications to impulsive difference equations. Adv. Differ. Equ. **2013**, 282 (2013)
40. Tariboon, J., Ntouyas, S.K.: Quantum integral inequalities on finite intervals. J. Inequal. Appl. **2014**, 121 (2014)
41. Youness, E.A.: E-convex sets, E-convex functions, and Econvex programming. J. Optim. Theory Appl. **102**, 439–450 (1999)

A Digital Signature Scheme Based on Two Hard Problems

Dimitrios Poulakis and Robert Rolland

Abstract In this paper we propose a signature scheme based on two intractable problems, namely the integer factorization problem and the discrete logarithm problem for elliptic curves. It is suitable for applications requiring long-term security and provides smaller signatures than the existing schemes based on the integer factorization and integer discrete logarithm problems.

Keywords: Digital signature • Integer factorization • Elliptic curve discrete logarithm • Supersingular elliptic curves • Pairing • Map to point function • Long-term security

1 Introduction

Many applications of the Information Technology, such as encryption of sensitive medical data or digital signatures for contracts, need long-term cryptographic security. Unfortunately, today's cryptography provides strong tools only for short-term security [5]. Especially, digital signatures do not guarantee the desired long-term security. In order to achieve this goal Maseberg [20] suggested the use of more than one sufficiently independent signature schemes. Thus, if one of them is broken, then it can be replaced by a new secure one. Afterward the document has to be re-signed. Again we have more than one valid signatures of our document. Of course, a drawback of the method is that the document has to be re-signed.

In order to avoid this problem, it may be interesting for applications with long-term, to base the security of cryptographic primitives on two difficult problems, so if any of these problems is broken, the other will still be valid and hence the signature will be protected. We propose in this paper an efficient signature scheme built taking

D. Poulakis (✉)
Department of Mathematics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
e-mail: poulakis@math.auth.gr

R. Rolland
Institut de Mathématiques de Marseille, Université d'Aix-Marseille, Case 907,
13288 Marseille Cedex 9, France
e-mail: robert.rolland@acrypta.fr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_19

into account this constraint. The following signature scheme is based on the integer factorization problem and the discrete logarithm problem on a supersingular elliptic curve. Remark that these two problems have similar resistance to attack, thus they can coexist within the same protocol. The use of a supersingular curve allows us to easily build a pairing that we use to verify the signature.

Several signature schemes combining the intractability of the integer factorization problem and integer discrete logarithm problem were proposed but they have proved either to be enough to solve the one of two problems for breaking the system or to have other security problems [6, 9, 16–19, 22, 27]. An interesting scheme based on the above problems is GPS [8]. Furthermore, some recent such schemes are given in [12, 13, 19, 24, 25, 27].

In Sect. 2 we describe the infrastructure for the implementation of the scheme. Then we present the key generation, the generation of a signature and the verification. In Sect. 3 we show how to build an elliptic curve adapted to the situation and how to define a valuable pairing on it. In Sect. 4 we address the problem of the map to point function and give a practical solution. We deal with the performance of our scheme and compare it with others in Sect. 5. In Sect. 6 we give a complete example that shows that the establishment of such a system can be made in practice. In Sect. 7 we study the security of the scheme. Finally Sect. 8 concludes the paper.

2 The Proposed Signature Scheme

In this section we present our signature scheme.

2.1 Public and Private Key Generation

A user $\mathscr{A}$, who wants to create a public and a private key selects:

1. primes p_1 and p_2 such that the factorization of $n = p_1p_2$ is unfeasible;
2. an elliptic curve E over a finite field $\mathbb{F}_q$, a point $P \in E(\mathbb{F}_q)$ with $\mathrm{ord}(P) = n$ and an efficiently computable pairing e_n such that $e_n(P, P)$ is a primitive nth root of 1;
3. $g \in \{1, \ldots, n-1\}$ with $\gcd(g, n) = 1$, $a \in \{1, \ldots, \phi(n) - 1\}$ and computes $Q = g^a P$;
4. two one-way, collision-free hash functions, $h : \{0,1\}^* \to \{0, \ldots, n-1\}$ and $H : \{0,1\}^* \to < P >$, where $< P >$ is the subgroup of $E(\mathbb{F}_q)$ generated by P.

$\mathscr{A}$ publishes the elliptic curve E, the pairing e_n, and the hash functions h and H. The public key of $\mathscr{A}$ is (P, Q, g, n) and his private key (a, p_1, p_2).

2.2 Signature Generation

The user $\mathscr{A}$ wants to sign a message $m \in \{0,1\}^*$. Then he chooses at random $k, l \in \{1, \ldots, \phi(n) - 1\}$ such that $k + l = a$. Next, he computes

$$s = k + h(m) + n \bmod \phi(n) \quad \text{and} \quad S = g^l H(m).$$

Let $x(S)$ be the x-coordinate of S and b a bit determining S. The signature of m is $(s, x(S), b)$.

2.3 Verification

Suppose that (s, x, b) is the signature of m. The receiver uses b in order to determine y such that $S = (x, y)$ is a point of $E(\mathbb{F}_q)$. He accepts the signature if and only if

$$e_n(g^s P, S) = e_n(g^{h(m)+n} Q, H(m)).$$

Proof of Correctness of Verification. Suppose that the signature (x, s, b) is valid and $S = (x, y)$ is a point of $E(\mathbb{F}_q)$. Then we get

$$e_n(g^s P, S) = e_n(g^{k+h(m)+n} P, g^l H(m)) = e_n(g^{h(m)+n} Q, H(m)).$$

Suppose now we have a couple (s, S), where $s \in \{1, \ldots, \phi(n)\}$ and $S \in < P >$, such that

$$e_n(g^s P, S) = e_n(g^{h(m)+n} Q, H(m)).$$

Since $H(m), S \in < P >$, there are $u, v \in \{0, \ldots, n-1\}$ such that $S = uP$ and $H(m) = vP$. Thus we get

$$e_n((g^s u - g^{h(m)+n+a} v)P, P) = 1.$$

The element $e_n(P, P)$ is a primitive nth root of 1 and so, we obtain

$$uv^{-1} \equiv g^{a+h(m)+n-s} \pmod{n},$$

Putting $l = a + h(m) + n - s \bmod \phi(n)$ and $k = a - l \bmod \phi(n)$, we get

$$s = k + h(m) + n \bmod \phi(n) \quad \text{and} \quad S = g^l H(m).$$

It follows that $(s, x(S), b)$ is the signature of m (where b is a bit determining S).

3 The Elliptic Curve and the Pairing

In this section we show how we can construct an elliptic with the desired properties in order to implement our signature scheme. This task is achieved by the following algorithm:

1. select two large prime numbers p_1 and p_2 such that the factorization of $p_1 - 1$, $p_2 - 1$ are known and the computation of the factorization of $n = p_1 p_2$ is unfeasible;
2. select a random prime number p and compute $m = \mathrm{ord}_n(p)$;
3. find, using the algorithm of [4], a supersingular elliptic curve E over $\mathbb{F}_{p^{2m}}$ with trace $t = 2p^m$;
4. return $\mathbb{F}_{p^{2m}}$ and E.

Since the trace of E is $t = 2p^m$, we get $|E(\mathbb{F}_{p^{2m}})| = (p^m - 1)^2$. On the other hand, we have $m = \mathrm{ord}_n(p)$, whence $n | p^m - 1$, and so n is a divisor of $|E(\mathbb{F}_{p^{2m}})|$. Therefore $E(\mathbb{F}_{p^{2m}})$ contains a subgroup of order n.

By Bróker [4, Theorem 1.1], we obtain, under the assumption that the Generalized Riemman Hypothesis is true, that the time complexity of Step 3 is $\tilde{O}((\log p^{2m})^3)$. Furthermore, since the factorization of $\phi(n) = (p_1 - 1)(p_2 - 1)$ is known, the time needed for the computation of m is $O((\log n)^2 / \log\log n)$ [15, Section 4.4].

For the implementation of our signature scheme we also need a point P with order n and an efficiently computable pairing e_n such that $e_n(P, P)$ is a primitive nth root of 1. The Weil pairing does not fulfill this requirement and also, in many instances, the Tate pairing; the same happens for the eta pairing (the eta and omega pairings can be computed only on the ordinary elliptic curves) [1, 10, 28]. Let ϵ_n be one of the previous pairings on $E[n]$. Following the method introduced by Verheul [23], we use a distortion map ϕ such that the points P and $\phi(P)$ is a generating set for $E[n]$ and we consider the pairing $e_n(P, Q) = \epsilon_n(P, \phi(Q))$. The algorithm of [7, Section 6] provides us a method for the determination of P and ϕ.

Another method for the construction of the elliptic curve E which is quite efficient in practice is given by the following algorithm:

1. draw at random a prime number p_1 of a given size l (for example, l is 1024 bits);
2. draw at random a number p_2 of size l;
3. repeat $p_2 = \mathrm{NextPrime}(p_2)$ until $4p_1p_2 - 1$ is prime;
4. return $p = 4p_1p_2 - 1$.

It is not proved that this algorithm will stop with a large probability. This is an open problem which is for $p_1 = 2$ the Sophie Germain number problem. But in practice we obtain a result p which is a prime of length $2l$.

Since $p \equiv 3 \pmod 4$, the elliptic curve defined over $\mathbb{F}_p$ by the equation

$$y^2 = x^3 + ax,$$

where $-a$ is not a square in $\mathbb{F}_p$, is supersingular with $p + 1 = 4p_1p_2$ points. By Vladut [26, Theorem 2.1], the group $E(\mathbb{F}_p)$ is either cyclic or $E(\mathbb{F}_p) \simeq \mathbb{Z}/2p_1p_2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$. In each case the group $E(\mathbb{F}_p)$ has only one subgroup of order $n = p_1p_2$, and this subgroup is cyclic.

If ϵ_n is one of the Weil, Tate, or eta pairings on $E[n]$, then we use the distortion map $\phi(Q) = \phi(x, y) = (-x, iy)$ with $i^2 = -1$ (cf. [14]) and so, we obtain the following pairing: $e_n(P, Q) = \epsilon_n(P, \phi(Q))$.

4 The Map to Point Function

Let G be the subgroup of order $n = p_1p_2$ of $E(\mathbb{F}_q)$ introduced in the previous section. In order to sign using the discrete logarithm problem on this group, we have to define a hash function into the group G, namely a map to point function. This problem was studied by various authors giving their own method, for example in [3] or [11]. We give here the following solution. Let us denote by $|n| = \lfloor \log_2(n) \rfloor + 1$ the size of n. Let h be a key derivation function, possibly built using a standard hash function. We recall that h maps a message M and a bitlength l to a bit string $h(M, l)$ of length l. Moreover we will suppose that h acts as a good pseudo-random generator. Let Q be a generator of the group G. Let us denote by $(T_i)_{i \geq 0}$ the sequence of bit strings defined by $T_0 = 0$ and for $i \geq 1$

$$T_i = a_u \cdots a_0,$$

where $i = \sum_{j=0}^{u} a_j 2^j$ and $a_u = 1$.

To map the message m to a point $H(m)$ we run the following algorithm:

$i := 0$;
Repeat
$k := h(m||T_i, |n|)$;
$i := i + 1$;
Until $k < n$;
Output $H(M) = k.Q$;

This Las Vegas algorithm has a probability zero to never stop. In practice this algorithm stops quickly, namely as $2^{|n|-1} < n < 2^{|n|}$ then the expected value of the number of iterations is <2. If one can find a collision for H it is easy to find a collision for h.

5 Performance Analysis

In this section we analyze the performance of our scheme. The computation of s requires two additions modulo $\phi(n)$. The computation of S needs a modular exponentiation $g^l \pmod n$ and the computations of $H(m)$ and $g^l H(m)$. Note that

the computation of $g^l \bmod n$ and $k + n \bmod \phi(n)$ can be done off-line. Thus, the signature generation requires only a modular addition and a point multiplication on the elliptic curve. The signature verification needs two modular exponentiations, two points multiplications on the elliptic curves, and two pairing computations. Moreover note that the length of the signature of a message is the double of its length.

The signature generation in the GPS scheme [8] needs only one modular exponentiation and the signature verification two. The signature length is the triple of the message length. The most efficient of the schemes given in [12, 13, 19, 24, 25, 27] requires three modular exponentiations for the signature generation and four modular exponentiations for the signature verification. The signature length of the above schemes is larger than the double of the message length.

Hence we see that the signature length in our scheme is smaller than that in GPS and the other schemes. Moreover, the performance of the proposed algorithm is competitive to the performance of the above schemes.

6 Example

In this section we give an example of our signature scheme. We consider the 1024-bits primes

$$\begin{aligned} p_1 := {} & 6108796057503878981698853611415079226637763631843177587564 \\ & 3192462711995704175406099915839974976783389653390629685931 1 \\ & 2548516341523155127521258304405215057761482861700580373038 9 \\ & 4387740068924296027884510970369084302618887384791344223443 2 \\ & 3659125568423449336215957210074769940424533921400807874383 6 \\ & 7162669180839 \end{aligned}$$

and

$$\begin{aligned} p_2 := {} & 9507945757890361939852894941002382717649136493419364464410 81 \\ & 3770725005780357545382689025181429829602340553197183481715 64 \\ & 5318353480131696755985754343945282697291263271281907117581 93 \\ & 4870883956965030903071113034338701551145996172171056480400 05 \\ & 3445067968984228979774891961106102606656645536560010740680 87 \\ & 13249343. \end{aligned}$$

We take $n = p_1 p_2$. The number $q = 4n - 1$ is a prime. Since $q \equiv 3 \pmod 4$, the elliptic curve E defined by the equation $y^2 = x^3 + x$ over $\mathbb{F}_q$ is supersingular. The point $P = (x(P), y(P))$, where $x(P) = 2^{1500} + 2$ and

$$y(P) = 9262933472009648539425022902353147312856121030374736987117 0$$

532503591346084781038053790347765721405539373837575715741111302632
222520728502603977901582753916707479492439228918725855423715991340
003621514555505206507732534242013847767107764800751435936328543137
789247911179152023276247696951339536945505339588067200491193957998
044975563046555194785086909103272771864842171753848435480722850484
547366650914307823107502201128733622163636510656608071825566283432
994640380462713709910638633429178083083878848700277309884412794341
026781057881112432733889255328105052291841518470922081921433382412
472012678120546125640726148962.

has order n. We take $g = 2$,

$$a = 2^{256} + 2^9 + 1 = 115792089237316195423570985008687907853269984665640564039457584007913129640449$$

and we compute

$$g^a \bmod n = 291246612437704212466554616370488460582482345412043139387071627568366461190658309237330580043030838224854789252968050905018578440545530480131761225347896913705349073419345335895868832920014327349522957752032149784650672578527400186028060209053035728070430079944852013985987562947197675511448867860271390438151997510376157277527652722786834963496843487625119512000324307142997876216044005309541179123902262183075125684914484636806915549910481194533920018176890664864601123368083711476432553316859751469426810204461407620204756483516542976417259702626996120442929825569733396712622105195095244311593920926256171476744 3.$$

Next, we compute $Q = g^a P = (x(Q), y(Q))$, where

$$x(Q) = 492906626963089094011867684016548035835802792163377707597056795455537761970341320418289803336076175870732053896841006011789243411173491601076264818884432777686675649566399360544060115589059409495626348669253033853643920668587107209662122339196308521380419432$$

$$3958767770010377591298098261888264447928963024835312975003285776611156441376633776947815847988008319196552077880554266338219162536485455422641818199238687159366040776610195158709092926451452926125820820564544916736264069574112504476158054648006035374272664210840670688899424879273678267062426009254707550914157923366582588873582336648011173165127581579893233$$

and

$$y(Q) = 9251640006679849414362134638435628671328426925266395037136231007610587593256539123868607426378281972116750233717652921901662256889076587632786360429521239281996051884310217309505235221721760612499163363529422455175409284709873271636908991699714235667300461460401314617119825149525737613057257718590923730935907182295497757283180913934597216850220500675730525413684644075563296631876920873257853188066562736344515028989009339090827154585880138328472819829180452504062174178921959822834145697232804630292818810258440117103130036374232447169484309288773766481841241697043304934210730109599042000468957343998962535886947.$$

Therefore $(P, Q, 2, n)$ and (a, p_1, p_2) are a public key and the corresponding private key for our signature scheme. Moreover, we can use the Tate pairing with the distortion map $\phi(x, y) = (-x, iy)$ with $i^2 = -1$.

7 Security of the Scheme

In this section we shall discuss the security of our system. First, we remark that if an attacker wants to compute the private key (a, p_1, p_2) from the public key, he has to factorize n and to compute the discrete logarithm g^a of Q to the base P and next to calculate the discrete logarithm a of g^a to the base g in the group $\mathbb{Z}_n$. Note that an algorithm which computes the discrete logarithm modulo n implies an algorithm which breaks the Composite Diffie–Hellman key distribution scheme for n and any algorithm which breaks this scheme for a non-negligible proportion of the possible inputs can be used to factorize n [2, 21].

In order to study the security of the scheme we are going to look at the two worst cases:

1. the factorization problem is broken but the elliptic curve discrete logarithm problem is not;
2. the elliptic curve discrete logarithm problem is broken but the factorization problem is not.

In each case we will prove that if an attacker is able to generate a valid signature for any given message m, then it is able to solve, in the first case the elliptic curve discrete logarithm problem and in the second case the factorization problem.

1. Let us suppose that the attacker is able to factorize n. Then he can compute $\phi(n)$. But he is unable to compute a since a is protected by the elliptic curve discrete logarithm problem and by the discrete logarithm problem modulo n, because the only known relation involving a is $Q = g^a P$. So, in order to produce a valid signature of a message m the attacker has only two possibilities: he can arbitrary choose k, and then he can compute s but not S, or choose arbitrary l and he can compute S but not s.
2. Let us suppose now that the attacker is able to solve the elliptic curve discrete logarithm problem. Then he can compute g^a but as the factorization problem is not broken the discrete logarithm problem modulo n is not broken and consequently he cannot compute a (cf. the beginning of this section). Then as in (1) he cannot compute simultaneously s and S.

8 Conclusion

In this paper we defined a signature system based on two difficult arithmetic problems. In the framework chosen, these problems have similar resistance to known attacks. We explained how to implement in practice all the basic functions we need for the establishment and operation of this system. This strategy has an interest in any application that includes a signature to be valid for long. Indeed, it is hoped that if any of the underlying problems is broken, the other will still be valid. In this case, the signature should be regenerated with a new system, without the chain of valid signatures being broken. Finally, the signature length of our scheme is smaller than that of the schemes based on integer factorization and integer discrete logarithm problems, and its performance is competitive to that of these schemes.

References

1. Barreto, P.S.L., Galbraith, S.D., Ó'hÉigeartaigh, C., Scott, M.: Efficient pairing computation on supersingular Abelian varieties. Des. Codes Crypt. **42**, 239–271 (2007)
2. Biham, E., Boneh, D., Reingold, O.: Breaking generalized Diffie-Hellman is no easier than factoring. Inf. Proces. Lett. **70**, 83–87 (1999)

3. Boneh, D., Lynn, B., Shacham, H.: Short signatures from the Weil pairing. Lect. Notes Comput. Sci. **2248**, 514–532 (2001)
4. Bróker, R.: Constructing supersingular elliptic curves. J. Combin. Number Theory **1**(3), 269–273 (2009)
5. Buchmann, J., May, A., Vollmer, U.: Perspectives for cryptographic long term security. Commun. ACM **49**(9), 50–55 (2006)
6. Chen, T.-H., Lee, W.-B., Horng, G.: Remarks on some signature schemes based on factoring and discrete logarithms. Appl. Math. Comput. **169**, 1070–1075 (2005)
7. Galbraith, S.D., Rotger, V.: Easy decision Diffie-Hellman groups. LMS J. Comput. Math. **7**, 201–218 (2004)
8. Girault, M., Poupard, G., Stern, J.: Global Payment System (GPS): un protocole de signature à la volée. In: Proceedings of Trusting Electronic Trade, 7–9 June 1999
9. Harn, L.: Enhancing the security of ElGamal signature scheme. IEE Proc. Comput. Digital **142**(5), 376 (1995)
10. Hess, F., Smart, N.P., Vercauteren, F.: The Eta pairing revisited. IEEE Trans. Inf. Theory **52**(10), 4595–4602 (2006)
11. Icart, T.: How to Hash into elliptic curves. In: CRYPTO 2009. Lecture Notes in Computer Science, vol. 5677, pp. 303–316. Springer, New York (2009)
12. Ismail, E.S., Tahat, N.M.F., Ahmad, R.R.: A new digital signature scheme based on factoring and discrete logarithms. J. Math. Stat. **4**(4), 22–225 (2008)
13. Ismail, E.S., Tahat, N.M.F.: A new signature scheme based on multiple hard number theoretic problems. ISRN Commun. Netw. **2011**, 3 pp. (2011). Article ID 231649. http://dx.doi.org/10.5402/2011/231649
14. Joux, A.: The weil and tate pairings as building blocks for public key cryptosystems (Survey). In: ANTS 2002. Lecture Notes in Computer Science, vol. 2369, pp. 20–32. Springer, Berlin (2001)
15. Karagiorgos, G., Poulakis, D.: Efficient algorithms for the basis of finite Abelian groups. Discret. Math. Algoritm. Appl. **3**(4), 537–552 (2011)
16. Lee, N.Y.: Security of Shao's signature schemes based on factoring and discrete logarithms. IEE Proc. Comput. Digital Tech. **146**(2), 119–121 (1999)
17. Lee, N.Y., Hwang, T.: The security of He and Kiesler's signature scheme. IEE Proc. Comput. Digital Tech. **142**(5), 370–372 (1995)
18. Li, J., Xiao, G.: Remarks on new signature scheme based on two hard problems. Electron. Lett. **34**(25), 2401 (1998)
19. Madhur, K., Yadav, J.S., Vijay, A.: Modified ElGamal over RSA Digital Signature Algorithm (MERDSA). Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(8), 289–293 (2012)
20. Maseberg, J.S.: Fail-safe konzept fur public-key infrastrukturen. Thesis, Technische Universitat Darmstadt (2002)
21. McCurley, K.S.: A key distibution system equivalent to factoring. J. Cryptol. **1**, 95–105 (1988)
22. Shao, Z.: Security of a new digital signature scheme based on factoring and discrete logarithms. Int. J. Comput. Math. **82**(10), 1215–1219 (2005)
23. Verheul, E.: Evidence that XTR is more secure than supersingular elliptic curve cryptosystems. In: Advances in Cryptology-Eurocrypt '01. Lecture Notes in Computer Science, vol. 2045, pp. 195–210. Springer, New York (2001)
24. Verma, S., Sharma, B.K.: A new digital signature scheme based on two hard problems. Int. J. Pure Appl. Sci. Technol. **5**(2), 55–59 (2011)
25. Vishnoi, S., Shrivastava, V.: A new digital signature algorithm based on factorization and discrete logarithm problem. Int. J. Comput. Trends Tech. **3**(4), 653–657 (2012)
26. Vladut, S.: Cyclicity statistics for elliptic curves over finite fields. Finite Fields Appl. **5**(1), 13–25 (1999)
27. Wei, S.: A new digital signature scheme based on factoring and discrete logarithms. In: Progress on Cryptography. Kluwer International Series in Engineering and Computer Science, vol. 769, pp. 107–111. Kluwer Academic Publishers, Boston (2004)
28. Zhao, C.-A., Xie, D., Zhang, F., Zhang, J., Chen, B.-L.: Computing bilinear pairing on elliptic curves with automorphisms. Des. Codes Crypt. **58**, 35–44 (2011)

Randomness in Cryptography

(Invited Talk)

Robert Rolland

Abstract This talk is a short overview [This overview is partially based on the paper (Ballet and Rolland, Cryptogr. Commun. **3**(4), 189–206, 2011)] on the use of randomness in cryptography. Firstly we give some indications on building and using the randomness and pseudo randomness in a cryptographic context. In the second step, we study more formally the notion of pseudo-random sequence. We introduce the notion of distinguisher and prediction algorithms and we compare these two notions.

Keywords: Cryptography • Distinguisher • Prediction • Pseudo-random generator • Ramdomness • Seed • Yao theorem

1 Introduction

Randomness is among the main tools in cryptography. Many cryptographic primitives or protocols include a random part. It is the case for stream ciphers, construction of keys, key exchange in the Ephemeral Unified Model, construction of an initial value, etc. Usually, randomness is simulated by a pseudo-random generator or occasionally, when we only need a small number of isolated values, by a built-in physical generator.

We must distinguish between two typical mode of use. On the one hand, the random draw of a number of medium size, for example a secret key, on the other hand the random draw of a very large sequence of bits as, for example, in the case of a stream cipher. In the first case we can use a primitive of general interest, such as a hash function or a block cipher. The second case is more difficult. Indeed, a typical application leading to this situation is the stream cipher for which the main interest is speed. Then we must build a system faster than usual block ciphers without compromising security, which is difficult because we must treat each bit upon arrival. Then we cannot, in order to increase the security, iterate a round

R. Rolland (✉)
Institut de Mathématiques de Marseille, Université d'Aix-Marseille, Case 907,
13288 Marseille Cedex 9, France
e-mail: robert.rolland@acrypta.fr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_20

function as in the case of a block cipher. For this case we refer to the following European Project eSTREAM:

http://www.ecrypt.eu.org/stream/

In the first part of this talk we present a practical study of the concept of randomness in cryptography. This includes a practical way to construct a seed and a pseudo-random generator for medium size data in a Linux environment.

The second part is theoretical. In that part we precisely define the notion of pseudo-random generator. Then we define the notion of distinguisher and the notion of prediction. Yao's theorem [7] gives an equivalence between the indistinguishability of a pseudo-random generator and the unpredictability of the next bit from an asymptotic point of view. In this paper we present modified versions of Yao's theorem (see [1]) which can be of interest for the study of practical cryptographic primitives. In particular we consider non-asymptotic versions. We study the case of one pseudo-random generator, then the case of a family of pseudo-random generators with the same fixed length, and finally we consider the asymptotic case. We compute in each case the cost of the reduction (in the sense of complexity theory) between the two algorithms.

Some books on pseudo-random generators as well as probabilistic algorithms and proofs are given in [2–4].

2 Pseudo-Random Number Generator

2.1 How to Generate a Seed

Generally, operating systems provide a physical source of randomness based on different component behaviors: keyboard, mouse, clock, processes, etc. As this source of randomness does not contain a large amount of bits, it is only used to generate an occasional number as a seed. For example, under a linux system the device /dev/random plays this role. Then to get a (printable) seed under linux we can give the following instructions:

```
head -c 128 /dev/random |
openssl dgst -sha256 -binary |
openssl enc -base64 > seed.b64
```

The first line extracts 128 bits of the /dev/random device and sends them in the pipe. The second line reads the pipe, hashes these 128 bits, and returns the 256 bits of the hash in the pipe (binary format). The third line reads the pipe and transforms in base64 format the 256 bits in 44 symbols included the symbol "=" at the end in order to have a printable result.

Note that the linux device /dev/random can be improved by using the *haveged* deamon based on the Havege algorithm (see [5]).

2.2 Example: How to Construct a Pseudo-Random Generator

Let H be a hash function (for example sha256). From a seed s (at least 128 bits) we construct the following pseudo-random sequence S_n of bits :

$$s_0 = h(s), s_1 = h(s||s_0), \cdots, s_n = h(s||s_{n-1}), \cdots$$
$$S_n = s_1||s_2||\cdots||s_n$$

It is also possible to use AES (as for counter mode) to construct a pseudo-random generator. Let us remark that this type of pseudo-random generator is also called a key derivation function as its main interest is building cryptographic keys. We refer to the standard ISO 18033-2 to see how to implement in practice such a system (see [6]).

However, as remarked before, for stream ciphers it is mandatory to use a specific construction if a faster encryption device than AES is hoped (beware the attacks against stream ciphers).

3 Theoretical Point of View

3.1 Definition of a Pseudo-Random Number Generator

Definition 1. Let k and n be two integers such that $n > k$. A pseudo-random generator (prng) is a function f from a subset $\mathbf{U}$ of $\{0,1\}^k$ into $\{0,1\}^n$:

$$f : \mathbf{U} \subset \{0,1\}^k \to \{0,1\}^n,$$

mapping a seed $X_0 \in \mathbf{U}$ to a pseudo-random finite sequence

$$f(X_0) = (x_1, x_2, \ldots, x_n).$$

We shall denote by (f, U, k, n) this prng.

A typical case is when u is a bijection from $\mathbf{U}$ onto itself, X_i is a secret internal state built recursively from X_0 by $X_i = u(X_{i-1})$ and the bit x_i is extracted from X_i by a function v: $x_i = v(X_i)$:

$$f(X_0) = \left(v \circ u(X_0), v \circ u^2(X_0), \ldots, v \circ u^n(X_0)\right).$$

3.2 Distinguisher

Roughly speaking a distinguisher is a probabilistic algorithm able to distinguish a true random sequence of a pseudorandom one. Let us specify this informal definition. Let (f, U, k, n) be a prng. Let $\mathscr{A}$ be a probabilistic algorithm that applies to a binary vector $Y = (Y_1, \ldots, Y_n)$ and which outputs one bit.

True randomness experiment:

Let us denote by $p_{f,0}$ the probability of the following event:
we draw at random an element Y of $\{0,1\}^n$ and $\mathscr{A}(Y)$ is 1.

Pseudo randomness experiment:

Let us denote by $p_{f,n}$ the probability of the following event:
we draw at random an element of U, compute $Y = f(U)$ and $\mathscr{A}(Y)$ is 1.

Now we can define the Advantage of a distinguisher:

The Advantage of $\mathscr{A}$ to distinguish a true random sequence of n bits from the pseudo-random sequence given by f is

$$\mathrm{Adv}_f^{\mathrm{dist}}(\mathscr{A}) = |p_{f,0} - p_{f,n}|.$$

Then, we define a (T, ϵ)-distinguisher:

Definition 2. Let f be a pseudo-random generator. Let T and ϵ be positive real numbers. A (T, ϵ)-distinguisher for f is a probabilistic algorithm $\mathscr{A}$ such that

1. the maximal running time of $\mathscr{A}$ is $\leq T$,
2. the input of $\mathscr{A}$ is an element of $\{0,1\}^n$,
3. the output of $\mathscr{A}$ is a bit b,
4. the algorithm $\mathscr{A}$ can distinguish the pseudo-random generator from the uniform distribution, namely

$$\mathrm{Adv}_f^{\mathrm{dist}}(\mathscr{A}) > \epsilon.$$

3.3 Prediction

Let f a pseudo-random generator whose image is in $\{0,1\}^l$. A prediction algorithm is a probabilistic algorithm which has the ability to predict the next bit of a finite sequence. Let us specify this informal definition. Let $1 \leq s < l$. The following random experiment involves a probabilistic algorithm $\mathscr{B}$ having for input a sequence of s bits and for output a bit.

Experiment B:

$\mathbf{Expt}_{\mathbf{f,s}}^{\mathbf{pred}}(\mathscr{B})$
$X_0 \leftarrow \mathbf{U} \subseteq \{0,1\}^k$
$X \leftarrow f(X_0)$ (notation : $X = (x_1, \cdots, x_l)$)
$Y \leftarrow (x_1, x_2, \cdots, x_s)$
$b \leftarrow \mathscr{B}(Y)$
if $b = x_{s+1}$
then return 1
else return 0
fi
End.

Let $r_{f,s}$ be the probability that the experiment $\mathbf{Expt}_{\mathbf{f,s}}^{\mathbf{pred}}(\mathscr{B})$ returns 1.

Definition 3. The advantage of the algorithm $\mathscr{B}$ to predict the bit of index $(s+1)$ computed by f is:

$$\mathrm{Adv}_{f,s}^{\mathrm{pred}}(\mathscr{B}) = \left| r_{f,s} - \frac{1}{2} \right|.$$

Definition 4. Let f be a pseudo-random generator. Let T and ϵ be positive real numbers and s be an integer such that $1 \leq s < l$. A (T, s, ϵ)-prediction algorithm $\mathscr{B}$ is a probabilistic algorithm such that:

1. the maximal running time of $\mathscr{B}$ is $\leq T$,
2. the input of $\mathscr{B}$ is an element of $\{0,1\}^s$,
3. the output of $\mathscr{B}$ is a bit,
4. the algorithm $\mathscr{B}$ can predict the next bit, namely

$$\mathrm{Adv}_{f,s}^{\mathrm{pred}}(\mathscr{B}) > \epsilon.$$

We define now the notion of (T, s, ϵ)-unpredictable pseudo-random generator.

Definition 5. Let f be a pseudo-random generator and s an integer such that $1 \leq s < l$. The generator f is (T, s, ϵ)-unpredictable, if there does not exist any (T, s, ϵ)-prediction algorithm.

3.4 A Static Version of Yao's Theorem

We give here two theorems that summarize in a static context the relations between prediction algorithms and distinguishers. These results are proved in [1].

Theorem 1. *We consider the following pseudo-random generator:*

$$f : \mathbf{U} \subset \{0,1\}^k \rightarrow \{0,1\}^l.$$

If we have a

$$(T, s, \epsilon)\text{-}\textit{prediction algorithm}$$

for f, we can build a

$$(T + c, \epsilon)\text{-}\textit{distinguisher},$$

where c is the constant time needed to compare two bits.

Theorem 2. *Let f be a pseudo-random generator:*

$$f : \mathbf{U} \subset \{0,1\}^k \rightarrow \{0,1\}^l.$$

Let us suppose that there is a (T, ϵ)-distinguisher for f, then there exist a s such that $1 \leq s \leq l$ and a $(T + (c_1 l + c_2), s, \epsilon/l)$-prediction algorithm where c_1 is the constant time needed to draw one bit at random, and c_2 is the constant time needed to test the value of a bit and then return a bit depending upon the result of the test.

3.5 Family of Pseudo-Random Generators

In a realistic situation we must, in the random experiment which defines the attacker's advantage, draw at random the function f from a family Γ according to a probability law δ. For example, let us define the family of Blum, Blum, Shub generators. Let us choose two Blum prime p and q, namely p and q are such that $p \equiv 3 \bmod 4$ and $q \equiv 3 \bmod 4$. From a secret seed s_0 we construct a sequence such that $s_k = s_{k-1}^2 \bmod(pq)$. Note that s_k is an internal state which must remain secret.

Then the bit x_k of the pseudo random sequence is the last bit of s_k. If we fix a size for the product pq, we can consider the family of pseudo-random generators constructed with all the couple (p, q) of distinct primes such that pq has the required size.

If we slightly modify the definitions according to this new context, we obtain similar results.

Theorem 3. *Let Γ be a family of pseudo-random generators having the same size where each $f \in \Gamma$ is a function*

$$f : \mathbf{U}_f \subset \{0,1\}^k \to \{0,1\}^l.$$

If we have a

$$(T, s, \epsilon)\text{-prediction algorithm for } \Gamma$$

we can build a

$$(T + c, \epsilon)\text{-distinguisher for } \Gamma,$$

where c is the constant time needed to compare two bits.

Theorem 4. *Let Γ be a family of pseudo-random generators having the same size where each $f \in \Gamma$ is a function*

$$f : \mathbf{U}_f \subset \{0,1\}^k \to \{0,1\}^l.$$

If we have a

$$(T, \epsilon)\text{-distinguisher algorithm for } \Gamma$$

we can build a

$$(T + c_1 l + c_2, s, \epsilon/l)\text{-prediction algorithm for } \Gamma$$

for some value of s ($1 \leq s < l$), where c_1 is the constant time needed to draw one bit at random, and c_2 is the constant time needed to test the value of a bit and then to return a bit.

3.6 Asymptotic Behavior

As a consequence of the previous results for fixed k and l, we can deduce results on the asymptotic theory of the pseudo-random generators, namely k growing to infinity and $l = l(k) > k$ a polynomial function of k.

Let k be a positive integer (the security parameter) and $l(k)$ a polynomial function of k such that $l(k) > k$. For any k we have a set Γ_k of deterministic functions such that

1. if $f \in \Gamma_k$, then f is a function from a subset $\mathbf{U}_f$ of $\{0,1\}^k$ into $\{0,1\}^{l(k)}$;
2. there exists a polynomial function $t(k)$ such that for any k, any $f \in \Gamma_k$ and any $X \in \mathbf{U}_f$ the computation time of $f(X)$ is upper-bounded by $t(k)$;
3. for any k we provide a probability δ_k on the set Γ_k.

The asymptotic notions of indistinguishability and unpredictability are derived from the previous definitions. We define now a distinguisher $\mathscr{A}$ to be a probabilistic polynomial algorithm having for inputs the security parameter k, a function $f \in \Gamma_k$ and a vector $Y \in \{0,1\}^{l(k)}$, and which outputs a bit. Let k be an integer, we will denote by $\mathscr{A}_k$ the probabilistic algorithm obtained from $\mathscr{A}$ by fixing the first entry to the value k.

Definition 6. The family $\Gamma = (\Gamma_k)_{k>0}$ of sets of pseudo-random generators is said asymptotically secure if for any polynomial $S(k)$, any positive integer u and any distinguisher $\mathscr{A}$ with running time $\leq S(k)$, the advantage of the algorithm $\mathscr{A}_k$ is a negligible function of $\frac{1}{k^u}$, namely

$$\lim_{k \to +\infty} k^u \operatorname{Adv}^{\text{dist}}_{\Gamma_k}(\mathscr{A}_k) = 0.$$

Let $s = (s_k)_{k \geq 1}$ be an increasing sequence of positive integers such that $1 \leq s_k < l(k)$. We define now an s-prediction algorithm to be a probabilistic polynomial algorithm $\mathscr{B}$ having for inputs the security parameter k, a function $f \in \Gamma_k$ and a vector $Z \in \{0,1\}^{s_k}$, and which outputs a bit. Let k be an integer, we will denote by $\mathscr{B}_k$ the probabilistic algorithm obtained from $\mathscr{B}$ by fixing the first entry to the value k.

Definition 7. The family $\Gamma = (\Gamma_k)_{k>0}$ of sets of pseudo-random generators is said asymptotically unpredictable if for any polynomial $S(k)$, any sequence s and any s-prediction algorithm $\mathscr{B}$ with running time $\leq S(k)$, the advantage of the s_k-prediction algorithm $\mathscr{B}_k$ is a negligible function of $\frac{1}{k^u}$, namely

$$\lim_{k \to +\infty} k^u \operatorname{Adv}^{\text{pred}}_{\Gamma_k, s_k}(\mathscr{A}_k) = 0.$$

We can now state the asymptotic version of the Yao's theorem:

Theorem 5. *Let $l(k)$ be a polynomial function of one integer variable k such that $l(k) > k$. Let $\Gamma = (\Gamma_k)_{k>0}$ a family of sets, where any set Γ_k is a probability set of random generators mapping a subset of $\{0,1\}^k$ into $\{0,1\}^{l(k)}$ (more precisely, each $f \in \Gamma_k$ has its own definition subset $\mathbf{U}_f \subseteq \{0,1\}^k$). The family Γ is asymptotically secure if and only if it is asymptotically unpredictable.*

Proof. See [1].

References

1. Ballet, S., Rolland, R.: A note on yao's theorem about pseudo-random generators. Cryptogr. Commun. **3**(4), 189–206 (2011)
2. Barthélemy, P., Rolland, R., Véron, P.: Cryptographie: principes et mises en œuvre - 2e édition revue et augmentée. Lavoisier (2012)
3. Goldreich, O.: Modern Cryptography, Probabilistic Proofs and Pseudo-Randomness. Number 17 in Algorithms and Combinatorics. Springer, Berlin (1999)
4. Luby, M.: Pseudorandomness and Cryptographic Applications. Princeton University Press, Princeton (1996)
5. Seznec, A., Sendrier, N.: Havege: a user-level software heuristic for generating empirically strong random numbers. ACM Trans. Model. Comput. Simul. **13**(4), 334–346 (2003)
6. Shoup, V.: Iso/iec fcd, 18033-2–information technology– security techniques–encryption algorithms–part 2: asymmetric ciphers. Technical Report, International Organization for Standardization (2004)
7. Yao, A.C.: Theory and applications of trapdoor functions. In: Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science, pp. 80–91. IEEE Computer Society, Washington, DC, USA (1982)

Current Challenges for IT Security with Focus on Biometry

Benjamin Tams, Michael Th. Rassias, and Preda Mihăilescu

Abstract In this paper we give a survey of biometrical applications in security context. We start with a brief overview of the different biometric modalities which are most frequently used and compare their security contribution with classical cryptographic primitives. We then consider the case of fingerprints when used as password surrogates. We discuss the main security concerns of biometry in more detail on this practical example and make a point that the *false accept error probability* should be considered as the de facto measure of security.

Keywords: Cryptography • IT security • Biometry • Fuzzy commitment scheme • Fuzzy vault Scheme • Slow-down function • Biometric hash

1 Introduction

Confidential communication is a request with an old tradition, mostly with military applications. Two parties wish to communicate in such a way that no unauthorized (by them) third party may have *a slight chance* to reveal the content of the communication. Some side-requirements in such a setting are

- The request for *secure authentication*.
- The request for provable *signatures*, or, more generally, insurance of the impossibility to repudiate the origin of a message.

A common answer to these requirements was provided by cryptography. A logical art for dealing with this problem is known from early Antiquity; until recent times. It was commonly accepted that for confidentiality, one needed some *secret keys* that were shared only by the authorized parties. The algorithm by which these secret keys were used should also preferably contain some private tricks to make it

B. Tams • P. Mihăilescu
Mathematisches Institut der Universität Göttingen, Bunsenstrasse 3-5, 37073 Göttingen, Germany
e-mail: btams@math.uni-goettingen.de; preda@uni-math.gwdg.de

M.Th. Rassias (✉)
Department of Mathematics, ETH-Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
e-mail: michail.rassias@math.ethz.ch

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_21

more reliable. Since the ideas for encryption were based on a common collection of techniques, one could not require completely private algorithms; but it was assumed that by adding some special tricks and complexity, an algorithm would become more resistant to attacks. The general attitude in this respect was completely reversed in modern cryptography, and since decades we prefer to use publically known algorithms, that have resisted the scrutiny of a world-wide community of specialists, thus proving their reliability. It is believed that additional private tricks can often lead to providing a false impression of security, which may lead to errors and attacks.

Transposing the alphabet of a spoken language into a sequence of numeric codes is always useful for discussing cryptographic ideas. Suppose thus that the Latin alphabet $a, b, \ldots, z$ is encoded in ascending order by the numbers $0, 1, \ldots, 25$. The idea of permuting the letters cyclically by a constant σ was purportedly used by Caesar in the Gallic wars—hence the name of *Caesar* code. For instance, for $\sigma = 4$, the word

ATHENS

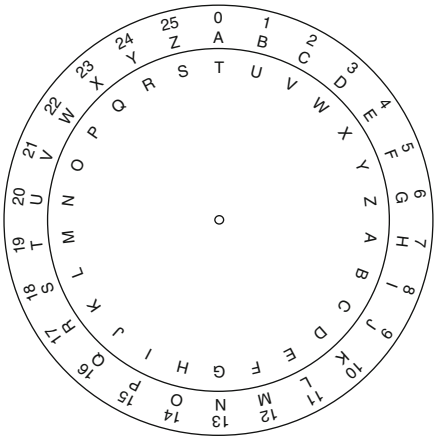
becomes

EYLIRW.

For decryption, use $\sigma = 25 - 4 = 21$. The main idea of this approach to confidentiality, which is based on the sharing of a secret key—hence the term *secret key algorithms*—is some kind of key triggered permutation. One may permute the very alphabet in which the message is written—the seminal idea used in the so-called Caesar Code. More sophisticated variants will first translate the written text by means of a code, and then use chains of key-driven permutations for encryption. This approach is applied even by modern secret key algorithms such as the internationally used *data encryption algorithm* DEA. While *confidentiality* is obtained by protecting the secret key, the authenticity of the peer is only deduced from the fact that he possess the secret key which should not be obtained by any other person. And there is no possible means to bind messages to the identity of their issuer in secret key mode—by the very fact that at least two peers must share the same key for communication, it becomes obvious that there is no individual information available for identifying the author of a message (Fig. 1).

Therefore, secret key cryptography may still be used successfully even in modern computer times, for protecting the message's content. At the advent of computer networks, the alternative of *public key cryptography* was invented independently by two groups of young American researchers involved in the incipient ARPA net of the 1970s and by an engineer working for the MI5, who was allowed to disclose his discovery in the year 2000. The common idea is to split the key of a peer, say A, for Alice, in two parts, a private part $S(A)$ and a public part $P(A) \subsetneq S(A)$, which is available to the world. Encryption does work both ways like in the secret case. Only, for writing to A, peers B will use the public key, creating messages that

Fig. 1 Caesar code



can only be decrypted using the private key $S(A)$. In addition, Alice will now be able to authenticate herself, by encrypting *any public message*—for instance "hello world"—with her private key. Since nobody else should be able to find the private key of Alice, upon description with the public key, Bob or anyone else, can be convinced of the fact that it was indeed Alice that generated the encryption—in this situation, the encryption with the private key stands for something like a handwritten signature; it is therefore also called *digital signature*. The same procedure is used to obtain non-repudiation of the origin of a message. These facilities are essential for private secure electronic communication.

Therefore public key cryptography has found its exponential spread at the time of the opening of the world wide web to the large public, in the mid-1990s. While public key cryptography is found in more and more applications, several new important problems arise

1. The *reliability* of public keys obtained in the public domain.
2. The multitude of secret key protections required.
3. The very reliability of hardware used in transactions that require personal authentication.
4. User-friendliness.

The purpose of this paper is to discuss these challenges of modern IT security. We shall focus hereby on a new facility which receives increasing attention and use in this context, namely the use of biometric traits for identifying humans. After explaining the context in which new challenges to information security arise and discussing the possibilities and limitations of cryptography, we give a brief introduction to the classical aspects of biometry, related to image identification. After that we approach the core subject of this survey, which is the application of biometry to secure applications, giving an overview of attempts that have been done in this direction, their limitations, and discussing some new algorithms that circumvent problems and vulnerabilities found with some state-of-the-art algorithms.

2 Trends and Challenges in Information Security

In the last three decades, *cryptology* has become a major field of research, together with its Janus—faced duality: *cryptography*, for the design of algorithms and protection principles and *cryptanalysis* for investigation of possible attacks against these algorithms. The primitive algorithms are divided into:

A. Secret Key Algorithms
B. Public Key Algorithms
C. One way functions, hashes and
D. Key management.

We have already discussed briefly the first two. One way functions or hashes have the paradoxical property of being highly non-injective maps, since they map the realm of all possible messages to fixed length blocks, of, say, 192 bits. Such a hash would be a map $\chi : \mathbb{N} \to \mathbb{Z}/(192 \cdot \mathbb{Z})$. However, the size of the image set is large enough to ensure that it is not computationally feasible to find even one collision, i.e. $x \neq y$ with $\chi(x) = \chi(y)$. Little to say about a match, which would require to find, for a given hash of an unknown value, say $h = \chi(x)$ a value $y \in \mathbb{N}$ with $\chi(y) = h$. The collision problem is easier, since it only requires two random hashes to match; in the second case, one hash value is already fixed. One way functions must fulfill certain properties related to the conditions discussed. If they do, they are used both for saving passwords in a protected way, without the use of encryption: just substitute a password by its hash value, so that the stored data will reveal no information about the initial password. Hash words are also used in connection with *digital signatures*: Messages to bind to a digital signature are sometimes very large, so one prefers to replace them by their unique hash value and place a digital signature on this hash value.

Key management is less of a cryptographic primitive and more of a set of requirements for the privacy and reliability of keys and passwords used in secure communication. Key management draws on standards of key-authentication, as well as hardware token such as chip cards or other devices, carrying sensitive keys, etc. It is the task of key management to provide not only for secure key storage—either on encrypted memory or chip cards or similar devices—but also for *trust diffusion*. By this we mean that two peers, say Alice and Bob, who start communication by exchanging public keys, should be provided with means to trust that the received public key does indeed belong to either Alice or Bob. Avoiding attacks by *masquerading* false keys is thus an important task of key management. The provisions for this task are a mixture of cryptography and protocol administration.

It is probably the most important achievement of modern cryptography, that the problems of secure information exchange have been reduced to primitives, endowed with well-defined properties, and security is asserted on the base of such properties which can be verified by the cryptologist in the whole world. Hence, the possibility of attacks to a cryptographically secured environment can be also grouped in types

of attacks based on well-defined *attack scenarios*. It is the presence of these attack scenarios which helps establish the trust into cryptographic solutions, which end up being standardized and used world-wide. A typical, very important *standard* in this context is the *TLS/SSL standard*, which is the cryptographic standard of the world wide web and provides secure communication facilities based on variable tool-kit primitives.

One may conclude that the first decades of public key cryptography provided a reliable system of well-scrutinized primitives for addressing each of the problems A–D. The algorithms for public key encryption, hashes, and secret key algorithms as well as the protocols for key management of the last decades are resistant to direct attacks, beyond reasonable doubt.[1]

At the present day, cryptology offers protocols and primitives that are

C1. **Reliable**: They are well researched and secure within any reasonable doubt.
C2. Providing **scalable security** in the sense that it is possible in any of the primitives, to adapt to increased performance of computers, by modifying the length of keys in such a way that the expected time necessary to perform well-defined attacks on a given primitive stays unchanged.
C3. **deterministic** in the sense that on the same input they will always produce the same output. The notion of security if based on the provision that an attack on a primitive should require computation time which stretches beyond hundreds of years, under the most favorable circumstances and using the best algorithms to date. *Even the lowest accepted level of security is beyond doubt*, and the primitives are rejected as soon as theoretical advances show any vulnerability allowing for attacks which can be performed in less than decades or even centuries.

2.1 Recent Evolution

After these achievements were completed in the 1990s, the challenges of security moved to more volatile topics. The most important ones being:

H1. The definition of trust: in an open environment, who should security protect against whom? Can one trust the user more or the vendor providing some token or hardware, that requires secure identification, which may be stolen?
H2. Viruses and denial of service attacks: both are attacks against an operation system that can either spread over the whole internet or focus on certain target intranets, leading to a blockage of their functionality by overload.

[1] We should not mistake *beyond reasonable doubt* with *provable certainty*. There is no mathematical proof for the lack of efficient attacks to the state-of-the-art primitives, and even if such one would be provided, it would always be connected to a fixed context of application. But new attacks can be invented, which were not thought of. Confidence relays on the intensive long time research in the public academic domain, spent on the related cryptologic questions.

H3. Hacker intrusions of intranets. These are often performed with the purpose of commercial espionage and use any kind of vulnerabilities of operating system, security implementations of even individual authentic users of the intranet.

Developing countermeasures to these very real and corrosive kinds of attacks is an endeavor that requires all the apparatus of cryptology but reaches well beyond: it is the modern task of security engineering.

One may thus observe that cryptology has offered its best and became now part of the more complex task of IT security engineering. Paradoxically, the development and spread of secure applications lead at the opposite end of complexity to new challenges. Since applications are mostly independent and coming from various vendors, the typical user of a large intranet becomes soon confronted with the requirement to secure his identification with respect to a multitude of software, each requiring *safe passwords* from him. This challenges human memory and it mostly happens that users choose to bypass security prescriptions for passwords, by either writing them down or using multiple passwords. This leads to user-driven vulnerabilities.

2.2 The Advent of Biometry

In this context, biometry entered the scene by raising an expectation which is best reflected in the paradigm *you are what you are* as opposed to *you are what you know or what you have*. Indeed, in a cryptographical frame, the user is authenticated either by knowledge of some secret, such as the password of some key, or by means of a token which carries this secret information for him. Assuming she has control on this access modalities, cryptology guarantees secure use. However, the control is relativized by the reasons presented above. Therefore biometry suggests to identify a person physically, by some unique traits that distinguish him uniquely. This can be fingerprints or iris, face or writing mechanics, vein geometry or voice—a multitude of physical and behavioral traits have been proposed and investigated in order to uniquely identify a person. The wish becomes one to remove the responsibility for identification information from the user and defer it to technology. The user presents his physical appearance and trusts the system that it may well identify him and not allow intrusions or any other kind of abuse of information related to him. The approach was motivated by the success achieved in image processing during the previous decades, which made the identification by means such as fingerprinting, iris, or face recognition quite reliable. However, the advent of biometry and its increasing **actual use** in security contexts raises a series of important questions.

B1. Unlike cryptographic authentication, the biometric one is not deterministic. It is based on comparing real time acquired data—such as images of fingers, iris, or a face, against *templates* stored in a database. It is certainly likely that the new data best matches the one of the template of the individual stored in the database; this is however only true within a certain stochastic measure of

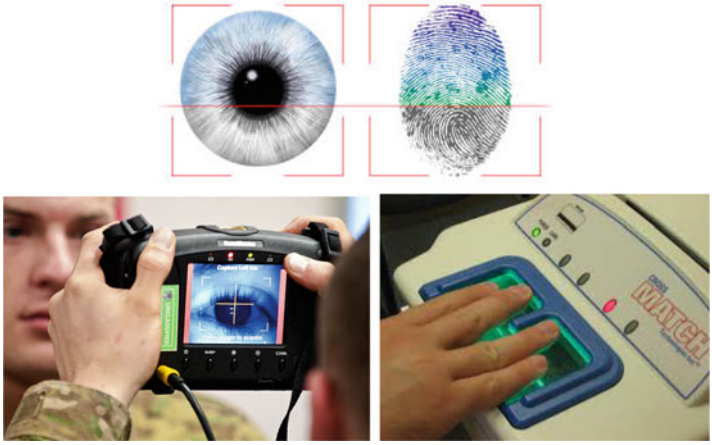


Fig. 2 Iris and fingerprint recognition

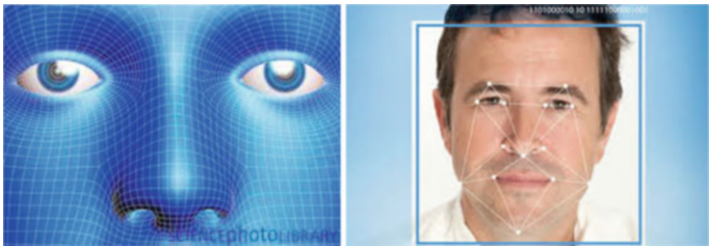


Fig. 3 Face recognition

doubt. The actual deterministic certainty is replaced by an error distribution of false acceptances and false rejects. By deciding acceptance scores, the system can adjust false accepts against false rejects—it will not be able though, to remove any probability of error. This is the **stochastic nature** of biometric identification (Figs. 2 and 3).

B2. Unlike passwords, which can be replaced when compromised, biometric traits cannot be changed. As a consequence, the world wide system using one kind of biometry cannot be reliably factored into more secure areas, by use of advanced and expensive technology: a template acquired in a weak system can be used for impersonating a user in any other system.

B3. Most important, the presence of a non-vanishing probability of false acceptance becomes the de facto measure of **active information entropy** present in some type of biometric recognition. We describe in this paper some attacks we performed, which confirm practically what one can well imagine by common sense: in presence of a certain probability of false acceptance, one can use databases of templates for successfully impersonating a stranger.

Although these concerns raise serious caveats for the use of biometry, its user friendliness leads to a continuous propagation of the idea of using it in secure applications. Certainly, the concerns are known, but the vague idea of *application* of lower or medium security concerns was brought as an argument. This breaks the fundamental principle C3. of security: suddenly one seems willing to accept secure contexts, in which attacks can be performed in very short time, yet expecting that the outcome is not sufficiently important for motivating such attacks. It is a defendable point of view, when the security context is a small intranet in which users are satisfied with a formal protection; or when biometry protects access to some protected areas or institutions. However, the consequences in view of B2, namely the uniqueness and irreplaceability of biometric traits are poorly thought through.

As an alternative, a separate branch of activity has been dedicated to a mixture of cryptography and biometry, in which cryptography is supposed to well protect the templates of biometry. However, this quite theoretical area of research operates with the questionable notion of biometric traits being *public data*. This assumption does take into account B2 and the possibility of compromising biometric traits—it is though questionable, what the overall amount of security based on public data may be. Most problematic is the fact that despite intense work, the possibility to uncouple biometric from cryptographic security in these settings, and thus breaking the weakest part in the chain has received no convincing answer yet.

Another approach, which we shall discuss in more detail in this paper, considers biometry as some kind of passwords. They allow access to resources, and, like passwords, should be stored under some one way transformation. This works without problem in the deterministic context: the user presents a password, and its hash value is stored. The hash value for one specific password will always be the same, but an attacker cannot recover the password from knowing the hash. In biometry though, any transformation of the template that can allow both to hide the data from intruders and to perform identification will lead to a notable loss of accuracy in the identification process, as compared to identification by means of "cleartext templates." Therefore, even in the context in which clear template matching provides quite low entropy and thus protection levels, the requirement for password protection leads to an additional loss of entropy, and thus even lower security.

These questions are actively discussed in the literature of the last decade. However, the community of biometry security is a mixed one, ranging from engineers with good expertise in image processing and practical implementations of biometrical matching systems, to specialists of information theory and cryptology who bring new ideas from their domains, while treating biometry as a black box yielding an amount of entropy. The responsibility that the entropy be measurable and sufficient is deferred to applicants—which often are not trained for establishing such complex measures. In fact no mathematical or statistical stringent definition of entropy can be accurately applied in the context. It is one of the points which shall make in this paper, that the de facto entropy of some biometric template is simply given by the equal error rate of the system, i.e. the balanced probability of

false accepts and false rejects. It is a realistic, albeit quite low quantity. A further concern of the incipient biometric security research should be the one of giving some accurate definitions of attacks. Like in the case of cryptographic security, these attacks should specify clearly:

A1. What resources one assumes that the intruder may dispose of.
A2. What advantage the intruder wishes to gain.

After providing a brief overview of the currently most frequently used types of biometric identification, we shall focus on the oldest and most spread fingerprint recognition. We shall discuss in this context more in depth the various concerns listed above and provide some partial answers.

3 Overview of Biometry

In our context, biometry is the scientific domain which is concerned with measurements and images of (parts of the) human body, that are to high extent reproducible and may also practically be used for the identification of individuals. In order to be useful in applications, biometry should enjoy some fundamental properties, like

BM1. Universality, meaning that all potential users should possess this biometric trait.
BM2. Uniqueness, in the sense that the biometric trait is different from person to person, and thus helps distinguish individuals and authenticate them correctly.
BM3. Permanence, meaning that the trait will not change in time, and thus, an individual can be identified even on base of templates gathered long periods of time before.
BM4. Some practical properties, such as performance, acceptability, and lack of circumvention. The processing time for identification should be low for reach "acceptable" recognition rates. The acceptability addresses a subjective, social issue: it should be accepted by the bulk of society that presenting one's biometric traits is acceptable. For instance, in some culture, showing the face of a woman and taking pictures of it, might appear as unacceptable, and even presenting one's eye into a camera may require some preparation. Biometric traits may often be imitated by fakes, so it is a requirement mainly for the authentication system, that it be capable of distinguishing between artificial fakes and living biometric sources.

3.1 Fingerprints

It was established already before the turn of the last century, that the fingerprints of humans contain sufficient information for distinguishing between any two

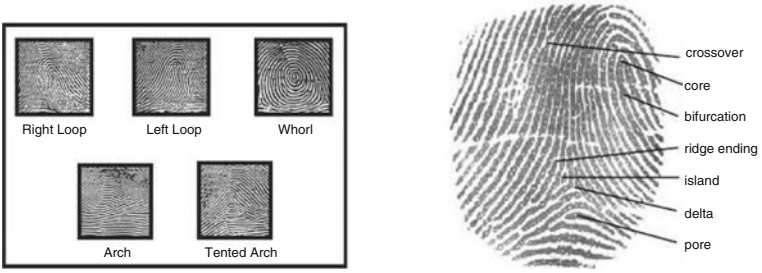


Fig. 4 Distinguishing fingerprints

individuals (Fig. 4). Since, in addition, human leave everywhere there fingerprint, due to the sweat and skin fat, the fingerprint became an important identification method in forensics: techniques for gathering *latent fingerprints* from crime sites developed, together with the science of *dactyloscopy*, which is the craft of fingerprint recognition, in the practice. The fingerprint can be seen as an overall picture of a flaw of ridge lines, induced, in detail with natural endings and bifurcations of the lines. These are called *minutiae*—while cores and delta, visible points of maximal curvature, respectively, of divergence of the ridge flaw, are in general easily identified and used for orientation of fingerprint images and templates. George Dalton classified the types of ridge flaws in five main types: left and right loops, whirls and arches, which may be plain or tented. Experience shows that both fingerprints help distinguish even one eyed twins, and the combination of types for the ten fingers is also highly individual. Therefore a first step in matching fingerprints out of large data bases will always begin with a matching of the 10-tuple of types of the ten fingers. This will lead to a small selection within which a detailed identification based on matching of minutiae can be performed by the specialized dactyloscopist. It is agreed that a reliable matching of between twelve and eighteen minutiae is an acceptable base in court, for acknowledging the identity of a person (Fig. 5). The precise number of identical minutiae may vary slightly from country to country, and one may even encounter some other classifications of ridge flaws than the one of Dalton—but the main features are the same.

With the advent of computers, the machine-identification of fingerprints became a task of study in image processing; dedicated methods were developed and towards the turn of the century sufficiently reliable plaintext matching system had been developed. For very good quality pictures, an error rate of around 0.1 % is frequent, while for pictures of poor quality, even an accuracy of 0.5–1.0 % is acceptable in practice. Encouraged by the improving quality of matching, the idea of applying the password paradigm to biometry was brought in the field, first by Juels and Wattenberg [1] and then again by Juels and Sudan [2, 3]. In this paradigm, biometric templates—fingerprint or others—should be stored in the hashed way, and identification should happen on base of hash values. While in deterministic mode, this approach is very natural, the stochastic nature of biometric matching poses series of problems, and the invention of fuzzy vaults in [2, 3] was the most successful

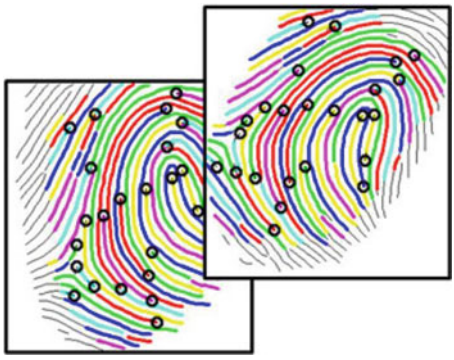


Fig. 5 Fingerprint matching

approach for satisfying this requirement. This comes however together with a loss of matching accuracy that may pose serious problems and leads to a difficult decision problem, pondering security against accuracy of matching. As mentioned above, the last is, in the end, also a matter of security—since one must estimate the entropy by the de facto error rate of the system, so when the error rate increases, the security drops too. We shall discuss these issues in more detail in the following chapters.

3.2 Iris

While fingerprint recognition has an old, forensic born history, the identification based on the human iris is a one-man show. It was the mathematician John Daugman, presently teaching at Cambridge, who recognized the identification potential in the human iris and developed after a lot of work the algorithms and patents for turning this insight into a practical biometrical identification procedure. The human iris has the advantage of a perfect crown-circular geometry, making its localization in images an easy task. The base for recognition are a system of log-like lines which are different in thickness and frequency, from person to person. Daugman had the bright idea of performing plain Fourier transforms on the iris picture, after having processed it and enhanced image qualities, while unfolding the circle along a line. The result of the analysis is a code of 256 which was standardized and patented by Daugman, as the *iris code*. Claims are that between 20 % and 30 % of identical bits in this code helps ascertain the identity of a person with error rates with in the one per million. Iris has been implemented at various airports, due to its claimed accuracy. Since biometry is not a deterministic science, as soon as iris recognition went public and entered scrutiny of various university research groups, new questions were raised, the claimed recognition rate slightly dropped and even the question was raised, if the iris imprint is permanent in time and if it did not chance after diseases and other organic disturbances. After all, the permanence of the fingerprint had been empirically watched in forensics over more than 100 years, while iris identification is only two decades old. Despite these dis-

Fig. 6 Palm recognition

cussions, iris recognition is certainly among the leading biometric identification resources, and it has possibly the most impressive accuracy among all. On the other hand, fingerprint recognition can easily improve its performance by using multiple finger recognition.

3.3 Palm

Palm recognition is a good alternative to fingers, which gained much popularity in the last decade. The identification artifacts are similar to those for the fingers, but the advantages stem from the fact that palms are easy orientable, better protected from scars and optical disturbances which are a source of poor image quality for fingers, and, finally, have a high amount of information (Fig. 6).

3.4 Face

Face recognition is an application as old as computers. There are multiple approaches to face recognition, from flat, two dimensional images, to three dimensional simulations gained by the use of multiple cameras and angles of image acquisition. However, the challenges are very high, since face is the biometrics with the highest dynamics—it may vary due to momentary expression, but also to usual changes, such as make up, eyelid enhancements for women, or beard growing for men. As a consequence, an identification error rate below 5 % is quite rare for face recognition. The practical applications are less in the authentication context, and more for the identification of faces in moved contexts and real life scenes.

Fig. 7 Hand veins identification

3.5 Hand Veins

Hand veins are the youngest type of biometric identification method. It has been claimed that comparing hand vein geometry can lead to identification rates much superior to the one of fingerprints, and reaching in the area of accuracy known from the iris recognition. Picture of hand veins can be taken by means of infrared cameras, which became affordable in the last years, due to technological development. Since hand veins are not exposed, the pictures are very stable and uninfluenced by wounds or physical condition of the scanners. These facts speak in favor of deployment of hand veins as biometrical identification method. Unfortunately, producers of vein-scanners have started a new trend by producing also their proprietary, system embedded, matching algorithms. As a consequence, the academic research with hand veins is at most incipient, and encountering practical problems, being reduced to develop own hard- and software (Fig. 7).

3.6 Various Other Biometries

The above are the most important and widely spread biometric traits used for identification. However, numerous other typical and distinctive traits have been researched, for the purpose of biometric identification. Voice has an important role in applications of telephony and its potential has thus been thoroughly studied. Thermoscans of hands or other parts of the body can also be used for identification, as well as can the mechanics of human gait help recognize individuals with a certain accuracy. Inspired by the hand signature, engineers have built special purpose pens which integrate the *writing mechanics* of individuals, while, for instance, writing down their signature. It is then the mechanical plot and not the actual signature which is used for identification. These and other biometric investigation either have specific ranges of application where they can be of use, or are a matter of pure research. Their identification rate is in general quite poor, in the range of face recognition.

3.7 Present Applications

In the last two decades, the applications of biometry reached most diverse areas of social life. In several countries, the drivers' license or id card carries a fingerprint for identification. Meanwhile ATM machines using biometry for identification, based either on iris, palm or fingerprint, are used in several, mostly Asian countries. Fingerprints are used as replacement for signatures especially in third world countries with a high rate of illiteracy. The same kind of biometric traits may also be used for access control in hotels, museums, clubs or lounges, as car openers or weapon activators, etc. In the area of surveillance techniques, face and gait recognition naturally play an important role. While the introduction of biometry in international passports is being pushed ahead world wide, it becomes more and more important to achieve some reliable security standard in the domain of biometric applications in security contexts.

4 "Hashes" for Biometry

Cryptographic password hashes are common solutions for storing passwords in a protected form, while enabling verification of genuine users. However, as discussed above, merely relying on a person's ability to reproduce a password in order to verify her authenticity leads to certain problems—e.g., key management. For this reason, biometry came to be considered as an alternative, possibly in combination to passwords. While the requirements for *biometric template protection solutions* are similar to those for user password protection, they are more difficult to achieve: Passwords are deterministic whereas biometric templates, at the contrary, are typically subjected to noise, i.e., multiple matching samples are expected to be different while they also have some reasonable similarity. These differences can be usefully conceptualized as errors. In this way, biometric template protection schemes have been proposed, that combine techniques known from traditional cryptography with techniques from the discipline of *error-correcting codes* to allow error-tolerant verification.

4.1 The Fuzzy Commitment Scheme

One of the conceptually simplest approaches for generating protected data from biometric templates that allows error-tolerant verification was proposed by Juels and Wattenberg in 1999 as the *fuzzy commitment scheme* [1].

Let $\mathbf{F}$ be a finite field and assume that we are given the decoder of an *error-correcting code* $\mathbf{C} \subset \mathbf{F}^n$,[2] that is a function

$$\mathsf{dec} : \mathbf{F}^n \to \mathbf{C} \cup \{\text{FAILURE}\},$$

for which there exists an integer $\epsilon \geq 0$ such that $\mathsf{dec}(v) = c$ for all $c \in \mathbf{C}$ and $v \in \mathbf{F}^n$, if the *Hamming distance* fulfills $|c - v| \leq \epsilon$.[3]

4.1.1 Enrollment

On enrollment, given a biometric template encoded as an n-length bit feature vector $x \in \mathbf{F}^n$, its cryptographic hash value $h(x)$ is computed. Then, a codeword $c \in \mathbf{C}$ is chosen at random and the offset $c + x$ is computed. Finally, the hash value together with the offset is stored as the *private template* $(h(x), c + x)$.

4.1.2 Verification

On verification, given a query template $x' \in \mathbf{F}^n$ of the (alleged) same user, an attempt for recovering the protected vector x is performed by computing $(c+x)-x'$. If x' differs in no more than ϵ positions from x, i.e., if $|x - x'| \leq \epsilon$, then $\mathsf{dec}((c + x) - x') = c$ due to the error-correcting property of the decoder. If in this way the correct codeword c can be recovered, then the correct feature vector can be computed according to $x = (c + x) - c$. The correctness of the result can be verified by using its hash value $h(x)$. Otherwise, if $|x - x'| \geq \epsilon$, any verification attempt results, with high probability, in FAILURE or, otherwise, the decoder may output another candidate for the correct feature vector; in both cases, the verification attempt is rejected.

4.1.3 Security

If we assume that the feature vectors x are distributed uniformly and independently among all elements from $\mathbf{F}^n$, then the complexity of the operation of recovering the correct feature set $\mathbf{A}$ from a fuzzy commitment $x + c$ is provably of $O(|\mathbf{F}|^k)$—or the complexity of breaking the hash $h(x)$ [1]. However, it is not realistic to assume in

[2] For more details on error-correcting codes, we refer the reader to [4] or any other good textbook on the subject.

[3] Here $|\cdot|$ denotes the *Hamming weight* of a vector in $\mathbf{F}^n$, i.e., the number of positions at which the vector has non-zero entries.

biometric disciplines that the templates are distributed uniformly within the feature space. We will later emphasize that optimistically estimating security using i.i.d. assumptions easily leads to a severe overestimation of effective security (Sect. 4.4).

4.1.4 Designing Problems

In order for the fuzzy commitment scheme to be applicable for protecting biometric templates of a certain biometric modality, the following conditions have to be fulfilled:

(1) It must be possible to encode biometric templates as fixed-length feature vectors from $\mathbf{F}^n$.
(2) The similarity between biometric templates must be correlated with the Hamming distance of their corresponding feature vectors.
(3) An error-correcting code $\mathbf{C} \subset \mathbf{F}^n$ of sufficient size for which there is a known efficient decoder dec must exist.

Encoding biometric templates as fixed-length feature vectors is usually not a big problem for it is possible to adopt the binary representation of the biometric templates thereby working in the field with two elements. However, ensuring that this binary representation allows comparison via the Hamming distance represents one of the main challenges when designing fuzzy commitment-based biometric template protection. Furthermore, a generic consideration of the concept of error-correcting is not sufficient for implementing a practical fuzzy commitment scheme. First, the code must have a sufficient size in order to allow a certain protection against reversibility attacks; second, it is not known whether there exist codes with efficient decoders for arbitrary block length n.

Even though the generic concept of the fuzzy commitment scheme is very simple, yet clever, the design of a certain fuzzy commitment-based biometric template protection may be challenging. In fact, we have to focus on the specific biometric modalities for which biometry hashes is to be implemented. Thereby, we set our focus to fingerprints even though similar but individual problems exist for other modalities.

4.2 Fingerprints and its Minutiae

A *fingerprint* is given by the traces that the ridges of a finger leave on a surface. In these modern days, digital scanners can be used (including specific fingerprint sensors) to obtain a digitized image, i.e., the *fingerprint image* of these traces (see Fig. 8a for an example). Typically, features are extracted from fingerprint images on which base two fingerprints can be compared. A standardized type of fingerprint features are *minutiae*, i.e., the positions at which a fingerprint ridge ends abruptly or where it bifurcates, i.e., a *minutiae ending* or *minutiae bifurcation*, respectively.

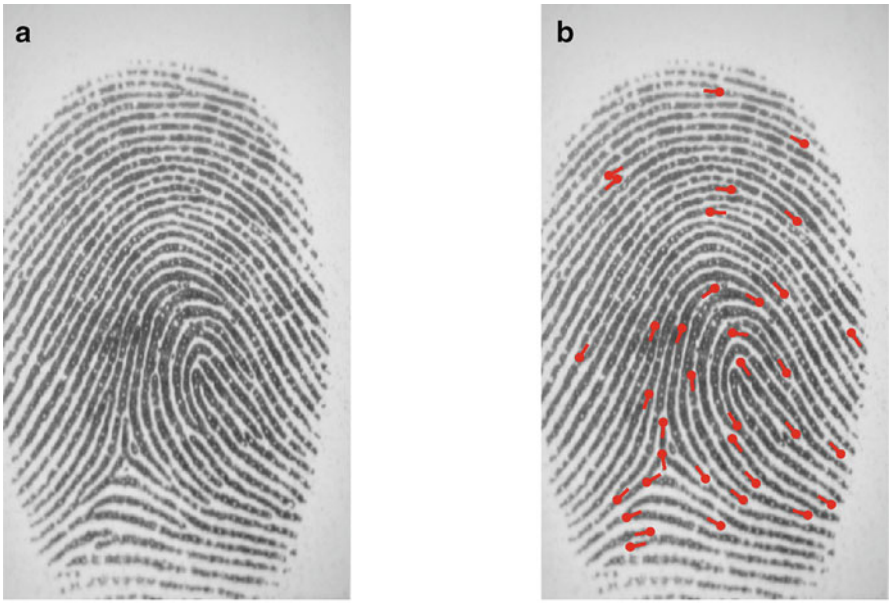


Fig. 8 A fingerprint (**a**) and its minutiae (**b**)

Furthermore, these minutiae positions are typically attached with a *minutia angle* (see Fig. 8b for a visualization of an example). Given two minutiae feature sets, i.e., two *minutiae templates*, comparison may be performed through the adoption of two-dimensional point registering methods accounting for the minutia angles. For further details as well as for a comprehensive overview on fingerprints, we refer the reader to [5].

4.2.1 Fuzzy Commitment Scheme for Fingerprint Minutiae

In order to apply the fuzzy commitment scheme for protecting a fingerprint's minutiae template it is necessary to find an encoding of a minutiae set to a space of fixed-length feature vectors in which similarity between minutiae sets is reflected by the Hamming distance. The probably simplest approach to extract a binary n-length feature vector from a minutiae template may work as follows:

(1) The fingerprint image is partitioned into n disjoint regions and each region is attached with a unique index varying between 0 and $n-1$;
(2) For the feature vector $x = (x_0, \ldots, x_{n-1})$ we set $x_i = 1$ if a minutia is contained in the i-th grid region and, otherwise, if no minutia is contained in the i-th region, then $x_i = 0$.

The above approach is, however, not very eligible for fingerprint minutiae since usable binary error-correcting codes typically must have a block length n that

matches a certain form, e.g., $2^m - 1$ for BCH codes. This explicitly and implicitly leads to limitations when designing fuzzy commitment scheme-based template protection for fingerprint minutiae. There exists an implementation of the fuzzy commitment scheme to fingerprint minutiae [6]; nevertheless, the scheme is better tailored for use in other biometric modalities, such as human irises [7].

4.3 The Fuzzy Vault Scheme

In 2002, Juels and Sudan [2, 3] have proposed the *fuzzy vault scheme* solving some of the problems that we may encounter when attempting to implement a fuzzy commitment scheme for fingerprint minutiae. Like in the case of the fuzzy commitment scheme, the fuzzy vault scheme uses techniques from coding theory in order to conceptualize differences between biometric samples and it has been formulated in quite general terms.

In the following, we shall restrict our considerations to the fingerprint modality. Roughly speaking, the vault works as follows in this case: The minutiae of the to-be-hashed fingerprint, called *genuine minutiae*, are hidden within a large number of randomly chosen non-authentic minutiae, called *chaff minutiae* (see Fig. 9a). The genuine minutiae are attached with some information by means of *Reed-Solomon error-correcting codes* while the chaff minutiae are attached with random information deemed to be indistinguishable from the information with

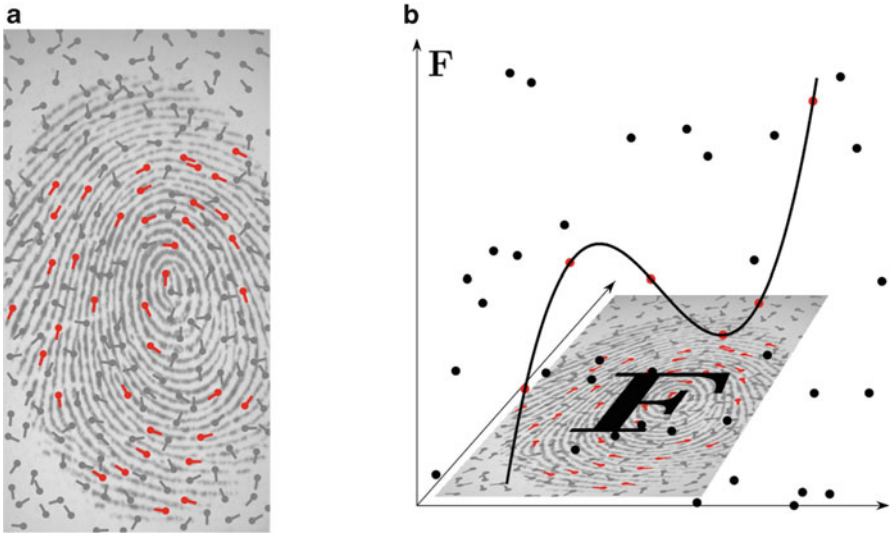


Fig. 9 (**a**) A fingerprint and its genuine minutiae (*red*) hidden among a large number of chaff minutiae (*gray*). (**b**) Visualization of the genuine minutiae being bound to a Reed-Solomon codeword

which genuine minutiae are constituted thereby providing a certain protection against recovery of the genuine minutiae set from the »minutiae cloud«, i.e., *vault minutiae*. On verification, the minutiae of the query fingerprint are used to extract the *unlocking minutiae*, i.e., those vault minutiae with which query minutiae are of reasonable agreement; if the query fingerprint stems from the same finger, then we may have reason to assume that the unlocking minutiae dominantly consists of genuine minutiae in which case we can recover the entire genuine minutiae set exploiting the error-correcting property; otherwise, if the query minutiae stem from another finger, the unlocking minutiae is expected to contain too few genuine minutiae for recovery.

The eligibility of protecting minutiae with the fuzzy vault scheme has been analyzed by Clancy et al. in 2003 [8] which resulted in a series of minutiae-based fuzzy vault implementations [9–12]. In the following, we present the functioning of a minutiae-based fuzzy vault mainly following the description of Nandakumar, Jain, and Pankanti [12].

4.3.1 Enrollment

On enrollment, a minutiae template is given that we want to protect. These minutiae, called genuine minutiae, are mapped to an encoding of a fixed finite field $\mathbf{F}$ by some fixed convention such that there is a one-to-one correspondence between minutiae and the finite field element encoding them.[4] We thus obtain the so-called set of genuine features, also called *feature set* $\mathbf{A} \subset \mathbf{F}$, encoding the minutiae to be protected. We assume that the number of minutiae, namely the size of the feature set, is public and denote it by $t = |\mathbf{A}|$. A secret polynomial $f \in \mathbf{F}[X]$ of degree smaller k is chosen uniformly at random and will be later dismissed. Using f, the genuine pairs $\mathbf{G} = \{(x, f(x)) \mid x \in \mathbf{A}\}$ are computed; this produces a binding of the genuine template to the secret polynomial f. After this, a random set of $n - t$ *chaff features* is generated $\mathbf{A}_{\text{chaff}} \subset \mathbf{F}$ which should be indistinguishable from genuine features $\mathbf{A}$. Finally, chaff pairs $\mathbf{C}$ are generated; they are pairs $(x, y) \in \mathbf{F} \times \mathbf{F}$, in which $x \in \mathbf{A}_{\text{chaff}}$ and y is chosen uniformly among all elements in $\mathbf{F}$ with the constraint that $f(x) \neq y$. The union $\mathbf{V} = \mathbf{G} \cup \mathbf{C}$ finally builds the vault of size n, to which one typically attaches a cryptographic hash $h(f)$ of the secret polynomial, in order to allow secure verification. Consequently, the protected record is given by the pair $(\mathbf{V}, h(f))$.

[4] For example, the integral encodings of the abscissa coordinate, ordinate coordinate and angle of a minutia can be concatenated thereby obtaining a single integer that can be used to encode an element of the finite field if of sufficient size. It is important to note that from the finite field element encoded by an integer, the minutia coordinates and minutia angles can be recovered.

4.3.2 Verification

Upon verification, a query feature set $\mathbf{B} \subset \mathbf{F}$ encoding a query minutiae set is provided. Based upon this set, vault pairs are extracted from $\mathbf{V}$, such that the abscissae (encoding a genuine/chaff vault minutia) are well approximated. The unlocking pair set $\mathbf{U} \subset \mathbf{F} \times \mathbf{F}$ is thus determined. If we assume that the query minutiae stem from the same finger as the protected minutiae, then we may expect that a significant amount of query minutiae agree with the genuine minutiae protected by the vault record. In such case, $\mathbf{U}$ may consist of a significant amount of genuine pairs $(x, y) \in \mathbf{G}$, which lie on the graph of the secret polynomial $f \in \mathbf{F}[X]$, i.e., $f(x) = y$. In particular, if $\mathbf{U}$ consists of at least $(|\mathbf{U}| + k)/2$ genuine pairs, then the secret polynomial can be efficiently recovered using an algorithm for decoding Reed-Solomon codes [4].

4.3.3 Brute-Force Security

An intruder who has intercepted a vault record $(\mathbf{V}, h(f))$ may attempt to guess k vault pairs from $\mathbf{V}$ and hope that they are genuine. In case they are genuine, their interpolation polynomial will reveal the correct polynomial of which correctness can be verified with $h(f)$.[5] There are $\binom{n}{k}$ possibilities for an attacker to choose vault pairs of which $\binom{t}{k}$ will reveal the correct polynomial. Hence, with probability $\binom{t}{k} \cdot \binom{n}{k}^{-1}$ an attacker can guess the correct polynomial. This yields a notion of *brute-force security*.

It is important to note that the brute-force attack is based on the unrealistic assumption that minutiae are distributed uniformly and independently from each other. We later emphasize that merely relying on brute-force security as a notion for the security of a fuzzy vault will yield a strong overestimation of the effective security (Sect. 4.4).

4.3.4 Pre-alignment

A very delicate problem with which implementations of minutiae-based fuzzy vault schemes have to cope with is the problem of fingerprint alignment during a genuine verification process. A common approach is to store unprotected helper data of the protected fingerprint (e.g., points of high ridge curvature) along with the vault records which can be used on verification to pre-align the query templates coarsely in a preliminary step [10–14]. Then, the query minutiae may be adjusted to the vault minutiae to obtain the final alignment with which the unlocking set is extracted [12–14].

[5] Even if the hash were not available, an intruder has still the opportunity to check whether the candidate polynomial interpolates $t = |\mathbf{A}|$ vault pairs; a wrong candidate polynomial will with overwhelming probability not fulfill this requirement for parameters that we expect to encounter in practice, thereby yielding a reliable criteria to an attacker to identify the correct secret polynomial.

From a security perspective, the use of public auxiliary alignment data is problematic because it *does* leak information about the protected fingerprints. Li et al. [15] proposed to use features from the fingerprint that do not depend on the fingers rotation and placement; for instance, features derived from minutiae triangle constellations, thereby removing the issue of information leakage from auxiliary alignment data.

4.3.5 Implementations

One of the first automatic implementations of a fingerprint-based fuzzy vault has been presented by Uludag and Jain in 2006 [11]. They bound the number of minutiae that are protected in the vault by $n = 18$ and bind them to a polynomial of degree smaller than $k = 9$, thereby yielding a brute-force security of 2^{-36}. The genuine acceptance rate that the authors achieved was 73 % at which no false accepts have been observed.

In 2007, Nandakumar, Jain, and Pankanti [12] improved the genuine acceptance rate to 86 % (again, for no observed false accepts) in which at most $t = 24$ genuine minutiae bounded to a polynomial of degree smaller than $k = 11$ are protected within $n = 224$ vault minutiae thereby yielding a brute-force security of 2^{-39}. Nandakumar, Nagar, and Jain suggested that the security of their vault implementation can be furthermore improved via a user password [13].

In 2010, Nagar, Nandakumar, and Jain showed how additional features of the fingerprint can be used to improve brute-force security by protecting the vaults' ordinate values with a fuzzy commitment scheme [14]. In particular, they showed that a genuine acceptance rate of 92 % is achievable at a brute-force security of approximately 2^{-40}.

The above three implementations all require a preliminary alignment step which is supported by public auxiliary alignment data stored along with the vault records. Public data which does leak information can also be exploited by an adversary to improve attacks. Therefore, Li et al. designed a fuzzy vault for fingerprints protecting features that do not depend on the fingerprint's alignment. In this implementation, $t = 40$ genuine features bound to a polynomial of degree smaller than $k = 14$ are hidden within $n = 440$ vault features; this yields a brute-force security of 2^{-52}. The authors measured a genuine acceptance rate of 92 %, again at no observed false accepts.

4.4 Fundamental Security Limit

Above, the security analyses of the respective fuzzy vault implementations are based on the assumptions that fingerprint features are distributed uniformly and independently from each other in the vault. In this section we give simple but yet irrefutable arguments why brute-force security is not even a close measure of the implementation's effective security.

We start with a simple exemplary observation. Consider the implementation of Nandakumar, Pankanti, and Jain [12]. For the parameter configuration in which $n = 224$ vault minutiae hide at most $t = 24$ genuine minutiae being bound to a secret code polynomial of degree smaller than $k = 9$, the genuine acceptance rate evaluates as 91 % while the false acceptance rate estimates as FAR ≈ 0.01 %. Now, an intruder who has intercepted a vault that he aims to break, i.e., recover the genuine minutiae from it, may establish a large database containing real fingerprints. With these fingerprints he may simulate verification attempts successively until he successfully breaks the vault. Given the computational complexity for simulating an impostor's verification attempt IDT the adversary can expect to break the vault after a time of

$$\mathrm{IDT} \cdot \frac{\log(0.5)}{\log(1 - \mathrm{FAR})} \tag{1}$$

yielding the notion of *false-accept security*. In [12] it has been reported that a verification lasted

$$\mathrm{IDT} \approx 33 \text{ “Lagrange interpolations”}. \tag{2}$$

Consequently, in terms of Lagrange interpolation for $k = 9$ the false accept security is estimated as approximately

$$2^{18} \text{ “Lagrange interpolations”} \tag{3}$$

which, however, strongly contrasts with an estimated brute-force security of 2^{31} Lagrange interpolations as a realistic measure for the implementation's overall security.

Even in case that no false accepts have been observed during performance evaluation, this does not imply that the false acceptance rate is negligible or even zero: The false acceptance rate is not negligible and the above observation emphasizes more than clearly that *brute-force security is not more than a coarse upper bound for the security of current biometric template protection schemes such as fuzzy vault. Each measure that significantly overestimates false-accept security should be seriously questioned.*

The situation is quite serious. Even a barely usable protection scheme for single finger typically only provides quite-a-low brute-force security of order 2^{31}, say, which is very weak from a cryptographic point of view: It is absolutely no problem to reveal the protected minutiae templates from such a vault within a few minutes. The situation is even worse as an attacker can exploit the statistics of fingerprint minutiae features. An indication of the maximal achievable security bound is given by the false-accept attack. The complexity of such attacks can be estimated to be in the order of 2^{18}: this amount of operation is a matter of only a few seconds for the attacker—even when using personal computers. In view of these observations, it seems questionable whether there exists sufficient information

on a finger in order to achieve a reasonable amount of security. The methods of fingerprint feature extraction and matching have been upgraded, so, one expects substantial improvements of the security. But even if the false acceptance rate can be reduced to the half—a tremendous improvement, indeed—the false-accept security only slightly improves by a single bit. One can thus hardly expect that fingerprint recognition can evolve in such a way that template protection of single fingerprints may become secure in a cryptographically acceptable way. It is important to note that also for other biometric modalities, such as a human's iris, that can provide higher genuine acceptance rates at lower false acceptance rates than fingerprints, have a security that is still rather low from a cryptographic point of view [7].

4.4.1 Combination with Passwords

A possible countermeasure may be to combine passwords with a biometric template, e.g., fingerprint minutiae, to improve security. Such an approach has been implemented and tested in which the minutiae-based fuzzy vault implementation [12] was additionally protected via a 64 bit user password [13]: Using a user password, the vault minutiae are shuffled and on verification given the correct user password, the vault minutiae can be transformed back to their original position. One may argue that the incorporation of user passwords may result in key management problems that were meant to be resolved with biometry: Again, the user of a system has to remember passwords which can be forgotten or, if written down, drop security. On the other hand, biometry can be used to improve password security by a certain amount, for example, 18 bits in case fingerprint minutiae are used—even for easily memorable passwords such as 4-digit *person identification numbers* PINs, say. The weak security of 13 bits provided by a 4-digit PIN can consequently be improved to $32 = 13 + 18$ bits using a single fingerprint's minutiae. For such an approach it must be guaranteed that correctly decrypted vault data is indistinguishable from falsely decrypted vault data—which, in fact, was not guaranteed in [13].

4.4.2 Slow-Down Functions

Another approach to improving the low security of biometric template protection is to implement slow-down mechanisms. In a password-based scheme, the password hashes may be hashed multiple, say a million, times. This yields virtual 20 bits of additional security while also increasing the verification time which is, especially in view of genuine verification attempts, a disadvantage.

In biometric template protection schemes, such as fuzzy vault, merely repeating the hashing process of the secret key's data bound to the template would not be a valid solution. An attacker running a false-accept attack may most likely be able to distinguish the correct secret polynomial from false polynomials without computing its hash. The correct polynomial is of known degree k interpolating $t \gg k$ vault pairs; other query templates will most likely not fulfill this requirement. Similar

observations apply to the fuzzy commitment scheme and other constructions based on error-correcting codes [1, 16]. Note that this observation has not been accounted for an iris-based fuzzy commitment scheme implementation which was proposed by Hao, Anderson, and Daugman [7]. There, the possibility of repeated hashing of the secret codeword has in fact been proposed for improving the security. Nevertheless, this measure yields virtually no additional security due to typically negligible *sphere packing densities*[6] of most *error-correcting codes* [4].

The following may be a valid approach to artificially slow-down the verification process in which the possibility of additionally encrypting the protected biometric templates with password is exploited. A *quiz* $\kappa \in \{0, \ldots, K-1\}$ is chosen at random during the generation of a protected template and is used to encrypt it. The data of the quiz q is then dismissed. Herewith, the verification process can be artificially slowed down—in particular an impostor verification attempt. Upon a failing impostor verification attempt, since the correct quiz q is unknown, all possible quizzes $q' = 0, \ldots, K-1$ must be used to temporarily decrypt the protected reference template and against each temporarily decrypted protected reference template a verification is performed. Consequently, the false-accept security increases by $\log_2(K)$ bits while, on average, the genuine decoding complexity is also increased by a factor of $K/2$. Thus, the slow-down factor K must be chosen carefully in order to achieve sufficient security while still keeping genuine verification feasible. Consequently, the relation between system security and genuine verification time cannot be changed by slow-down measures, i.e., the *security factor* remains unaffected and is potentially low.

4.4.3 Multiple Fingerprint/Multiple Biometric Modalities

To overcome the problem of low security factors in a fingerprint-based fuzzy vault, we may consider to fuse multiple fingerprints acquired from a user and protect them with the fuzzy vault scheme. On genuine verification, more than one fingerprint may be required for an accept. On the other hand, breaking a fuzzy vault to multiple fingerprints may also be more secure against attacks, in particular, false-accept attacks. An implementation of a multi-finger fuzzy vault has been proposed by Merkle et al. in 2011 [17]. It is important to note that the implementation has not been evaluated in terms of genuine acceptance rate and false acceptance rate and it is still unclear how a multi-finger implementation can perform.

It is also possible to fuse multiple biometric modalities of which the fusion of fingerprints is a special case. In 2012, Nagar et al. [18] analyzed fusion strategies of fingerprints, irises, and face using the fuzzy vault and fuzzy commitment scheme and reported that it is possible to achieve a 75 % genuine acceptance rate at a security level of 53 bits. In principle, this is an interesting result. It remains, however, unclear if and in which applications a fusion of fingerprints, iris, and face, which may be

[6] Sphere packing density: $|\mathbf{F}|^{n-k} \sum_{j=0}^{\epsilon} (|\mathbf{F}|-1)^j \binom{k}{j}$ where $k = \dim(\mathbf{C})$.

related with several inconveniences for users, will be of interest. Especially under the circumstance that the moderately high security of 53 bits is compensated by the very low genuine acceptance rate of merely 75 %.

4.5 Attacks Via Record Multiplicity

Even if we can assume that it is possible to implement a usable biometric template protection scheme, possibly based on multiple fingerprints or more generally on multiple biometric modalities, there are, however, other risks that must be considered. In addition to mere off-line attack in which an adversary aims at revealing the protected templates from intercepted data, there exist another serious scenario in which an adversary who has intercepted two (or more) protected records attempts to decide whether they stem from the same finger, say, i.e., whether they are *related*. The process of distinguishing related from unrelated records is commonly called *cross-matching* and is a privacy risk with which an intruder having intercepted the content of multiple application's database could trace particular users activity. For this reason international standards explicitly require from biometric template protection schemes to be *unlinkable*, i.e., cross-matching must not be possible (ISO/IEC IS 24745 [19]).

4.5.1 Correlation Attack in a Fuzzy Vault Scheme

In general, the fuzzy vault scheme is vulnerable to a very serious cross-matching attack [20]. Observe that in a fuzzy vault, the genuine features stem from a biometric sample while the chaff features have been generated at random. If two fuzzy vault record can be intercepted by an intruder protecting templates that stem from the same instance (e.g., finger) we may observe that the genuine vault features in the first record (e.g., red-colored minutiae in Fig. 10a) well agree with the genuine vault features in the second record (blue-colored in Fig. 10b), i.e., the correlate well as compared to the chaff features (Fig. 10c, d). This property can be exploited by an intruder to distinguish related from unrelated vault correspondences. Even worse, an intruder who has intercepted two related vault records may even unlock the vaults using the candidate sets of genuine vault features. In fact, for a minutiae-based fuzzy vault implementation, Kholmatov and Yanikoglu [21] demonstrated that an intruder can break two related vault correspondences with success probability of order 60 %, which is much too high for a system to fulfill the unlinkability and irreversibility requirement. The possibility of running the so-called *correlation attack* calls for a valid countermeasure.

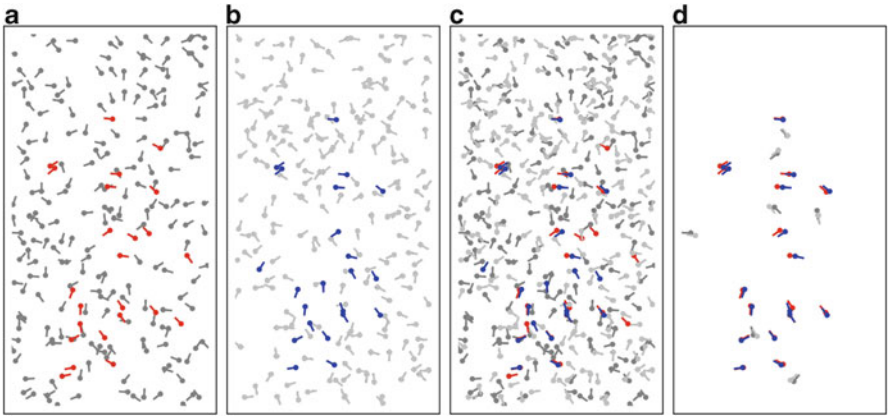


Fig. 10 Visualization of the correlation attack process in a minutiae-based fuzzy vault: those vault minutiae of two related vaults (**a**) and (**b**) are correlated (**c**) and those vault minutiae that well agree have a quite good chance to be genuine minutiae (**d**) being colored *red* and *blue* for the first and second vault, respectively

4.5.2 Decodability Attack in a Fuzzy Commitment Scheme

At a glance the serious vulnerability of the fuzzy vault scheme not to fulfill the unlinkability requirement advocates to base the protection on the fuzzy commitment scheme (see Sect. 4.1). However, the fuzzy commitment scheme is vulnerable to a linkability attack, too. With the notation of Sect. 4.1, assume that an intruder has intercepted two related records of the fuzzy commitment scheme $c + x$ and $c' + x'$, i.e., where $c, c' \in \mathbf{C}$ are random elements of a linear code $\mathbf{C} \subset \mathbf{F}^n$ and $x, x' \in \mathbf{F}^n$ are feature vectors with Hamming distance within the code's error-correcting capability ϵ, i.e., $|x - x'| \leq \epsilon$. The intruder has the possibility to compute the difference

$$(c + x) - (c' + x') = (c - c') + (x - x') \tag{4}$$

and exploit the observation that $c - c'$ is a codeword, due to the linearity of $\mathbf{C}$, and the bound $|x - x'| \leq \epsilon$. Hence, the difference can be decoded to the codeword $c - c'$ given two related records of the fuzzy commitment scheme. For non-related records, i.e., where $|x - x'| > \epsilon$, the difference may be decodable with probability equal to the sphere packing density of $\mathbf{C}$; this is typically negligible for most linear codes used in implementations of the fuzzy commitment scheme. Thus, just from decodability of the difference, an intruder may distinguish related from non-related records thereby conflicting with the unlinkability requirement. It is known that capturing two related records based on the linear code brings no advantage for breaking the fuzzy commitment scheme. But if an intruder has intercepted two related records based on different linear codes, the irreversibility requirement cannot be guaranteed [22].

In a binary fuzzy commitment scheme, Kelkboom et al. [23] proposed to pass the feature vectors through a record-specific but public permutation process, in order to prevent the decodability attack. Unfortunately, it has been overlooked that by implementing the measure, two related records of the fuzzy commitment scheme being subjected to different public permutation processes can be considered as having been built by means of different linear codes. This makes them susceptible to the reversibility attack mentioned above [24]. It has furthermore been shown in [24] that in a binary fuzzy commitment scheme, the problem cannot be solved by passing the feature vectors through a public transformation process that preserves the Hamming distance between two feature vectors. Fortunately, there may exist such transformations for a non-binary fuzzy commitment scheme. However, most implementations of the fuzzy commitment scheme are used to protect binary biometric feature vectors and thus the problem of designing an effective binary template protection scheme remains a challenge.

4.5.3 Unlinkable Minutiae-Based Fuzzy Vault

The correlation attack in a fuzzy vault scheme yields an advantage to an attacker in linking and breaking two related records due to the fact that the chaff is generated at random while the genuine features stem from the same instance thereby essentially being fixed (up to tolerable noise). We may avoid this inconvenience in a fuzzy vault scheme by a simple yet effective variation, which we describe next, in informal terms, for fingerprint minutiae. A grid (e.g., rectangular or hexagonal) is laid over the fingerprint image; each genuine minutiae is rounded to grid coordinates that are then used to build the genuine features, thereby passing the genuine minutia through a quantization scheme (we may also quantize the minutiae angles in a similar manner). All other, unoccupied grid coordinates are used as the chaff. Consequently, there is no correlation that can be exploited in an attack, since the feature sets are equal for any two records (Fig. 11).

Some challenges remain with this approach. Upon verification, it is not possible to adjust query minutiae to vault minutiae in order to unlock the vault with a pre-aligned minutiae set. Alternatively, we have to ensure that the genuine minutiae and the query minutiae can be represented w.r.t. an intrinsic coordinate system that can be robustly extracted from a fingerprint. This introduces new error sources. In fact, the estimation of intrinsic coordinate system and an alignment-free representation of minutiae, i.e., *absolutely pre-aligned minutiae*, is a challenging problem for which no definite solution has been found [5].

Recently, some progress in automatic absolute minutiae pre-alignment has been presented [25] and evaluated for an unlinkable minutiae-based fuzzy vault. The genuine acceptance rate that is currently achievable with such an approach is of order 80 % at which no false accept has been observed while in a traditional minutiae-based fuzzy vault the genuine acceptance rate has been measured as 86 % on the same database. Consequently, the unlinkability requirement can be fulfilled at a genuine acceptance rate well comparable to a traditional approach which,

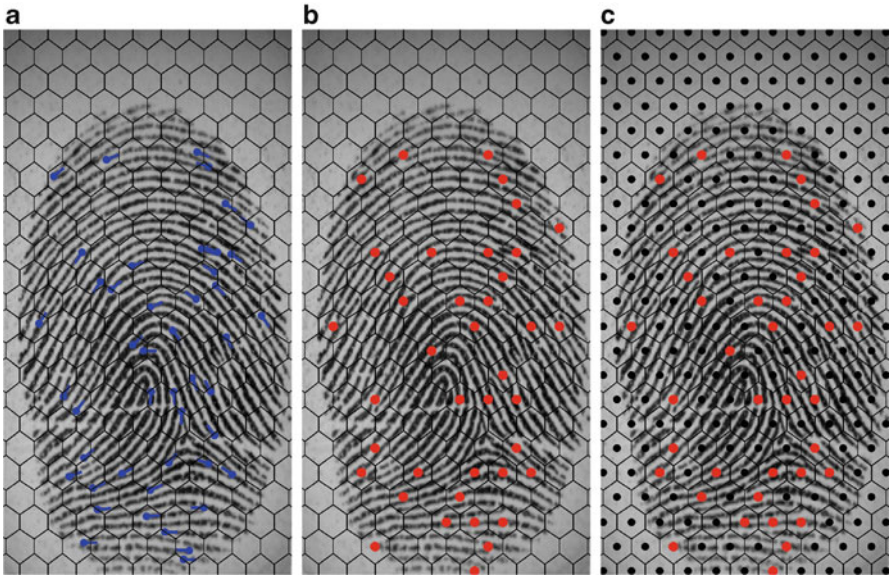


Fig. 11 Visualization of how to make a minutiae-based fuzzy vault resistant against the correlation attack: (**a**) the genuine minutiae are rounded/quantized as coordinates of a (for example) hexagonal grid (**b**) and all other unoccupied grid coordinates are used as the chaff (**c**)

however, is prone to record multiplicity attacks. It is important to note that both implementations following the traditional approach and by applying a quantization scheme to the minutiae are subject to the fundamental security limit discussed in Sect. 4.4. However, by incorporating a quantization scheme and robust methods for absolute fingerprint pre-alignment we may eventually obtain an unlinkable biometric template protection scheme for multiple fingerprints and/or even multiple biometric modalities. It remains to be seen how implementations for multiple fingerprints will be able to perform regarding verification performance and security.

4.5.4 A Compact Fuzzy Vault Scheme

Passing absolutely pre-aligned minutiae through a quantization process has another advantage, beyond merely achieving resistance against the correlation attack. We can apply a modified fuzzy vault construction proposed by Dodis et al. [16] for protecting quantized minutiae sets. This has the advantage of producing significantly more compact record sizes.

As above, the quantized minutiae are encoded as a subset $\mathbf{A}$ of the underlying finite field $\mathbf{F}$. Furthermore, as before, let $f \in \mathbf{F}[X]$ be a secret polynomial of degree smaller than k. Instead of chaff generation thereby yielding a set of vault pairs explicitly, they are encoded by the following polynomial

$$V(X) = f(X) + \prod_{x \in \mathbf{A}} (X - x). \tag{5}$$

If $x \in \mathbf{A}$, then $V(x) = f(x)$ and thus $(x, V(x))$ is a genuine pair lying on the graph of the secret polynomial; otherwise, if $x \notin \mathbf{A}$, then $V(x) \neq f(x)$ and then $(x, V(x))$ is a chaff pair. Consequently, by a single compact polynomial, genuine and chaff pairs are encoded in a smart manner. Note that $V(X)$ is a monic polynomial of degree $t = |\mathbf{A}|$. It only requires $t \cdot \log_2(|\mathbf{F}|)$ storage bits, while the traditional vault would need $2 \cdot (t + |\mathbf{C}|) \cdot \log_2(|\mathbf{F}|)$ bits for storing the vault pairs explicitly.

On the other hand, the following fact can be shown. Suppose that two related records of the compact fuzzy vault scheme can be intercepted:

$$V(X) = f(X) + \prod_{x \in \mathbf{A}} (X - x) \tag{6}$$

$$W(X) = g(X) + \prod_{x \in \mathbf{B}} (X - x), \tag{7}$$

and these records are protecting the feature sets $\mathbf{A}, \mathbf{B}$ bound to the polynomials $f, g \in \mathbf{F}[X]$ both of degree smaller than k, respectively. Here, without loss of generality, we assume that $|\mathbf{A}| \geq |\mathbf{B}|$ and $|\mathbf{A} \cap \mathbf{B}| \geq (|\mathbf{A}| + k)/2$ is fulfilled. Then the differences $\mathbf{A} \setminus \mathbf{B}$ and $\mathbf{B} \setminus \mathbf{A}$ can be recovered explicitly and efficiently by applying the extended Euclidean algorithm to $V(X)$ and $W(X)$. We refer to [26] for a proof of this fact. This would again conflict with the unlinkability requirement of effective biometric template protection calling for a countermeasure. Fortunately, by passing the feature elements through a record-specific random but public permutation $\mathbf{F} \rightarrow \mathbf{F}$ is a promising solution for preventing the extended Euclidean algorithm-based record multiplicity attack [26].

A record-specific random bit permutation process was considered to be incorporated in a fuzzy commitment scheme, in order to prevent the decodability attack [23]. In view of the fact that this measure has been shown to be forgeable [25], it would be highly desirable to prove or disprove the validity of the countermeasure in a reductionist sense, say. Yet, there is currently no attack known that can break two related records of the compact fuzzy vault scheme being subjected to a record-specific permutation process significantly better than breaking one of the records individually.

4.6 The Future of Biometric "Hashes"

The major issue in providing information security for biometric templates may lay in the design of implementations. In particular, for specific biometric modalities suitable feature extractions have to be developed that can decrease the most limiting factor *false acceptance rate* at a maintained and preferably high genuine acceptance

rate. However, even if the false acceptance rate can be reduced to its half for a certain biometric modality which would be a breakthrough, the security only increases by a single bit. Even though reducing false acceptance rates is certainly worth its trouble, it seems more reasonable to rely on the fusion of multiple biometric systems to achieve an acceptable amount of security. First steps have already been made [18], but they leave space for improvement.

References

1. Juels, A., Wattenberg, M.: A fuzzy commitment scheme. In: Proceedings of ACM Conference on Computer and Communications Security, 1999, pp. 28–36
2. Juels, A., Sudan, M.: A fuzzy vault scheme. In: Lapidoth, A. Teletar, E. (eds.) Proceedings of International Symposium on Information Theory, 2002, p. 408
3. Juels, A., Sudan, M.: A fuzzy vault scheme. Des. Codes Cryptogr. **38**(2), 237–257 (2006)
4. Berlekamp, E.R.: Algebraic Coding Theory. Aegean Park Press, Laguna Hills, CA (1984)
5. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of Fingerprint Recognition, 2nd edn. Springer, New York (2009)
6. Arakala, A., Jeffers, J., Horadam, K.: Fuzzy extractors for minutiae-based fingerprint authentication. In: Proceedings of International Conference on Biometrics. Lecture Notes on Computer Science, vol. 4642, pp. 760–769 (2007)
7. Hao, F., Anderson, R., Daugman, J.: Combining crypto with biometrics effectively. IEEE Trans. Comput. **55**(9), 1081–1088 (2006)
8. Clancy, T.C., Kiyavash, N., Lin, D.J.: Secure smartcard-based fingerprint authentication. In: Proceedings of ACM SIGMM Workshop on Biometrics Methods and Applications, WBMA '03, New York, NY, 2003, pp. 45–52
9. Yang, S., Verbaudwhede, I.: Automatic secure fingerprint verification system based on fuzzy vault scheme. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2005, pp. 609–612
10. Uludag, U., Pankanti, S., Jain, A.K.: Fuzzy vault for fingerprints. In: Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication, 2005, pp. 310–319
11. Uludag, U., Jain, A.K.: Securing fingerprint template: fuzzy vault with helper data. In: Proceedings of Workshop on Privacy Research in Vision, 2006, pp. 163–169
12. Nandakumar, K., Jain, A.K., Pankanti, S.: Fingerprint-based fuzzy vault: Implementation and performance. IEEE Trans. Inf. Forensics Secur. **2**(4), 744–757 (2007)
13. Nandakumar, K., Nagar, A., Jain, A.: Hardening fingerprint fuzzy vault using password. In: Proceedings of International Conference on Biometrics. Lecture Notes in Computer Science, vol. 4642, pp. 927–937 (2007)
14. Nagar, A., Nandakumar, K., Jain, A.K.: A hybrid biometric cryptosystem for securing fingerprint minutiae templates. Pattern Recogn. Lett. **31**, 733–741 (2010)
15. Li, P., Yang, X., Cao, K., Tao, X., Wang, R., Tian, J.: An alignment-free fingerprint cryptosystem based on fuzzy vault scheme. J. Netw. Comput. Appl. **33**, 207–220 (2010)
16. Dodis, Y., Ostrovsky, R., Reyzin, L., Smith, A.: Fuzzy extractors: how to generate strong keys from biometrics and other noisy data. SIAM J. Comput. **38**(1), 97–139 (2008)
17. Merkle, J., Ihmor, H., Korte, U., Niesing, M., Schwaiger, M.: Performance of the fuzzy vault for multiple fingerprints (extended version). CoRR, vol. abs/1008.0807v5 (2011)
18. Nagar, A., Nandakumar, K., Jain, A.K.: Multibiometric cryptosystems based on feature-level fusion. IEEE Trans. Inf. Forensics Secur. **7**(1), 255–268 (2012)
19. ISO/IEC JTC1 SC2 Security Techniques. ISO/IEC 24745:2011. Information Technology - Security Techniques - Biometric Information Protection. International Organization for Standardization (2011)

20. Scheirer, W.J., Boult, T.E.: Cracking fuzzy vaults and biometric encryption. In: Proceedings of Biometrics Symposium, 2007, pp. 1–6
21. Kholmatov, A., Yanikoglu, B.: Realization of correlation attack against the fuzzy vault scheme. In: Proceedings of SPIE, vol. 6819 (2008)
22. Simoens, K., Tuyls, P., Preneel, B.: Privacy weaknesses in biometric sketches. In: Proceedings of the 2009 30th IEEE Symposium on Security and Privacy (SP'09), pp. 188–203, Washington, DC. IEEE Computer Society, Silver Spring (2009)
23. Kelkboom, E.J.C., Breebaart, J., Kevenaar, T.A.M., Buhan, I., Veldhuis, R.N.: Preventing the decodability attack based cross-matching in a fuzzy commitment scheme. IEEE Trans. Inf. Forensics Secur. **6**(1), 107–121 (2011)
24. Tams, B.: Decodability attack against the fuzzy commitment scheme with public feature transforms. CoRR, vol. abs/1406.1154 (2014)
25. Tams, B.: Absolute fingerprint pre-alignment in minutiae-based cryptosystems. In: Proceedings of BIOSIG, 2013, pp. 75–86
26. Merkle, J., Tams, B.: Security of the improved fuzzy vault scheme in the presence of record multiplicity (full version). http://arxiv.org/pdf/1312.5225.pdf (2014). Also, in CoRR, vol. abs/1312.5225 (in review)

Generalizations of Entropy and Information Measures

Thomas L. Toulias and Christos P. Kitsos

Abstract This paper presents and discusses two generalized forms of the Shannon entropy, as well as a generalized information measure. These measures are applied on a exponential-power generalization of the usual Normal distribution, emerged from a generalized form of the Fisher's entropy type information measure, essential to Cryptology. Information divergences between these random variables are also discussed. Moreover, a complexity measure, related to the generalized Shannon entropy, is also presented, extending the known SDL complexity measure.

Keywords: Fisher's entropy type information measure • Shannon entropy • Rényi entropy • Generalized normal distribution • SDL complexity

1 Introduction

Since the time of Clausius, 1865, Entropy plays an important role in linking physical experimentation and statistical analysis. It was later in 1922, when Fisher developed in [9] the Experiment Design Theory, another link between Statistics with Chemistry, as well as in other fields. For the principle of maximum entropy, the normal distribution is essential and eventually it is related with the energy and the variance involved.

The pioneering work by Shannon [28] related Entropy with Information Theory and gave a new perspective to the study of information systems and of Cryptography, see [1, 14] among others. *Shannon entropy* (or *entropy*) measures the average uncertainty of a random variable (r.v.). In Information Theory, it is the minimum number of bits required, on the average, to describe the value x of the r.v. X. In Cryptography, entropy gives the ultimately achievable error-free compression in terms of the average codeword length symbol per source. There are two different roles of entropy measures: (a) positive results can be obtained in the form of security

T.L. Toulias (✉) • C.P. Kitsos
Technological Educational Institute of Athens, Egaleo 12243, Athens, Greece
e-mail: th.toulias@gmail.com; xkitsos@teiath.gr

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_22

proofs for (unconditionally secure) cryptographic systems, and (b) lower bounds on the required key sizes are negative results, in some scenarios, and follow from entropy-based arguments. See also [14].

Recall that the *relative entropy* or *discrimination* or *information divergence* between two r.v., say X and Y, measures the increase, or decrease, of information, about an experiment, when the probability $\Pr(X)$ (associated with the knowledge of the experiment) is changed to $\Pr(Y)$. Relative entropy is the underlying idea of the Authentication Theory which provides a level of assurance to the receiver of a message originating from a legitimate sender.

A central concept of Cryptography is that of *information measure* or *information*, as cryptographic scenarios can be modelled with information-theoretic methods. There are several kinds of information measures which all quantify the uncertainty of an outcome of a random experiment, and, in principle, information is a measure of the reduction of uncertainty.

Fisher's entropy type information measure is a fundamental one, see [5]. Poincaré and Sobolev Inequalities play an important role in the foundation of the generalized Fisher's entropy type information measure. Both classes of inequalities offer a number of bounds for a number of physical applications. The Gaussian kernel or the error function (which produce the normal distribution) usually has two parameters, the mean and the variance. For the Gaussian kernel an extra parameter was then introduced in [15], and therefore a generalized form of the Normal distribution was obtained. Specifically, the generalized Gaussian is obtained as an extremal for the Logarithm Sobolev Inequality (LSI), see [4, 30], and is referred here as the γ-order Normal Distribution, or $\mathcal{N}_\gamma$. In addition, the Poincaré Inequality (PI), offers also the "best" constant for the Gaussian measure, and therefore is of interest to see how Poincaré and Sobolev inequalities are acting on the Normal distribution.

In this paper we introduce and discuss two generalized forms of entropy and their behavior over the generalized Normal distribution. Moreover, the specific entropy measures as collision and the mean-entropy are discussed. A complexity measure for an r.v. is also evaluated and studied.

2 Information Measures and Generalizations

Let X be a multivariate r.v. with parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p) \in \mathbb{R}^p$ and p.d.f. $f_X = f_X(x;\ \theta)$, $x \in \mathbb{R}^p$. The parametric type Fisher's Information Matrix $\mathrm{I}_{\mathrm{F}}(X;\ \theta)$ (also denoted as $\mathrm{I}_\theta(X)$) defined as the covariance of $\nabla_\theta \log f_X(X;\ \theta)$ (where ∇_θ is the gradient with respect to the parameters θ_i, $i = 1, 2, \ldots, p$) is a parametric type information measure, expressed also as

$$\begin{aligned} \mathrm{I}_\theta(X) &= \mathrm{Cov}\left(\nabla_\theta \log f_X(X;\ \theta)\right) = \mathrm{E}_\theta\left[\nabla_\theta \log f_X \cdot (\nabla_\theta \log f_X)^{\mathrm{T}}\right] \\ &= \mathrm{E}_\theta\left[\|\nabla_\theta \log f_X\|^2\right], \end{aligned}$$

where $\|\cdot\|$ is the usual $\mathscr{L}^2(\mathbb{R}^p)$ norm, while $\mathrm{E}_\theta[\cdot]$ denotes the expected value operator applied to random variables, with respect to parameter θ.

Recall that the Fisher's entropy type information measure $\mathrm{I_F}(X)$, or $\mathrm{J}(X)$, of an r.v. X with p.d.f. f on $\mathbb{R}^p$, is defined as the covariance of r.v. $\nabla \log f(X)$, i.e. $\mathrm{J}(X) := \mathrm{E}[\|\nabla \log f(X)\|^2]$, with $\mathrm{E}[\cdot]$ now denotes the usual expected value operator of a random variable with respect to the its p.d.f. Hence, $\mathrm{J}(X)$ can be written as

$$\begin{aligned} \mathrm{J}(X) &= \int\limits_{\mathbb{R}^p} f(x)\|\nabla \log f(x)\|^2 dx = \int\limits_{\mathbb{R}^p} f(x)^{-1}\|\nabla f(x)\|^2 dx \\ &= \int\limits_{\mathbb{R}^p} \nabla f(x) \cdot \nabla \log f(x) dx = 4 \int\limits_{\mathbb{R}^p} \left\| \nabla \sqrt{f(x)} \right\|^2 dx. \end{aligned} \tag{1}$$

Generally, the family of the entropy type information measures $\mathrm{I}(X)$, of a p-variate r.v. X with p.d.f. f, are defined through the score function of X, i.e.

$$U(X) := \|\nabla \log f(X)\|,$$

as

$$\mathrm{I}(X) := \mathrm{I}(X;\ g, h) := g\left(\mathrm{E}[h(U(X))]\right),$$

where g and h being real-valued functions. For example, when g = i.d. and $h(X) = X^2$ we obtain the entropy type Fisher's information measure of X as in (1), i.e.

$$\mathrm{I_F}(X) = \mathrm{E}[\|\nabla \log f(X)\|^2]. \tag{2}$$

Besides $\mathrm{I_F}$, other entropy type information measures as the Vajda's, Mathai's, and Boeke's information measures, denoted with $\mathrm{I_V}$, $\mathrm{I_M}$, and $\mathrm{I_B}$, respectively, are defined as:

$$\begin{array}{llll} \mathrm{I_F}(X) := \mathrm{I}(X), & \text{with} & g := \mathrm{id}. & \text{and } h(U) := U^2, \\ \mathrm{I_V}(X) := \mathrm{I}(X), & \text{with} & g := \mathrm{id}. & \text{and } h(U) := U^\lambda,\ \lambda \geq 1, \\ \mathrm{I_M}(X) := \mathrm{I}(X), & \text{with} & g(X) := X^{1/\lambda} & \text{and } h(U) := U^\lambda,\ \lambda \geq 1, \\ \mathrm{I_B}(X) := \mathrm{I}(X), & \text{with} & g(X) := X^{\lambda-1} & \text{and } h(U) := U^{\frac{\lambda}{\lambda-1}},\ \lambda \in \mathbb{R}_+ \setminus 1. \end{array}$$

The notion of information "distance" or divergence of a p-variate r.v. X over a p-variate r.v. Y is given by

$$\mathrm{D}(X, Y) = \mathrm{D}(X, Y;\ g, h) := g\left(\int\limits_{\mathbb{R}^p} h(f_X, f_Y) \right),$$

where f_X and f_Y are the probability density functions (p.d.f) of X and Y, respectively. Some known divergences, such as the Kullback–Leibler $\mathrm{D_{KL}}$, the Vajda's $\mathrm{D_V}$, the

Kagan $\mathrm{D_K}$, the Csiszar $\mathrm{D_C}$, the Matusita $\mathrm{D_M}$, as well as the Rényi's $\mathrm{D_R}$ divergence, see also [8], are defined as follows:

$$
\begin{aligned}
&\mathrm{D_{KL}}(X,Y) := \mathrm{D}(X,Y), \text{ with } g := \mathrm{id.} && \text{and } h(f_X,f_Y) := f_X \log(f_X/f_Y),\\
&\mathrm{D_V}(X,Y) := \mathrm{D}(X,Y), \text{ with } g := \mathrm{id.} && \text{and } h(f_X,f_Y) := f_X\left|1-(f_Y/f_X)\right|^{\lambda},\ \lambda \geq 1,\\
&\mathrm{D_K}(X,Y) := \mathrm{D}(X,Y), \text{ with } g := \mathrm{id.} && \text{and } h(f_X,f_Y) := f_X\left|1-(f_Y/f_X)\right|^{2},\\
&\mathrm{D_C}(X,Y) := \mathrm{D}(X,Y), \text{ with } g := \mathrm{id.} && \text{and } h(f_X,f_Y) := f_Y\phi(f_X/f_Y),\ \phi \text{ convex},\\
&\mathrm{D_M}(X,Y) := \mathrm{D}(X,Y), \text{ with } g(A) := \sqrt{A} && \text{and } h(f_X,f_Y) := \left(\sqrt{f_X}-\sqrt{f_Y}\right)^2,\\
&\mathrm{D_R}(X,Y) := \mathrm{D}(X,Y), \text{ with } g(A) := \tfrac{\log A}{1-\lambda} && \text{and } h(f_X,f_Y) := f_X^{\lambda}f_Y^{1-\lambda},\ \lambda \in \mathbb{R}_+ \setminus 1.
\end{aligned}
$$

Consider now the Vajda's parametric type measure of information $\mathrm{I_V}(X;\ \theta,\alpha)$, which is in fact a generalization of $\mathrm{I_F}(X;\ \theta)$, defined as, [8, 33],

$$\mathrm{I}_V(X;\ \theta,\alpha) := \mathrm{E}_\theta[\|\nabla_\theta \log f(X)\|^\alpha], \quad \alpha \geq 1. \tag{3}$$

Similarly, the Vajda's entropy type information measure $\mathrm{J}_\alpha(X)$ generalizes Fisher's entropy type information $\mathrm{J}(X)$, defined as

$$\mathrm{J}_\alpha(X) := \mathrm{E}[\|\nabla \log f(X)\|^\alpha], \quad \alpha \geq 1, \tag{4}$$

see [15]. We shall refer to $\mathrm{J}_\alpha(X)$ as the generalized Fisher's entropy type information measure or α-GFI. The second-GFI is reduced to the usual J, i.e. $\mathrm{J}_2(X) = \mathrm{J}(X)$. Equivalently, from the definition of the α-GFI above we can obtain

$$
\begin{aligned}
\mathrm{J}_\alpha(X) &= \int\limits_{\mathbb{R}^p} \|\nabla \log f(x)\|^\alpha f(x)dx = \int\limits_{\mathbb{R}^p} \|\nabla f(x)\|^\alpha f^{1-\alpha}(x)dx\\
&= \alpha^\alpha \int\limits_{\mathbb{R}^p} \|\nabla f^{1/\alpha}(x)\|^\alpha dx.
\end{aligned} \tag{5}
$$

The Blachman–Stam inequality [2, 3, 31] still holds through the α-GFI measure J_α, see [15] for a complete proof. Indeed:

Theorem 1. *For two given p-variate and independent random variables X and Y, it holds*

$$\mathrm{J}_\alpha\left(\lambda^{1/\alpha}X + (1-\lambda)^{1/\alpha}Y\right) \leq \lambda \mathrm{J}_\alpha(X) + (1-\lambda)\mathrm{J}_\alpha(Y), \quad \lambda \in (0,1). \tag{6}$$

The equality holds when X and Y are normally distributed with the same covariance matrix.

As far as the superadditivity of J_α is concerned, the following Theorem it can be stated, see [19] for a complete proof.

Theorem 2. *Let an orthogonal decomposition $\mathbb{R}^p = \mathbb{R}^r \oplus \mathbb{R}^s$, $p = s + t$, with the corresponding marginal densities of a p.d.f. f on $\mathbb{R}^p$ being f_1 on $\mathbb{R}^s$ and f_2 on $\mathbb{R}^t$, i.e.*

$$f_1(x) = \int_{\mathbb{R}^s} f(x,y)d^s y, \quad f_2(y) = \int_{\mathbb{R}^t} f(x,y)d^t x, \tag{7}$$

Then, for r.v. X, X_1 and X_2 following f, f_1 and f_2, it holds

$$\mathrm{J}_\alpha(X) \geq \mathrm{J}_\alpha(X_1) + \mathrm{J}_\alpha(X_2), \tag{8}$$

with the equality holding when $f(x,y) = f_1(x)f_2(y)$ almost everywhere.

The Shannon entropy $\mathrm{H}(X)$ of a continuous r.v. X with p.d.f.f is defined as, [5],

$$\mathrm{H}(X) := \mathrm{E}[\log f(X)] = \int_{\mathbb{R}^p} f(x)\log f(x)dx, \tag{9}$$

(we drop the usual minus sign) and its corresponding entropy power $\mathrm{N}(X)$ is defined as

$$\mathrm{N}(X) := \nu e^{\frac{2}{p}\mathrm{H}(X)}, \tag{10}$$

with $\nu := (2\pi e)^{-1}$. The generalized entropy power $\mathrm{N}_\alpha(X)$, introduced in [15], is of the form

$$\mathrm{N}_\alpha(X) := \nu_\alpha e^{\frac{\alpha}{p}\mathrm{H}(X)}, \tag{11}$$

with normalizing factor ν_α given by the appropriate generalization of ν, namely

$$\nu_\alpha := \left(\tfrac{\alpha-1}{\alpha e}\right)^{\alpha-1} \pi^{-\frac{\alpha}{2}} \left[\frac{\Gamma\left(\frac{p}{2}+1\right)}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\frac{\alpha}{p}}, \quad \alpha \in \mathbb{R}\setminus[0,1]. \tag{12}$$

For the parameter case of $\alpha = 2$, (11) is reduced to the known entropy power $\mathrm{N}(X)$, i.e. $\mathrm{N}_2(X) = \mathrm{N}(X)$ and $\nu_2 = \nu$.

The known information inequality $\mathrm{J}(X)\mathrm{N}(X) \geq p$ still holds under the generalized entropy type Fisher's information, as $\mathrm{J}_\alpha(X)\mathrm{N}_\alpha(X) \geq p$, $\alpha > 1$, see [15]. As a result the Cramér–Rao inequality, $\mathrm{J}(X)\,\mathrm{Var}(X) \geq p$, can be extended to

$$\left[\tfrac{2\pi e}{p}\,\mathrm{Var}(X)\right]^{1/2} \left[\tfrac{\nu_\alpha}{p}\mathrm{J}_\alpha(X)\right]^{1/\alpha} \geq 1, \quad \alpha > 1, \tag{13}$$

see [15]. Under the normality parameter $\alpha = 2$, (13) is reduced to the usual Cramér–Rao inequality.

Furthermore, the classical entropy inequality

$$\mathrm{Var}(X) \geq p\mathrm{N}(X) = \tfrac{p}{2\pi e} e^{\frac{2}{p}\mathrm{H}(X)}, \tag{14}$$

can be extended, adopting our extension above, to the general form

$$\mathrm{Var}(X) \geq p(2\pi e)^{\frac{\alpha-4}{\alpha}} \nu_\alpha^{2/\alpha} \mathrm{N}_\alpha^{2/\alpha}(X), \quad \alpha > 1. \tag{15}$$

Under the "normal" parameter value $\alpha = 2$, the inequality (15) is reduced to the usual entropy inequality as in (14).

Through the generalized entropy power N_α a generalized form of the usual Shannon entropy can be produced. Indeed, consider the Shannon entropy of which the corresponding entropy power is N_α (instead of the usual N), i.e.

$$\mathrm{N}_\alpha(X) = \nu \exp\{\tfrac{2}{p}\mathrm{H}_\alpha(X)\}, \quad \alpha \in \mathbb{R} \setminus [0,\, 1]. \tag{16}$$

We shall refer to the quantity H_α as the *generalized Shannon entropy*, or *α-Shannon entropy*, see for details [17]. Therefore, from (11) a linear relation between the generalized Shannon entropy $\mathrm{H}_\alpha(X)$ and the usual Shannon entropy $\mathrm{H}(X)$ is obtained, i.e.

$$\mathrm{H}_\alpha(X) = \tfrac{p}{2} \log \tfrac{\nu_\alpha}{\nu} + \tfrac{\alpha}{2}\mathrm{H}(X), \quad \alpha \in \mathbb{R} \setminus [0,\, 1]. \tag{17}$$

Essentially, (17) represents a linear transformation of $\mathrm{H}(X)$ which depends on the parameter α and the dimension $p \in \mathbb{N}$. It is also clear that the generalized Shannon entropy with $\alpha = 2$ is the usual Shannon entropy, i.e. $\mathrm{H}_2 = \mathrm{H}$.

3 Entropy, Information, and the Generalized Gaussian

For a p-variate random vector X the following known Proposition bounds the Shannon entropy using only the covariance matrix of X.

Proposition 1. *Let the random vector X have zero mean and covariance matrix Σ. Then*

$$\mathrm{H}(X) \leq \tfrac{1}{2} \log \left\{(2\pi e)^p |\det \Sigma|\right\},$$

with equality holding if and only if $X \sim \mathcal{N}(0, \Sigma)$.

This Proposition is crucial and denotes that the entropy for the Normal distribution is depending, eventually, only on the variance–covariance matrix, while equality holds when X is following the (multivariate) normal distribution, a result widely applied in engineering problems and information systems.

A construction of an exponential-power generalization of the usual Normal distribution can be obtained as an extremal of (an Euclidean) LSI. Following [15], the Gross Logarithm Inequality with respect to the Gaussian weight, [13], is of the form

$$\int_{\mathbb{R}^p} \|g\|^2 \log \|g\|^2 dm \leq \tfrac{1}{\pi} \int_{\mathbb{R}^p} \|\nabla g\|^2 dm, \tag{18}$$

where $\|g\|_2 := \int_{\mathbb{R}^p} \|g(x)\|^2 dx = 1$ is the norm in $\mathscr{L}^2(\mathbb{R}^p, dm)$ with $dm := \exp\{-\pi|x|^2\}dx$. Inequality (18) is equivalent to the (Euclidean) LSI,

$$\int_{\mathbb{R}^p} \|u\|^2 \log \|u\|^2 dx \leq \tfrac{p}{2} \log \left\{ \tfrac{2}{\pi p e} \int_{\mathbb{R}^p} \|\nabla u\|^2 dx \right\}, \tag{19}$$

for any function $u \in \mathscr{W}^{1,2}(\mathbb{R}^p)$ with $\|u\|_2 = 1$, see [15] for details. This inequality is optimal, in the sense that

$$\tfrac{2}{\pi p e} = \inf \left\{ \frac{\int_{\mathbb{R}^p} \|\nabla u\|^2 dx}{\exp\left(\tfrac{2}{n} \int_{\mathbb{R}^p} \|u\|^2 \log \|u\|^2 dx \right)} : \quad u \in \mathscr{W}^{1,2}(\mathbb{R}^p), \quad \|u\|_2 = 1 \right\},$$

see [34]. Extremals for (19) are precisely the Gaussians

$$u(x) = (\pi\sigma/2)^{-p/4} \exp\left\{ -\left| \frac{x-\mu}{\sigma} \right|^2 \right\},$$

with $\sigma > 0$ and $\mu \in \mathbb{R}^p$, see [3, 4] for details.

Now, consider the extension of Del Pino and Dolbeault in [6] for the LSI as in (19). For any $u \in \mathscr{W}^{1,2}(\mathbb{R}^p)$ with $\|u\|_\gamma = 1$, the γ-LSI holds, i.e.

$$\int_{\mathbb{R}^p} \|u\|^\gamma \log \|u\| dx \leq \tfrac{p}{\gamma^2} \log \left\{ K_\gamma \int_{\mathbb{R}^p} \|\nabla u\|^\gamma dx \right\}, \tag{20}$$

with the optimal constant K_γ being equal to

$$K_\gamma = \tfrac{\gamma}{p} (\tfrac{\gamma-1}{e})^{\gamma-1} \pi^{-\gamma/2} \left[\frac{\Gamma(\frac{p}{2}+1)}{\Gamma(p\frac{\gamma-1}{\gamma}+1)} \right]^{\gamma/p}, \tag{21}$$

where $\Gamma(\cdot)$ is the usual gamma function.

Inequality (20) is optimal and the equality holds when $u(x) := f_X(x)$, $x \in \mathbb{R}^p$ where X is an r.v. following the multivariate distribution with p.d.f. f_X defined as

$$f_X(x) = f_X(x;\ \mu, \Sigma, \gamma) := C_\gamma^p(\Sigma) \exp\left\{-\tfrac{\gamma-1}{\gamma} Q_\theta(x)^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad x \in \mathbb{R}^p, \tag{22}$$

with normalizing factor

$$C_\gamma^p(\Sigma) := \frac{(\frac{\gamma-1}{\gamma})^{p\frac{\gamma-1}{\gamma}}}{\pi^{p/2}\sqrt{|\det \Sigma|}} \left[\frac{\Gamma(\frac{p}{2}+1)}{\Gamma(p\frac{\gamma-1}{\gamma}+1)}\right] = \max f_X, \tag{23}$$

and p-quadratic form $Q_\theta(x) := (x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}$, $x \in \mathbb{R}^p$ where $\theta := (\mu, \Sigma) \in \mathbb{R}^{p\times(p\times p)}$. The function $\phi(\gamma) = f_{X_\gamma}(x)^{1/\gamma}$ with $\Sigma = (\sigma^2/\alpha)^{2(\gamma-1)/\gamma}\mathbb{I}_p$ corresponds to the extremal function for the LSI due to [6]. The essential result is that the defined p.d.f f_X works as an extremal function to a generalized form of the Logarithmic Sobolev Inequality.

We shall write $X_\gamma \sim \mathscr{N}_\gamma^p(\mu, \Sigma)$ where $\mathscr{N}_\gamma^p(\mu, \Sigma)$ is an exponential-power generalization of the usual p-variate Normal distribution $\mathscr{N}^p(\mu, \Sigma)$ with location parameter vector $\mu \in \mathbb{R}^{1\times p}$ and positive definite scale matrix $\Sigma \in \mathbb{R}^{p\times p}$, involving a new shape parameter $\gamma \in \mathbb{R} \setminus [0, 1]$. These distributions shall be referred to as the γ-order Normal distributions. It can be easily seen that the parameter vector μ is, indeed, the mean vector of the $\mathscr{N}_\gamma^p$ distribution, i.e. $\mu = \mathrm{E}[X_\gamma]$ for all parameters $\gamma \in \mathbb{R} \setminus [0, 1]$, see [20]. Notice also that for $\gamma = 2$ the second-ordered Normal $\mathscr{N}_2^p(\mu, \Sigma)$ is reduced to the usual multivariate Normal $\mathscr{N}^p(\mu, \Sigma)$, i.e. $\mathscr{N}_2^p(\mu, \Sigma) = \mathscr{N}^p(\mu, \Sigma)$. One of the merits of the γ-order Normal distribution defined above belongs to the symmetric Kotz type distributions family, [21], as $\mathscr{N}_\gamma^p(\mu, \Sigma) = \mathscr{K}_{m,r,s}(\mu, \Sigma)$ with $m := 1$, $r := (\gamma-1)/\gamma$ and $s := \gamma/(2\gamma-2)$.

It is worth noting that the introduced univariate γ-order Normal $\mathscr{N}_\gamma(\mu, \sigma^2) := \mathscr{N}_\gamma^1(\mu, \sigma^2)$ coincides with the existent generalized normal distribution introduced in [23], with density function

$$f(x) = f(x;\ \mu, a, b) := \frac{b}{2a\Gamma(1/b)} \exp\left\{-\left|\tfrac{x-\mu}{a}\right|^b\right\}, \quad x \in \mathbb{R},$$

where $a = \sigma[\gamma/(\gamma-1)]^{(\gamma-1)/\gamma}$ and $b = \gamma/(\gamma-1)$, while the multivariate case of the γ-order Normal $\mathscr{N}_\gamma^p(\mu, \Sigma)$ coincides with the existent multivariate power exponential distribution $\mathscr{PE}^p(\mu, \Sigma', b)$, as introduced in [10], where $\Sigma' = 2^{2(\gamma-1)/\gamma}\Sigma$ and $b := \frac{1}{2}\gamma/(\gamma-1)$. See also [11, 22]. These existent generalizations are technically obtained (involving an extra power parameter b) and there are not resulting from a strong mathematical background, as the Logarithmic Sobolev Inequalities offer. Moreover, they cannot provide application to the generalized Fisher Information or entropy power, etc. as their form does not really contribute to technical proofs we have already provided, see [15, 18, 20].

Denote with $\mathbb{E}_\theta$ the area of the p-ellipsoid $Q_\theta(x) \leq 1$, $x \in \mathbb{R}^p$. The family of $\mathcal{N}_\gamma^p(\mu, \Sigma)$, i.e. the family of the elliptically contoured γ-order Normals, provides a smooth bridging between the multivariate (and elliptically countered) Uniform, Normal and Laplace r.v. U, Z and L, i.e. between $U \sim \mathcal{U}^p(\mu, \Sigma)$, $Z \sim \mathcal{N}^p(\mu, \Sigma)$ and Laplace $L \sim \mathcal{L}^p(\mu, \Sigma)$ r.v. as well as the multivariate degenerate Dirac distributed r.v. $D \sim \mathcal{D}^p(\mu)$ (with pole at the point μ), with density functions

$$f_U(x) = f_U(x;\ \mu, \Sigma) := \begin{cases} \dfrac{\Gamma(\frac{p}{2}+1)}{\pi^{p/2}\sqrt{|\det \Sigma|}}, & x \in \mathbb{E}_\theta, \\ 0, & x \notin \mathbb{E}_\theta, \end{cases} \tag{24}$$

$$f_Z(x) = f_Z(x;\ \mu, \Sigma) := \frac{1}{(2\pi)^{p/2}\sqrt{|\det \Sigma|}} \exp\left\{-\tfrac{1}{2}Q_\theta(x)\right\}, \quad x \in \mathbb{R}^p, \tag{25}$$

$$f_L(x) = f_L(x;\ \mu, \Sigma) := \frac{\Gamma(\frac{p}{2}+1)}{p!\pi^{p/2}\sqrt{|\det \Sigma|}} \exp\left\{-\sqrt{Q_\theta(x)}\right\}, \quad x \in \mathbb{R}^p, \tag{26}$$

$$f_D(x) = f_D(x;\ \mu) := \begin{cases} +\infty, & x = \mu, \\ 0, & x \in \mathbb{R}^p \setminus \mu, \end{cases} \tag{27}$$

respectively, see [20]. That is, the $\mathcal{N}_\gamma^p$ family of distributions generalizes not only the usual Normal but also two other significant distributions, as the Uniform and Laplace ones. The above discussion is summarized in the following Theorem, [20].

Theorem 3. *The elliptically contoured p-variate γ-order Normal distribution $\mathcal{N}_\gamma^p(\mu, \Sigma)$ for order values of $\gamma = 0, 1, 2, \pm\infty$ coincides with*

$$\mathcal{N}_\gamma^p(\mu, \Sigma) = \begin{cases} \mathcal{D}^p(\mu), & \text{for } \gamma = 0 \text{ and } p = 1, 2, \\ 0, & \text{for } \gamma = 0 \text{ and } p \geq 3, \\ \mathcal{U}^p(\mu, \Sigma), & \text{for } \gamma = 1, \\ \mathcal{N}^p(\mu, \Sigma), & \text{for } \gamma = 2, \\ \mathcal{L}^p(\mu, \Sigma), & \text{for } \gamma = \pm\infty. \end{cases} \tag{28}$$

Remark 1. Considering the above Theorem, the definition values of the shape parameter γ of $\mathcal{N}_\gamma^p$ distributions can be extended to include the limiting extra values of $\gamma = 0, 1, \pm\infty$, respectively, i.e. γ can now be considered as a real number outside the open interval $(0, 1)$. Particularly, when $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \Sigma)$, $\gamma \in \mathbb{R}\setminus(0, 1)\cup\{\pm\infty\}$, the r.v. $X_0, X_1 \sim \mathcal{U}^p(\mu, \Sigma)$ and $X_{\pm\infty} \sim \mathcal{L}^p(\mu, \Sigma)$ can be defined as

$$X_0 := \lim_{\gamma \to 0^-} X_\gamma, \quad X_1 := \lim_{\gamma \to 1^+} X_\gamma, \quad X_{\pm\infty} := \lim_{\gamma \to \pm\infty} X_\gamma. \tag{29}$$

Eventually, the Uniform, Normal, Laplace and also the degenerate distribution $\mathcal{N}_0^p$ (like the Dirac one for dimensions $p = 1, 2$) can be considered as members of the "extended" $\mathcal{N}_\gamma^p$, $\gamma \in \mathbb{R}\setminus(0,\ 1)\cup\{\pm\infty\}$, family of generalized Normal distributions.

Notice also that $\mathscr{N}_1^1(\mu,\sigma)$ coincides with the known (continuous) Uniform distribution $\mathscr{U}(\mu-\sigma,\mu+\sigma)$. Specifically, for every Uniform distribution expressed with the usual notation $\mathscr{U}(a,b)$, it holds that $\mathscr{U}(a,b) = \mathscr{N}_1^1(\frac{a+b}{2},\frac{b-a}{2}) = \mathscr{U}^1(\mu,\sigma)$. Also $\mathscr{N}_2(\mu,\sigma^2) = \mathscr{N}(\mu,\sigma^2)$, $\mathscr{N}_{\pm\infty}(\mu,\sigma^2) = \mathscr{L}(\mu,\sigma)$ and finally $\mathscr{N}_0(\mu,\sigma) = \mathscr{D}(\mu)$. Therefore the following holds.

Corollary 1. *For order values $\gamma = 0, 1, 2, \pm\infty$, the univariate γ-ordered Normal distributions $\mathscr{N}_\gamma^1(\mu,\sigma^2)$ coincides with the usual (univariate) degenerate Dirac $\mathscr{D}(\mu)$, Uniform $\mathscr{U}(\mu-\sigma,\mu+\sigma)$, Normal $\mathscr{N}(\mu,\sigma^2)$, and Laplace $\mathscr{L}(\mu,\sigma)$ distributions, respectively.*

Recall now the cumulative distribution function (c.d.f.) $\Phi_Z(z)$ of the standardized normally distributed $Z\sim\mathscr{N}(0,1)$, i.e.

$$\Phi_Z(z) = \tfrac{1}{2} + \tfrac{1}{2}\operatorname{erf}(\tfrac{z}{2}), \quad z\in\mathbb{R}, \tag{30}$$

with $\operatorname{erf}(\cdot)$ being the usual error function. For the c.d.f. of the $\mathscr{N}_\gamma$ family of distributions the generalized error function $\operatorname{Erf}_{\gamma/(\gamma-1)}(\cdot)$ or the upper (or complementary) incomplete gamma function $\Gamma(\cdot,\cdot)$ is involved, [12]. Indeed, the following holds, [19].

Theorem 4. *Let X be a γ-order normally distributed r.v., i.e. $X\sim\mathscr{N}_\gamma(\mu,\sigma^2)$ with p.d.f. f_γ. If F_X is the c.d.f. of X and Φ_Z the c.d.f. of the standardized $Z=\frac{1}{\sigma}(X-\mu)\sim\mathscr{N}_\gamma(0,1)$, then*

$$F_X(x) = \Phi_Z(\tfrac{x-\mu}{\sigma}) = \tfrac{1}{2} + \frac{\sqrt{\pi}}{2\Gamma(\frac{\gamma-1}{\gamma})\Gamma(\frac{\gamma}{\gamma-1})}\operatorname{Erf}_{\frac{\gamma}{\gamma-1}}\left\{(\tfrac{\gamma-1}{\gamma})^{\frac{\gamma-1}{\gamma}}\frac{x-\mu}{\sigma}\right\} \tag{31}$$

$$= 1 - \frac{1}{2\Gamma(\frac{\gamma-1}{\gamma})}\Gamma\left(\tfrac{\gamma-1}{\gamma},\tfrac{\gamma-1}{\gamma}\left(\frac{x-\mu}{\sigma}\right)^{\frac{\gamma}{\gamma-1}}\right), \quad x\in\mathbb{R}. \tag{32}$$

3.1 Shannon Entropy and Generalization

Applying the Shannon entropy on a γ-order normally distributed random variable we state and prove the following.

Theorem 5. *The Shannon entropy of a random variable $X\sim\mathscr{N}_\gamma^p(\mu,\Sigma)$, with p.d.f. f_X, is of the form*

$$\mathrm{H}(X) = p\tfrac{\gamma-1}{\gamma} - \log C_\gamma^p(\Sigma) = p\tfrac{\gamma-1}{\gamma} - \log\max f_X. \tag{33}$$

Proof. From (22) and the definition (9) we obtain that the Shannon entropy of X is

$$\mathrm{H}(X) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\Sigma)^{\frac{\gamma-1}{\gamma}} \int\limits_{\mathbb{R}^p} Q_\theta(x)^{\frac{\gamma}{2(\gamma-1)}} \exp\left\{-\tfrac{\gamma-1}{\gamma} Q_\theta(x)^{\frac{\gamma}{2(\gamma-1)}}\right\} dx.$$

Applying the linear transformation $z := (x-\mu)^{\mathrm{T}}\Sigma^{-1/2}$ with $dx = d(x-\mu) = \sqrt{|\det \Sigma|}dz$, the $\mathrm{H}(X_\gamma)$ above is reduced to

$$\mathrm{H}(X) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)^{\frac{\gamma-1}{\gamma}} \int\limits_{\mathbb{R}^p} \|z\|^{\frac{\gamma}{\gamma-1}} \exp\left\{-\tfrac{\gamma-1}{\gamma}\|z\|^{\frac{\gamma}{\gamma-1}}\right\} dz,$$

where $\mathbb{I}_p$ denotes the $p \times p$ identity matrix. Switching to hyperspherical coordinates, we get

$$\mathrm{H}(X) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)^{\frac{\gamma-1}{\gamma}} \omega_{p-1} \int\limits_{\mathbb{R}_+} \rho^{\frac{\gamma}{\gamma-1}} \exp\left\{-\tfrac{\gamma-1}{\gamma}\rho^{\frac{\gamma}{\gamma-1}}\right\} \rho^{p-1} d\rho,$$

where $\omega_{p-1} := 2\pi^{p/2}/\Gamma\left(\frac{p}{2}\right)$ is the volume of the $(p-1)$-sphere. Applying the variable change $du := d(\frac{\gamma-1}{\gamma}\rho^{\gamma/(\gamma-1)}) = \rho^{1/(\gamma-1)}d\rho$ we obtain successively

$$\begin{aligned}
\mathrm{H}(X) &= -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)\omega_{p-1} \int\limits_{\mathbb{R}_+} u e^{-u} \rho^{\frac{(p-1)(\gamma-1)-1}{\gamma-1}} du \\
&= \log C_\gamma^p(\Sigma) - C_\gamma^p(\mathbb{I}_p)\omega_{p-1} \int\limits_{\mathbb{R}_+} u e^{-u} \left(\rho^{\frac{\gamma}{\gamma-1}}\right)^{\frac{(p-1)(\gamma-1)-1}{\gamma}} du \\
&= -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)\omega_{p-1}(\tfrac{\gamma}{\gamma-1})^{p\frac{\gamma-1}{\gamma}-1} \int\limits_{\mathbb{R}_+} u^{p\frac{\gamma-1}{\gamma}} e^{-u} du \\
&= -\log C_\gamma^p(\Sigma) + p\tfrac{\gamma-1}{\gamma}\Gamma(p\tfrac{\gamma-1}{\gamma}) C_\gamma^p(\mathbb{I}_p)\omega_{p-1}.
\end{aligned}$$

Finally, by substitution of the volume ω_{p-1} and the normalizing factor $C_\gamma^p(\Sigma)$ and $C_\gamma^p(\mathbb{I}_p)$, as in (23), relation (33) is obtained. □

We state and prove the following Theorem which provides the results for the Shannon entropy of the elliptically contoured family of the $\mathcal{N}_\gamma$ distributions.

Theorem 6. *The Shannon entropy for the multivariate and elliptically countered Uniform, Normal, and Laplace distributed X (for $\gamma = 1, 2, \pm\infty$, respectively), with p.d.f. f_X, as well as for the degenerate $\mathcal{N}_0$ distribution, is given by*

$$\mathrm{H}(X) = \begin{cases} -\log\max f_X &= \log\dfrac{\pi^{p/2}\sqrt{|\det\Sigma|}}{\Gamma\left(\frac{p}{2}+1\right)}, & \text{for } X\sim\mathscr{U}^p(\mu,\Sigma),\\ \frac{p}{2}-\log\max f_X &= \log\sqrt{(2\pi e)^p|\det\Sigma|}, & \text{for } X\sim\mathscr{N}^p(\mu,\Sigma),\\ p-\log\max f_X &= \log\dfrac{p!e\pi^{p/2}\sqrt{|\det\Sigma|}}{\Gamma\left(\frac{p}{2}+1\right)}, & \text{for } X\sim\mathscr{L}^p(\mu,\Sigma),\\ +\infty, & & \text{for } X\sim\mathscr{N}_0^p(\mu,\Sigma). \end{cases} \tag{34}$$

Proof. Applying Theorem 3 into (33) we obtain the first three branches of (34) for $\gamma = 1$ (in limit), $\gamma = 2$ (normality), and $\gamma = \pm\infty$ (in limit), respectively. Consider now the limiting case of $\gamma = 0$. We can write (33) in the form

$$\mathrm{H}(X) = \log\left\{\frac{\pi^{p/2}\sqrt{|\det\Sigma|}}{\Gamma(\frac{p}{2}+1)}\cdot\frac{\Gamma(pg+1)}{(\frac{g}{e})^{pg}}\right\},$$

where $g := \frac{\gamma-1}{\gamma}$. We then have,

$$\lim_{\gamma\to 0^-}\mathrm{H}(X) = \log\left\{\frac{\pi^{p/2}\sqrt{|\det\Sigma|}}{\Gamma(\frac{p}{2}+1)}\lim_{k=p\lfloor g\rfloor\to\infty}\frac{p^k k!}{(\frac{k}{e})^k}\right\}, \tag{35}$$

and using the Stirling's asymptotic formula $k! \approx \sqrt{2\pi k}(\frac{k}{e})^k$ as $k\to\infty$, (35) finally implies

$$\lim_{\gamma\to 0^-}\mathrm{H}(X) = \log\left\{\sqrt{2\pi|\det\Sigma|}\frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)}\lim_{k\to\infty}p^k\sqrt{k}\right\} = +\infty,$$

which proves the Theorem. □

Example 1. For the univariate case $p = 1$, we are reduced to

$$\mathrm{H}(X) = \begin{cases} -\log\max f_X &= \log 2\sigma, & \text{for } X\sim\mathscr{N}_1(\mu,\sigma) &= \mathscr{U}(\mu-\sigma,\mu+\sigma),\\ \frac{1}{2}-\log\max f_X &= \log\sqrt{2\pi e}\sigma, & \text{for } X\sim\mathscr{N}_2(\mu,\sigma^2) &= \mathscr{N}(\mu,\sigma^2),\\ 1-\log\max f_X &= 1+\log 2\sigma, & \text{for } X\sim\mathscr{N}_{\pm\infty}(\mu,\sigma) &= \mathscr{L}(\mu,\sigma),\\ +\infty, & & \text{for } X\sim\mathscr{N}_0(\mu,\sigma) &= \mathscr{D}(\mu). \end{cases}$$

Figure 1 below illustrates the univariate case of Theorem 5. The Shannon entropy $\mathrm{H}(X_\gamma)$, of an r.v. $X_\gamma\sim\mathscr{N}_\gamma(\mu,\sigma^2)$ is presented as a bivariate function of $\sigma\in(0,3]$ and $\gamma\in[-10,0)\cup[1,10]$, which forms the appeared surface (for arbitrary $\mu\in\mathbb{R}$). The Shannon entropy values of Uniform ($\gamma = 1$) and Normal ($\gamma = 2$) distributions are denoted (as curves), recall Example 1. Moreover, the entropy values of the r.v. $X_{\pm 10}\sim\mathscr{N}_{\pm 10}(\mu,\sigma^2)$, which approximates the Shannon entropy of Laplace distributed r.v. $X_{\pm\infty}\sim\mathscr{L}(\mu,\sigma)$, as well as the entropy of the r.v.

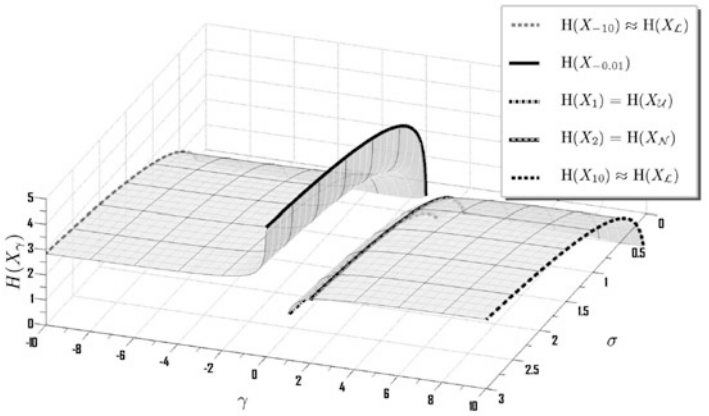


Fig. 1 Graph of all $\mathrm{H}(X_\gamma)$, $X_\gamma \sim \mathscr{N}_\gamma(\mu, \sigma^2)$ across the $\sigma(>0)$-semi-axis and γ-axis

$X_{-0.01} \sim \mathscr{N}_{-0.01}(\mu, \sigma^2)$ which approaches the degenerated Dirac r.v. $X_0 \sim \mathscr{D}(\mu)$, are also depicted. One can also notice the logarithmic increase of $\mathrm{H}(X_\gamma)$ as σ increases (for every fixed γ value), which holds due to the form of (33).

Due to the above proved Theorems, for the generalized Shannon entropy we obtain the following results.

Proposition 2. *The α-Shannon entropy H_α of the multivariate $X \sim \mathscr{N}_\gamma(\mu, \Sigma)$ is given by*

$$\mathrm{H}_\alpha(X) = \tfrac{2\gamma-\alpha}{2\gamma} p + \tfrac{p}{2} \log \left\{ 2\pi (\tfrac{\alpha-1}{\alpha})^{\alpha-1} (\tfrac{\gamma}{\gamma-1})^{\alpha \frac{\gamma-1}{\gamma}} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{\Gamma(p\frac{\alpha-1}{\alpha}+1)} \right]^{\frac{\alpha}{p}} |\det \Sigma|^{\frac{\alpha}{2p}} \right\}. \tag{36}$$

Moreover, in case of $\alpha = \gamma$, we have

$$\mathrm{H}_\gamma(X) = \tfrac{p}{2} \log \left\{ 2\pi e |\det \Sigma|^{\frac{\gamma}{2p}} \right\}. \tag{37}$$

Proof. Substituting (12) and (33) into (16) we obtain

$$\begin{aligned} \mathrm{H}_\alpha(X) &= \tfrac{p}{2} \log \left\{ 2\pi^{\frac{2-\alpha}{2}} e^{\frac{2\gamma-\alpha}{\gamma}} (\tfrac{\alpha-1}{\alpha})^{\alpha-1} \right\} + \\ &= \tfrac{\alpha}{2} \log \left\{ \pi^{p/2} (\tfrac{\gamma}{\gamma-1})^{p\frac{\gamma-1}{\gamma}} \frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{\Gamma(p\frac{\alpha-1}{\alpha}+1)} \sqrt{|\det \Sigma|} \right\}, \end{aligned}$$

and after some algebra we derive (36).

In case of $\alpha = \gamma$ we have $\mathrm{H}_\gamma(X) = \frac{p}{2} \log\{2\pi e |\det \Sigma|^{\gamma/(2p)}\}$, i.e. (37) holds. □

Proposition 3. *For a random variable X following the multivariate Uniform, Normal, and Laplace distributions ($\gamma = 1, 2, \pm\infty$, respectively), it is*

$$\mathrm{H}_\alpha(X) = \begin{cases} \frac{2-\alpha}{2}p + h(\Sigma), & \text{for } X \sim \mathscr{U}^p(\mu, \Sigma), \\ p + \frac{\alpha}{2}\log\left\{(2/e)^{p/2}\Gamma(\frac{p}{2}+1)\right\} + h(\Sigma), & \text{for } X \sim \mathscr{N}^p(\mu, \Sigma), \\ p + \frac{p}{2}\log p! + h(\Sigma), & \text{for } X \sim \mathscr{L}^p(\mu, \Sigma), \end{cases} \tag{38}$$

where

$$h(\Sigma) := \tfrac{\alpha}{2}\log\left\{(2\pi)^{p/\alpha}(\tfrac{\alpha-1}{\alpha})^{p\frac{\alpha-1}{\alpha}}[\Gamma(p\tfrac{\alpha-1}{\alpha}+1)]^{-1}\sqrt{|\det\Sigma|}\right\}, \tag{39}$$

while for the limiting degenerate case of $X \sim \mathscr{N}_0^p(\mu, \Sigma)$ we obtain

$$\mathrm{H}_\alpha(X) = \begin{cases} (\operatorname{sgn}\alpha)(+\infty), & \text{for } \alpha \neq 0, \\ p\log\sqrt{2\pi e}, & \text{for } \alpha = 0. \end{cases} \tag{40}$$

Proof. Recall (29) and let $X_\gamma := X$. The α-Shannon entropy of r.v. X_γ, with $\gamma = 1, \pm\infty$, can be considered as

$$\mathrm{H}_\alpha(X_1) := \lim_{\gamma\to 1^+}\mathrm{H}_\alpha(X_\gamma), \quad \text{and} \quad \mathrm{H}_\alpha(X_{\pm\infty}) := \lim_{\gamma\to\pm\infty}\mathrm{H}_\alpha(X_\gamma).$$

Hence, for order values $\gamma = 1$ (in limit), $\gamma = 2$ and $\gamma = \pm\infty$ (in limit), we derive (38).

Consider now the limiting case of $\gamma = 0$. We can write (36) in the form

$$\begin{aligned}\mathrm{H}_\alpha(X_\gamma) &= \tfrac{p}{2}(2-\alpha+\gamma g) + \tfrac{p}{2}\log\left\{2\pi(\tfrac{g-1}{g})^{\alpha-1}g^{-g\alpha}\left[\frac{\Gamma(pg+1)\sqrt{|\det\Sigma|}}{\Gamma(p\frac{\alpha-1}{\alpha}+1)}\right]^{\frac{\alpha}{p}}\right\} \\ &= \log\left\{(2\pi)^{p/2}(\tfrac{\alpha-1}{\alpha})^{p\frac{\alpha-1}{2}}\left[\frac{\Gamma(pg+1)}{(\frac{g}{e})^{pg}\Gamma(p\frac{\alpha-1}{\alpha}+1)}\right]^{\frac{\alpha}{2}}|\det\Sigma|^\alpha\right\},\end{aligned}$$

where $g := \frac{\gamma-1}{\gamma}$. We then have,

$$\mathrm{H}_\alpha(X_0) := \lim_{\gamma\to 0^-}\mathrm{H}_\alpha(X_\gamma) = \log\left\{(2\pi)^{p/2}(\tfrac{\alpha-1}{\alpha})^{p\frac{\alpha-1}{2}}\left[\lim_{k:=p[g]\to\infty}\frac{p^k k!}{(\frac{k}{e})^k}\right]^{\frac{\alpha}{2}}|\det\Sigma|^\alpha\right\}.$$

Using the Stirling's asymptotic formula (similar as in Theorem 6), the above relation for $\alpha \neq 0$ implies

$$\mathrm{H}_\alpha(X_0) = \log\left\{(2\pi)^{p/2}\left(\tfrac{\alpha-1}{\alpha}\right)^{p\frac{\alpha-1}{2}}|\det\Sigma|^\alpha\left(\lim_{k\to\infty}p^k\sqrt{k}\right)^{\frac{\alpha}{2}}\right\} = (\operatorname{sgn}\alpha)(+\infty),$$

where $\operatorname{sgn}\alpha$ is the sign of parameter α, and hence the first branch of (40) holds. For the limiting case of $\gamma = \alpha = 0$, (37) implies the second branch of (40). □

Notice that despite the rather complicated form of the $\mathrm{H}_\alpha(X)$ when $\alpha \neq \gamma$ in Proposition 2, the γ-Shannon entropy of a γ-order normally distributed X_γ has a very compact expression, see (37), while in (36) varies both the shape parameter γ (to decide the distributions "fat tails" or not) and the parameter α (of the Shannon entropy) vary.

Recall now the known relation of the Shannon entropy of a normally distributed random variable $Z \sim \mathcal{N}(\mu, \Sigma)$, i.e. $\mathrm{H}(Z) = \frac{1}{2}\log\{(2\pi e)^p|\det\Sigma|\}$. Therefore, $\mathrm{H}_\gamma(X_\gamma)$, where $X_\gamma \sim \mathcal{N}_\gamma(\mu, \Sigma)$ generalizes $\mathrm{H}(Z)$, or equivalently $\mathrm{H}_2(X_2)$, preserving the simple formulation for every γ, as parameter γ affects only the scale matrix Σ.

Another interesting fact about $\mathrm{H}_\gamma(X_\gamma)$ is that, $\mathrm{H}_0(X_0) = \frac{p}{2}\log\{2\pi e\}$ or $\mathrm{H}_0(X_0) = -\frac{p}{4}\log\nu$, recall Corollary (40) and (25). According to (40) the Shannon entropy diverges to $+\infty$ for the degenerated distribution $\mathcal{N}_0$. However, the 0-Shannon entropy H_0 (in limit), for an r.v. following $\mathcal{N}_0$, converges to $\log\sqrt{2\pi e} = -\frac{1}{2}\log\nu \approx 1.4189$, which is the same value as the Shannon entropy of the standardized normally distributed $Z \sim \mathcal{N}(0, 1)$. Thus, the generalized Shannon entropy, introduced already, can "handle" the degenerated $\mathcal{N}_0$ distribution in a more "coherent" way than the usual Shannon entropy (i.e., not diverging to infinity).

We can mention also that (36) expresses the generalized α-Shannon entropy of the multivariate Uniform, Normal, and Laplace distributions relative to each other. For example (recall Corollary 3), the difference of these entropies between Uniform and Laplace is independent of the same scale matrix Σ, i.e. $\mathrm{H}_\alpha(X_{\pm\infty}) - \mathrm{H}_\alpha(X_1) = \frac{p}{2}(\alpha + \log p!)$, while for the usual Shannon entropy, $\mathrm{H}(X_{\pm\infty}) - \mathrm{H}(X_1) = p + \frac{p}{2}\log p!$, i.e. their Shannon entropies differ by a dimension-depending constant. The difference ratio is then

$$\frac{\mathrm{H}_\alpha(X_{\pm\infty}) - \mathrm{H}_\alpha(X_1)}{\mathrm{H}(X_{\pm\infty}) - \mathrm{H}(X_1)} = \frac{\log(p!e^\alpha)}{\log(p!e^2)}.$$

3.2 Generalized Entropy Power

So far we have developed a generalized form for the Shannon entropy. We shall now discuss and provide general results about the generalized entropy power. The typical cases are presented in (46). Notice that, as N_α and $\mathrm{H}\alpha$ are related, some of the proofs are consequences of this relation, see Proposition 4. The following holds for different α and γ parameters.

Proposition 4. *The generalized entropy power* $\mathrm{N}_\alpha(X)$ *of the multivariate* $X_\gamma \sim \mathscr{N}_\gamma^p(\mu, \Sigma)$ *is given, for all defined parameters* $\alpha, \gamma \in \mathbb{R} \setminus [0,\ 1]$, *by*

$$\mathrm{N}_\alpha(X_\gamma) = \left(\tfrac{\alpha-1}{e\alpha}\right)^{\alpha-1} \left(\tfrac{e\gamma}{\gamma-1}\right)^{\alpha\frac{\gamma-1}{\gamma}} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\alpha/p} |\det\Sigma|^{\frac{\alpha}{2p}}. \tag{41}$$

Moreover, in case of $\alpha = \gamma \in \mathbb{R} \setminus [0,\ 1]$,

$$\mathrm{N}_\gamma(X_\gamma) = |\det\Sigma|^{\frac{\gamma}{2p}}. \tag{42}$$

Proof. Substituting (33) into (11), we obtain (41) and (42). □

Corollary 2. *For the usual entropy power of the* γ*-order normally distributed r.v.* $X_\gamma \sim \mathscr{N}_\gamma^p(\mu, \Sigma)$, *we have that*

$$\mathrm{N}(X_\gamma) = \tfrac{1}{2e}\left(\tfrac{e\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{\Gamma\left(\frac{p}{2}+1\right)}\right]^{2/p} |\det\Sigma|^{1/p}, \quad \gamma \in \mathbb{R}\setminus[0,\ 1]. \tag{43}$$

For the multivariate Uniform, Normal, and Laplace distributions ($\gamma = 1, 2, \pm\infty$, *respectively), as well as for the degenerate case of* $\gamma = 0$, *it is*

$$\mathrm{N}(X) = \begin{cases} \dfrac{|\det\Sigma|^{1/p}}{2e\Gamma\left(\frac{p}{2}+1\right)^{2/p}}, & \text{for } X \sim \mathscr{U}^p(\mu,\Sigma), \\ \sqrt[p]{|\det\Sigma|}, & \text{for } X \sim \mathscr{N}^p(\mu,\Sigma), \\ 2^{\frac{2-p}{p}} e \left[\dfrac{(p-1)!\sqrt{|\det\Sigma|}}{\Gamma(p/2)}\right]^{2/p}, & \text{for } X \sim \mathscr{L}^p(\mu,\Sigma), \\ +\infty, & \text{for } X \sim \mathscr{N}_0^p(\mu,\Sigma). \end{cases} \tag{44}$$

Proof. For the normality parameter $\alpha = 2$, (43) is obtained from (41).

Recall (29) and let $X_\gamma := X$. The usual entropy power of r.v. X_γ, with $\gamma = 1, \pm\infty$, can be considered as

$$\mathrm{N}(X_1) := \lim_{\gamma\to 1^+} \mathrm{N}(X_\gamma) \quad \text{and} \quad \mathrm{N}(X_{\pm\infty}) := \lim_{\gamma\to\pm\infty} \mathrm{N}(X_\gamma).$$

Hence, for order values $\gamma = 1$ (in limit), $\gamma = 2$ and $\gamma = \pm\infty$ (in limit), we derive the first three branches of (44).

Consider now the limiting case of $\gamma = 0$. We can write (43) in the form

$$\mathrm{N}(X_\gamma) = \frac{|\det\Sigma|^{1/p}}{2e\Gamma\left(\frac{p}{2}+1\right)^{2/p}} \left(\tfrac{e}{g}\right)^{2g} \Gamma(pg+1)^{2/p},$$

where $g := \frac{\gamma-1}{\gamma}$. We then have

$$\mathrm{N}(X_0) := \lim_{\gamma\to 0^-} \mathrm{N}(X_\gamma) = \frac{|\det \Sigma|^{1/p}}{2e\Gamma\left(\frac{p}{2}+1\right)^{2/p}} \lim_{k:=p[g]\to\infty} \left(\frac{ep}{k}\right)^{2k/p} (k!)^{2/p}.$$

Using the Stirling's asymptotic formula (similar as in Theorem 6), the above relation implies

$$\mathrm{N}(X_0) = \frac{(2\pi|\det \Sigma|)^{1/p}}{2e\Gamma\left(\frac{p}{2}+1\right)^{2/p}} \lim_{k\to\infty} p^{2k/p} k^{1/p} = +\infty,$$

and hence the last branch of (44) holds. □

Example 2. For the univariate $p = 1$, (43) implies

$$\mathrm{N}(X_\gamma) = \tfrac{2}{\pi e}\left(\tfrac{e\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}} \Gamma\left(\tfrac{\gamma-1}{\gamma}+1\right)^2 \sigma^2, \tag{45}$$

and thus we derive from (44) that

$$\mathrm{N}(X) = \begin{cases} \frac{b-a}{\pi e}, & \text{for } X \sim \mathscr{U}(a,b), \\ \sigma^2, & \text{for } X \sim \mathscr{N}(\mu,\sigma^2), \\ \frac{2e\sigma}{\pi}, & \text{for } X \sim \mathscr{L}(\mu,\sigma), \\ +\infty, & \text{for } X \sim \mathscr{D}(\mu). \end{cases} \tag{46}$$

Corollary 3. *For the generalized entropy power of the multivariate Uniform, Normal, and Laplace distributions ($\gamma = 1, 2, \pm\infty$, respectively), it is*

$$\mathrm{N}_\alpha(X) = \begin{cases} \left(\frac{\alpha-1}{e\alpha}\right)^{\alpha-1} \dfrac{|\det \Sigma|^{\frac{\alpha}{2p}}}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)^{\alpha/p}}, & \text{for } X \sim \mathscr{U}^p(\mu,\Sigma), \\ \left(\frac{\alpha-1}{e\alpha}\right)^{\alpha-1} (2e)^{\alpha/2} \left[\dfrac{\Gamma(\frac{p}{2}+1)}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\alpha/p} |\det \Sigma|^{\frac{\alpha}{2p}}, & \text{for } X \sim \mathscr{N}^p(\mu,\Sigma), \\ e\left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1} \left[\dfrac{p!}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\alpha/p} |\det \Sigma|^{\frac{\alpha}{2p}}, & \text{for } X \sim \mathscr{L}^p(\mu,\Sigma), \end{cases} \tag{47}$$

while for the degenerate case of $X \sim \mathscr{N}_0^p(\mu,\Sigma)$ we have

$$\mathrm{N}_\alpha(X) = \begin{cases} +\infty, & \text{for } \alpha > 1, \\ 0, & \text{for } \alpha < 0. \end{cases} \tag{48}$$

Proof. Recall (29) and let $X_\gamma \sim \mathcal{N}_\gamma(\mu, \Sigma)$. The generalized entropy power of r.v. X_γ, with $\gamma = 1, \pm\infty$, can be considered as

$$\mathrm{N}_\alpha(X_1) := \lim_{\gamma\to 1^+} \mathrm{N}_\alpha(X_\gamma), \quad \text{and} \quad \mathrm{N}_\alpha(X_{\pm\infty}) := \lim_{\gamma\to\pm\infty} \mathrm{N}_\alpha(X_\gamma),$$

and hence, for order values $\gamma = 1$ (in limit), $\gamma = 2$ and $\gamma = \pm\infty$ (in limit), we derive (47).

Consider now the limiting case of $\gamma = 0$. We can write (41) in the form

$$\mathrm{N}_\alpha(X_\gamma) = \left(\tfrac{\alpha-1}{e\alpha}\right)^{\alpha-1} \left(\tfrac{e}{g}\right)^{g\alpha} \left[\frac{\Gamma(pg+1)}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)}\right]^{\alpha/p} |\det\Sigma|^{\frac{\alpha}{2p}},$$

where $g := \frac{\gamma-1}{\gamma}$. We then have

$$\mathrm{N}_\alpha(X_0) := \lim_{\gamma\to 0^-} \mathrm{N}_\alpha(X_\gamma) = \frac{\left(\frac{\alpha-1}{e\alpha}\right)^{\alpha-1} |\det\Sigma|^{\frac{\alpha}{2p}}}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)^{\alpha/p}} \lim_{k:=p[g]\to\infty} \left(\tfrac{ep}{k}\right)^{\alpha k/p} (k!)^{\alpha/p}.$$

Using the Stirling's asymptotic formula, the above relation implies

$$\mathrm{N}_\alpha(X_0) = \frac{\left(\frac{\alpha-1}{e\alpha}\right)^{\alpha-1} |\det\Sigma|^{\frac{\alpha}{2p}}}{\Gamma\left(p\frac{\alpha-1}{\alpha}+1\right)^{\alpha/p}} \lim_{k\to\infty} (2\pi p^2 k)^{\alpha k/(2p)} = \begin{cases} +\infty, & \text{for } \alpha > 1, \\ 0, & \text{for } \alpha < 0, \end{cases}$$

and hence (48) holds. □

For the special cases $\alpha = 0, 1, \pm\infty$ of the parameter $\alpha \in \mathbb{R} \setminus [0, 1]$ of the generalized entropy power, the following holds.

Proposition 5. *The generalized entropy power* $\mathrm{N}_\alpha(X)$*, for the limiting values* $\alpha = 0, 1, \pm\infty$*, of the multivariate r.v.* $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \Sigma)$*, for all shape parameter values* $\gamma \in \mathbb{R} \setminus [0, 1]$*, is given by*

$$\mathrm{N}_0(X_\gamma) = p, \tag{49}$$

$$\mathrm{N}_1(X_\gamma) = \left(\tfrac{e\gamma}{\gamma-1}\right)^{\frac{\gamma-1}{\gamma}} \Gamma\left(p\tfrac{\gamma-1}{\gamma}+1\right)^{\frac{1}{p}} |\det\Sigma|^{\frac{1}{2p}}, \tag{50}$$

$$\mathrm{N}_{+\infty}(X_\gamma) = \begin{cases} +\infty, & \textit{for } |\det\Sigma| > S_\gamma^2, \\ 0, & \textit{for } |\det\Sigma| < S_\gamma^2, \end{cases} \tag{51}$$

$$\mathrm{N}_{-\infty}(X_\gamma) = \begin{cases} +\infty, & \textit{for } |\det\Sigma| < S_\gamma^2, \\ 0, & \textit{for } |\det\Sigma| > S_\gamma^2, \end{cases} \tag{52}$$

where

$$S_\gamma := \frac{e^p p!(\frac{\gamma-1}{e\gamma})^{p\frac{\gamma-1}{\gamma}}}{\Gamma(p\frac{\gamma-1}{\gamma}+1)}. \tag{53}$$

Proof. For the limiting value $\alpha = 0$, we can consider

$$\mathrm{N}_0(X_\gamma) := \lim_{\alpha\to 0^-} \mathrm{N}_\alpha(X_\gamma),$$

which can be written, through (41), into the form

$$\mathrm{N}_0(X_\gamma) = \lim_{\beta\to+\infty} (\tfrac{\beta}{e})^{\frac{\beta}{1-\beta}} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{\Gamma(p\beta+1)}\right]^{\frac{1}{p(1-\beta)}} |\det\Sigma|^{\frac{1}{2p(1-\beta)}},$$

where $\beta := \frac{\alpha-1}{\alpha}$, or

$$\mathrm{N}_0(X_\gamma) = \lim_{k:=p[\beta]\to\infty} (\tfrac{k}{pe})^{\frac{k}{p-k}} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)}{k!}\right]^{\frac{1}{p-k}} |\det\Sigma|^{\frac{1}{2(p-k)}}.$$

Applying the Stirling's asymptotic formula for $k!$, the above relation implies

$$\mathrm{N}_0(X_\gamma) = \lim_{k\to\infty} \left[\frac{\Gamma(p\frac{\gamma-1}{\gamma}+1)\sqrt{|\det\Sigma|}}{p^k\sqrt{2\pi k}}\right]^{\frac{1}{p-k}} = \lim_{k\to\infty} p^{\frac{k}{k-p}} k^{\frac{1}{2(k-p)}} = p\cdot 1 = p,$$

and therefore, (49) holds due to the fact that

$$\lim_{k\to\infty} k^{\frac{1}{2(k-p)}} = \exp\left\{\tfrac{1}{2}\lim_{k\to\infty}\frac{\log k}{k-p}\right\} = e^0 = 1. \tag{54}$$

For the limiting value $\alpha = 1$, we can consider

$$\mathrm{N}_1(X_\gamma) := \lim_{\alpha\to 1^+} \mathrm{N}_\alpha(X_\gamma),$$

and thus (50) hold, through (41).

For the limiting value $\alpha = \pm\infty$, we can consider

$$\mathrm{N}_{\pm\infty}(X_\gamma) := \lim_{\alpha\to\pm\infty} \mathrm{N}_\alpha(X_\gamma) = \lim_{\alpha\to\pm\infty} \left[\frac{A(\Sigma)}{\Gamma(p\frac{\alpha-1}{\alpha}+1)}\right]^{\alpha/p}, \tag{55}$$

where

$$A(\Sigma) := e^{-p}\left(\frac{e\gamma}{\gamma-1}\right)^{p\frac{\gamma-1}{\gamma}}\Gamma\left(p\frac{\gamma-1}{\gamma}+1\right)\sqrt{|\det\Sigma|}, \tag{56}$$

due to (41) and the fact that $\lim_{\alpha\to\pm\infty}[(\alpha-1)/\alpha]^{\alpha-1} = e^{-1}$. Moreover, $\Gamma(p\frac{\alpha-1}{\alpha}+1)\to p!$ as $\alpha\to\pm\infty$, and thus, from (55) we obtain (51) and (52). □

Corollary 2 presents the usual entropy power $\mathrm{N}(X) = \mathrm{N}_{\alpha=2}(X)$ when X follows a Uniform, Normal, Laplace, or a degenerated ($\mathscr{N}_0$) random variable. The following Proposition investigates the limiting cases of $\mathrm{N}_{\alpha=0,1,\pm\infty}(X)$, as it provides results for applications working with "extreme-tailed" distributions. Notice the essential use of the quantity S_2, as in (65), for the determinant of the distributions' scale matrix Σ, that alters the behavior of the extreme case of $\mathrm{N}_{\pm\infty}$.

Proposition 6. *For the multivariate Uniform, Normal, and Laplace distributions, i.e. $\mathscr{N}_{\gamma=1,2,\pm\infty}$, as well as for the degenerate $\mathscr{N}_{\gamma=0}$, the "limiting values" of the generalized entropy power $\mathrm{N}_{\alpha=0,1,\pm\infty}$, are given by*

$$\mathrm{N}_0(X) = \begin{cases} p, & \text{for } X\sim\mathscr{U}^p(\mu,\Sigma),\\ p, & \text{for } X\sim\mathscr{N}^p(\mu,\Sigma),\\ p, & \text{for } X\sim\mathscr{L}^p(\mu,\Sigma),\\ 1, & \text{for } X\sim\mathscr{N}_0^p(\mu,\Sigma), \end{cases} \tag{57}$$

$$\mathrm{N}_1(X) = \begin{cases} |\det\Sigma|^{\frac{1}{2p}}, & \text{for } X\sim\mathscr{U}^p(\mu,\Sigma),\\ \sqrt{2e}\Gamma(\frac{p}{2}+1)^{1/p}|\det\Sigma|^{\frac{1}{2p}}, & \text{for } X\sim\mathscr{N}^p(\mu,\Sigma),\\ e(p!)^{1/p}|\det\Sigma|^{\frac{1}{2p}}, & \text{for } X\sim\mathscr{L}^p(\mu,\Sigma),\\ +\infty, & \text{for } X\sim\mathscr{N}_0^p(\mu,\Sigma), \end{cases} \tag{58}$$

$$\mathrm{N}_{+\infty}(X) = \begin{cases} +\infty, & \text{for } |\det\Sigma| > (e^p p!)^2,\\ 0, & \text{for } |\det\Sigma| < (e^p p!)^2, \end{cases} \quad \text{and } X\sim\mathscr{U}^p(\mu,\Sigma), \tag{59}$$

$$\mathrm{N}_{-\infty}(X) = \begin{cases} +\infty, & \text{for } |\det\Sigma| < (e^p p!)^2,\\ 0, & \text{for } |\det\Sigma| > (e^p p!)^2, \end{cases} \quad \text{and } X\sim\mathscr{U}^p(\mu,\Sigma), \tag{60}$$

$$\mathrm{N}_{+\infty}(X) = \begin{cases} +\infty, & \text{for } |\det\Sigma| > S_2^2,\\ 0, & \text{for } |\det\Sigma| < S_2^2, \end{cases} \quad \text{and } X\sim\mathscr{N}^p(\mu,\Sigma), \tag{61}$$

$$\mathrm{N}_{-\infty}(X) = \begin{cases} +\infty, & \text{for } |\det\Sigma| < S_2^2,\\ 0, & \text{for } |\det\Sigma| > S_2, \end{cases} \quad \text{and } X\sim\mathscr{N}^p(\mu,\Sigma), \tag{62}$$

$$\mathrm{N}_{+\infty}(X) = \begin{cases} +\infty, & for \ |\det \Sigma| > 1, \\ 0, & for \ |\det \Sigma| < 1, \end{cases} \quad and \ X \sim \mathscr{L}^p(\mu, \Sigma), \tag{63}$$

$$\mathrm{N}_{-\infty}(X) = \begin{cases} +\infty, & for \ |\det \Sigma| < 1, \\ 0, & for \ |\det \Sigma| > 1, \end{cases} \quad and \ X \sim \mathscr{L}^p(\mu, \Sigma), \tag{64}$$

where

$$S_2 := \frac{e^p p!}{(2e)^{p/2}\Gamma(\frac{p}{2}+1)}. \tag{65}$$

Proof. For the limiting value $\alpha = 1$, the first three branches of (58) holds, through (47). Moreover, for the degenerate case of $\mathscr{N}_{\gamma=0}$, we consider $\mathrm{N}_1(X_0) := \lim_{\gamma\to 0^-} \mathrm{N}_1(X_\gamma)$, with $X_\gamma \sim \mathscr{N}_\gamma(\mu, \Sigma)$, i.e.

$$\mathrm{N}_1(X_0) = \lim_{g\to+\infty} (\tfrac{e}{g})^g \Gamma(pg+1)^{\frac{1}{p}} |\det \Sigma|^{\frac{1}{2p}},$$

where $g := \frac{\gamma-1}{\gamma}$. Then,

$$\mathrm{N}_1(X_0) = \lim_{k:=p[g]\to\infty} (\tfrac{ep}{k})^{k/p} (k!)^{\frac{1}{p}} |\det \Sigma|^{\frac{1}{2p}}.$$

Applying the Stirling's asymptotic formula of $k!$, the above relation implies

$$\mathrm{N}_1(X_0) = \lim_{k\to\infty} p^{k/p} (2\pi k |\det \Sigma|)^{\frac{1}{2p}} = +\infty,$$

and thus (58) holds.

For the limiting value $\alpha = \pm\infty$ and $X \sim \mathscr{N}_1(\mu, \Sigma) = \mathscr{U}^p(\mu, \Sigma)$, we consider $\mathrm{N}_{\pm\infty}(X) := \lim_{\alpha\to\pm\infty} \mathrm{N}_\alpha(X)$, i.e.

$$\mathrm{N}_{\pm\infty}(X) = \lim_{\alpha\to\pm\infty} \left[\frac{\sqrt{|\det \Sigma|}}{e^p \Gamma(p\frac{\alpha-1}{\alpha}+1)} \right]^{\alpha/p}, \tag{66}$$

from (47). Moreover, $\Gamma(p\frac{\alpha-1}{\alpha}+1) \to p!$ as $\alpha \to \pm\infty$, and thus, from (66), we obtain (59) and (60)

For the limiting value $\alpha = \pm\infty$ and $X \sim \mathscr{N}_2(\mu, \Sigma) = \mathscr{N}^p(\mu, \Sigma)$, relations (61) and (62) hold due to (51) and (52), where S_2 as in (53) with $\gamma = 2$.

For the limiting value $\alpha = \pm\infty$ and $X \sim \mathscr{N}_{\pm\infty}(\mu, \Sigma) = \mathscr{L}^p(\mu, \Sigma)$, relations (63) and (64) hold due to (42) with $\gamma \to \pm\infty$.

For the limiting value $\alpha = 0$ and $X \sim \mathscr{N}_1(\mu, \Sigma) = \mathscr{U}^p(\mu, \Sigma)$ we consider $\mathrm{N}_0(X) := \lim_{\alpha\to 0^-} \mathrm{N}_\alpha(X)$ which can be written, through the first branch of (47), into the form

$$\mathrm{N}_0(X) = \lim_{\beta\to+\infty} \frac{(\frac{\beta}{e})^{\frac{\beta}{1-\beta}}}{\Gamma(p\beta+1)^{\frac{1}{p(1-\beta)}}} |\det\Sigma|^{\frac{1}{2p(1-\beta)}},$$

where $\beta := \frac{\alpha-1}{\alpha}$, or

$$\mathrm{N}_0(X) = \lim_{k:=p[\beta]\to\infty} (\tfrac{k}{pe})^{\frac{k}{p-k}} (k!)^{\frac{1}{k-p}} |\det\Sigma|^{\frac{1}{2(p-k)}}.$$

Applying the Stirling's asymptotic formula for $k!$, the above relation implies

$$\mathrm{N}_0(X) = \lim_{k\to\infty} \left(\frac{\sqrt{|\det\Sigma|}}{p^k\sqrt{2\pi k}}\right)^{\frac{1}{p-k}} = \lim_{k\to\infty} p^{\frac{k}{k-p}} k^{\frac{1}{2(k-p)}} = p\cdot 1 = p,$$

and therefore the first branch of (57) holds due to (54).

For the limiting value $\alpha = 0$ and $X \sim \mathscr{N}_2(\mu,\Sigma) = \mathscr{N}^p(\mu,\Sigma)$, the second branch of (57) holds due to (49).

For the limiting value $\alpha = 0$ and $X \sim \mathscr{N}_{\pm\infty}(\mu,\Sigma) = \mathscr{L}^p(\mu,\Sigma)$ we consider $\mathrm{N}_0(X) := \lim_{\alpha\to 0^-} \mathrm{N}_\alpha(X)$ which can be written, through the last branch of (47), into the form

$$\mathrm{N}_0(X) = \lim_{\beta\to+\infty} (\tfrac{\beta}{e})^{\frac{\beta}{1-\beta}} \left[\frac{p!}{\Gamma(p\beta+1)}\right]^{\frac{1}{p(1-\beta)}} |\det\Sigma|^{\frac{1}{2p(1-\beta)}},$$

or

$$\mathrm{N}_0(X) = \lim_{k:=p[\beta]\to\infty} (\tfrac{k}{pe})^{\frac{k}{p-k}} (\tfrac{p!}{k!})^{\frac{1}{p-k}} |\det\Sigma|^{\frac{1}{2(p-k)}}.$$

Applying, again, the Stirling's asymptotic formula for $k!$, the above relation implies

$$\mathrm{N}_0(X) = \lim_{k\to\infty} \left(\frac{p!\sqrt{|\det\Sigma|}}{p^k\sqrt{2\pi k}}\right)^{\frac{1}{p-k}} = \lim_{k\to\infty} p^{\frac{k}{k-p}} k^{\frac{1}{2(k-p)}} = p\cdot 1 = p,$$

and therefore the third branch of (57) holds due to (54).

For the limiting value $\alpha = 0$ and $X \sim \mathscr{N}_0(\mu,\Sigma)$, the last branch of (57) holds due to (42) with $\gamma \to 0^-$. □

Recall Proposition 5 where $\gamma \in \mathbb{R}\setminus[0,1]$. For the limiting extra values of $\gamma = 1$ (Uniform case), $\gamma = \pm\infty$ (Laplace case), and $\gamma = 0$ (degenerate case), the results (50)–(52) still hold in limit, see (58) and from (59) to (64). Therefore, the relations (50)–(52) hold for all shape parameters γ taking values over its "extended" domain, i.e. $\gamma \in \mathbb{R}\setminus(0,1)\cup\{\pm\infty\}$. However, from (49) and (57), it holds that

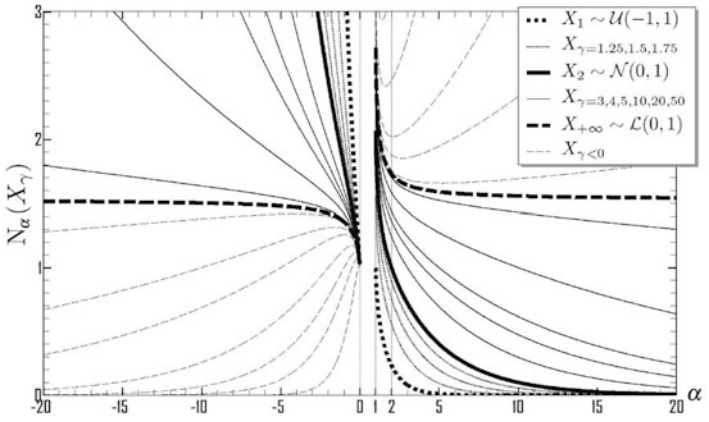


Fig. 2 Graphs of $\mathrm{N}_\alpha(X_\gamma)$ along α, for various γ values, where $X_\gamma \sim \mathscr{N}_\gamma(0,1)$

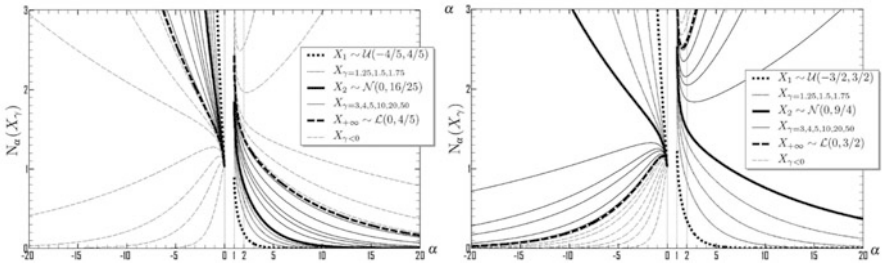


Fig. 3 Graphs of $\mathrm{N}_\alpha(X_\gamma)$ along α, for various γ values, where $X_\gamma \sim \mathscr{N}_\gamma(0, 0.8)$ (*left-side*) and $X_\gamma \sim \mathscr{N}_\gamma(0, 1.5)$ (*right-side*)

$$\mathrm{N}_0(X_\gamma) = \begin{cases} p, & \text{for } \gamma \in \mathbb{R} \setminus [0,\ 1) \cup \{\pm\infty\}, \\ 1, & \text{for } \gamma = 0. \end{cases} \tag{67}$$

while for the univariate case, the generalized entropy power N_0, as in (67), is always unity for all the members of the "extended" γ-order Normal distribution's family, i.e. $\mathrm{N}_0(X_\gamma) = 1$ with $\gamma \in \mathbb{R} \setminus (0,\ 1) \cup \{\pm\infty\}$.

Figure 2 presents the generalized entropy power $\mathrm{N}_\alpha(X_\gamma)$ as a function of its parameter $\alpha \in \mathbb{R} \setminus [0,\ 1]$ for various $X_\gamma \sim \mathscr{N}_\gamma(0,1)$ random variables, with the special cases of Uniform ($\gamma = 1$), Normal ($\gamma = 2$), and Laplace ($\gamma = \pm\infty$) r.v. being denoted. Figure 3 depicts the cases of $X_\gamma \sim \mathscr{N}_\gamma(0, \sigma^2)$ with $\sigma < 1$ (left sub-Figure) and $\sigma > 1$ (right sub-Figure).

3.3 Rényi Entropy

We discuss now the Rényi entropy, another significant entropy measure which also generalizes Shannon entropy, and can be best introduced through the concept of *generalized random variables*. These variables extend the usual notion of a random experiment that cannot always be observed. See for details the Rényi's original work in [25] and [26].

See [5]. For a p-variate continuous random variable, with p.d.f. f_X, the Rényi entropy $\mathrm{R}_\alpha(X)$ is defined, through the α-norm $\|\cdot\|_\alpha$ on $\mathscr{L}^\alpha(\mathbb{R}^p)$, by

$$\mathrm{R}_\alpha(X) := -\tfrac{\alpha}{\alpha-1}\log\|f_X\|_\alpha = \tfrac{1}{1-\alpha}\log\int_{\mathbb{R}^p}|f_X(x)|^\alpha dx, \tag{68}$$

with $\alpha \in \mathbb{R}_+^* \setminus 1$, i.e. $0 < \alpha \neq 1$. For the limiting case of $\alpha = 1$ the Rényi entropy converges to the usual Shannon entropy $\mathrm{H}(X)$ as in (9). Notice that we use the minus sign for R_α to be in accordance with the definition of (9), where we reject the usual minus sign of the Shannon entropy definition.

Considering now an r.v. from the $\mathscr{N}_\gamma^p$ family of the generalized Normal distributions, the following Theorem provides a general result to calculate the Rényi entropy for different α and γ parameters.

Theorem 7. *For the elliptically contoured γ-order normally distributed r.v. $X_\gamma \sim \mathscr{N}_\gamma^p(\mu, \Sigma)$, with p.d.f. f_{X_γ}, the Rényi R_α entropy of X_γ is given by*

$$\mathrm{R}_\alpha(X_\gamma) = p\tfrac{\gamma-1}{\gamma(\alpha-1)}\log\alpha - \log C_\gamma^p(\Sigma) = p\tfrac{\gamma-1}{\gamma(\alpha-1)}\log\alpha - \log\max f_{X_\gamma}, \tag{69}$$

for all the defined parameters $\alpha \in \mathbb{R}_+^* \setminus \{1\}$ and $\gamma \in \mathbb{R} \setminus [0,\ 1]$.

Proof. Consider the p.d.f. f_{X_γ} as in (22). From the definition (68) it is

$$\mathrm{R}_\alpha(X_\gamma) = \tfrac{\alpha}{1-\alpha}\log C_\gamma^p(\Sigma) + \tfrac{1}{1-\alpha}\log\int_{\mathbb{R}^p}\exp\left\{-\tfrac{\alpha(\gamma-1)}{\gamma}\left[(x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}\right]^{\frac{\gamma}{2(\gamma-1)}}\right\}dx.$$

Applying the linear transformation $z = (x-\mu)\Sigma^{-1/2}$ with $dx = d(x-\mu) = \sqrt{|\det\Sigma|}dz$, the R_α above is reduced to

$$\mathrm{R}_\alpha(X_\gamma) = \tfrac{\alpha}{1-\alpha}\log C_\gamma^p(\Sigma) + \tfrac{1}{1-\alpha}\log\int_{\mathbb{R}^p}\exp\left\{-\tfrac{\alpha(\gamma-1)}{\gamma}\|z\|^{\frac{\gamma}{\gamma-1}}\right\}dz.$$

Switching to hyperspherical coordinates, we get

$$\mathrm{R}_\alpha(X_\gamma) = \tfrac{\alpha}{1-\alpha}\log\left\{C_\gamma^p(\Sigma)\omega_{p-1}^{1/\alpha}\right\} + \tfrac{1}{1-\alpha}\log\int_{\mathbb{R}_+}\exp\left\{-\tfrac{\alpha(\gamma-1)}{\gamma}\rho^{\frac{\gamma}{\gamma-1}}\right\}\rho^{p-1}d\rho,$$

where $\omega_{p-1} = 2\pi^{p/2}/\Gamma\left(\frac{p}{2}\right)$ is the volume of the $(p-1)$-sphere. Assuming $du := d(\frac{\gamma-1}{\gamma}\rho^{\gamma/(\gamma-1)}) = \rho^{1/(\gamma-1)}d\rho$ we obtain successively

$$\begin{aligned}
\mathrm{R}_\alpha(X_\gamma) &= \tfrac{\alpha}{1-\alpha}\log M(\Sigma) + \tfrac{1}{1-\alpha}\log \int\limits_{\mathbb{R}_+} e^{-\alpha u}\rho^{\frac{(p-1)(\gamma-1)-1}{\gamma-1}}du \\
&= \tfrac{\alpha}{\alpha-1}\log M(\Sigma) + \tfrac{1}{1-\alpha}\log \int\limits_{\mathbb{R}_+} e^{-\alpha u}\left(\rho^{\frac{\gamma}{\gamma-1}}\right)^{\frac{(p-1)(\gamma-1)-1}{\gamma}}du \\
&= \tfrac{\alpha}{1-\alpha}\log M(\Sigma) + \tfrac{1}{1-\alpha}\log(\tfrac{\gamma}{\gamma-1})^{p\frac{\gamma-1}{\gamma}-1} + \tfrac{1}{1-\alpha}\log \int\limits_{\mathbb{R}_+} e^{-\alpha u}u^{p\frac{\gamma-1}{\gamma}-1}du \\
&= \tfrac{\alpha}{1-\alpha}\log M(\Sigma) + \tfrac{1}{1-\alpha}\log(\tfrac{\gamma}{\gamma-1})^{p\frac{\gamma-1}{\gamma}-1} - p\tfrac{\gamma-1}{\gamma}\cdot\tfrac{\log\alpha}{1-\alpha} + \tfrac{1}{1-\alpha}\log\Gamma(p\tfrac{\gamma-1}{\gamma}),
\end{aligned}$$

where $M(\Sigma) := C_\gamma^p(\Sigma)\omega_{p-1}^{1/\alpha}$. Finally, by substitution of the volume ω_{p-1} we obtain, through the normalizing factor $C_\gamma^p(\Sigma)$ as in (23),

$$\mathrm{R}_\alpha(X_\gamma) = -\tfrac{\alpha}{\alpha-1}\log C_\gamma^p(\Sigma) + \tfrac{1}{\alpha-1}\log C_\gamma^p(\Sigma) + p\tfrac{\gamma-1}{\gamma}\cdot\tfrac{\log\alpha}{\alpha-1},$$

and thus (69) holds true. □

For the limiting parameter values $\alpha = 0, 1, +\infty$ we obtain a number of results for other well-known measures of entropy, applicable to Cryptography, as the Hartley entropy, the Shannon entropy, and min-entropy, respectively, while for $\alpha = 2$ the collision entropy is obtained. Therefore, from Theorem 7, we have the following.

Corollary 4. *For the special cases of $\alpha = 0, 1, 2, +\infty$, the Rényi entropy of the elliptically contoured r.v. $X_\gamma \sim \mathscr{N}_\gamma(\mu, \Sigma)$ is reduced to*

$$\mathrm{R}_\alpha(X_\gamma) = \begin{cases} +\infty, & \text{for } \alpha = 0, \quad (\textit{Hartley entropy}) \\ p\frac{\gamma-1}{\gamma} - \log\max f_{X_\gamma}, & \text{for } \alpha = 1, \quad (\textit{Shannon entropy}) \\ p\frac{\gamma-1}{\gamma}\log 2 - \log\max f_{X_\gamma}, & \text{for } \alpha = 2, \quad (\textit{collision entropy}) \\ -\log\max f_{X_\gamma}, & \text{for } \alpha = +\infty, \quad (\textit{min-entropy}) \end{cases} \tag{70}$$

where $\max f_{X_\gamma} = C_\gamma^p(\Sigma)$.

The Rényi entropy $\mathrm{R}_\alpha(X_\gamma)$, as in (69), is a decreasing function of parameter $\alpha \in \mathbb{R}_+^* \setminus \{1\}$, and hence

$$\mathrm{R}_{+\infty}(X_\gamma) < \mathrm{R}_2(X_\gamma) < \mathrm{R}_1(X_\gamma) < \mathrm{R}_0(X_\gamma), \quad \gamma \in \mathbb{R} \setminus [0, 1],$$

while

$$\min_{0<\alpha\neq 1}\{\mathrm{R}_\alpha(X_\gamma)\} = \mathrm{R}_{+\infty}(X_\gamma) = -\log\max f_{X_\gamma} = -\log C_\gamma^p(\Sigma).$$

Corollary 5. *The Rényi entropy* R_α *of the multivariate and elliptically contoured Uniform random variable* $X \sim \mathscr{U}(\mu, \Sigma)$ *is* α*-invariant, as* $\mathrm{R}_\alpha(X)$ *equals to the logarithm of the volume* $\omega(\mathbb{E}_\theta)$ *of the* $(p-1)$*-ellipsoid* $\mathbb{E}_\theta : \ Q_\theta(x) = 1, x \in \mathbb{R}^p$*, in which the p.d.f. of the elliptically contoured Uniform r.v. X is actually defined, i.e.*

$$\mathrm{R}_\alpha(X) = \log \omega(\mathbb{E}_\theta) = \log \frac{\pi^{p/2} |\det \Sigma|^{-1/2}}{\Gamma(\frac{p}{2}+1)}, \quad \alpha \in \mathbb{R}_+^* \setminus \{1\}, \tag{71}$$

while for the univariate case of $X \sim \mathscr{U}(a, b)$ *it is reduced to*

$$\mathrm{R}_\alpha(X) = \log(b-a), \quad \alpha \in \mathbb{R}_+^* \setminus \{1\}.$$

Proof. Recall (29) and let $X_\gamma \sim \mathscr{N}_\gamma(\mu, \Sigma)$. Then, the Rényi entropy of the uniformly r.v. X can be considered as $\mathrm{R}_\alpha(X) := \lim_{\gamma \to 1^+} \mathrm{R}_\alpha(X_\gamma)$ and therefore, from (69), we obtain (71). □

Notice, from the above Corollary 5, that the Hartley, Shannon, collision, and min-entropy of a multivariate uniformly distributed r.v. coincide with $\log \omega(\mathbb{E}_\theta)$.

Corollary 6. *For the multivariate Laplace random variable* $X \sim \mathscr{L}(\mu, \Sigma)$*, the Rényi entropy is given by*

$$\mathrm{R}_\alpha(X) = p \tfrac{\log \alpha}{\alpha - 1} + L(S), \tag{72}$$

and the Hartley, Shannon, collision, and the min-entropy are then given by

$$\mathrm{R}_\alpha(X) = \begin{cases} +\infty, & \text{for } \alpha = 0, \quad \text{(Hartley entropy)} \\ p + L(\Sigma), & \text{for } \alpha = 1, \quad \text{(Shannon entropy)} \\ p \log 2 + L(\Sigma), & \text{for } \alpha = 2, \quad \text{(collision entropy)} \\ L(\Sigma), & \text{for } \alpha = +\infty, \quad \text{(min-entropy)} \end{cases} \tag{73}$$

where $L(\Sigma) := \log\{p! \pi^{p/2} |\det \Sigma|^{1/2} \Gamma(\frac{p}{2}+1)^{-1}\}$.

Proof. Recall (29) and let $X_\gamma \sim \mathscr{N}_\gamma(\mu, \Sigma)$. Then, the Rényi entropy of the Laplace r.v. X can be considered as $\mathrm{R}_\alpha(X) := \lim_{\gamma \to \pm\infty} \mathrm{R}_\alpha(X_\gamma)$ and therefore, from (69), we obtain (72), while through (70), relation (73) is finally derived. □

Relations (71) and (72) below provide a general compact form of Rényi entropy R_α (for the Uniform and Laplace r.v.) and can be compared with the α-Shannon entropy H_α (for the such r.v.), as in (38).

3.4 Generalized Fisher's Entropy Type Information

As far as the generalized Fisher's entropy type information measure $\mathrm{J}_\alpha(X_\gamma)$ is concerned, for the multivariate and spherically contoured r.v. $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \sigma^2\mathbb{I}_p)$, it holds, [18],

$$\mathrm{J}_\alpha(X_\gamma) = \left(\tfrac{\gamma}{\gamma-1}\right)^{\frac{\alpha}{\gamma}} \frac{\Gamma\left(\frac{\alpha+p(\gamma-1)}{\gamma}\right)}{\sigma^\alpha \Gamma\left(p\frac{\gamma-1}{\gamma}\right)}. \tag{74}$$

More general, the following holds [19].

Theorem 8. *The generalized Fisher's entropy type information* J_α *of a* γ*-order normally distributed r.v.* $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \Sigma)$*, where* Σ *is a definite positive real matrix consisted of orthogonal vectors (matrix columns) with the same norm, is given by*

$$\mathrm{J}_\alpha(X_\gamma) = \frac{\left(\frac{\gamma}{\gamma-1}\right)^{\frac{\alpha}{\gamma}} \Gamma\left(\frac{\alpha+p(\gamma-1)}{\gamma}\right)}{|\det \Sigma|^{\frac{\alpha}{2p}} \Gamma(p\frac{\gamma-1}{\gamma})}. \tag{75}$$

Therefore, for the spherically contoured case, (74) holds indeed, through Theorem 8.

Corollary 7. *The generalized Fisher's information* J_α *of a spherically contoured r.v.* $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \sigma^2\mathbb{I}_p)$*, with* $\alpha/\gamma \in \mathbb{N}^*$*, is reduced to*

$$\mathbf{J}_\alpha(X_\gamma) = \sigma^{-\alpha}(\gamma-1)^{-\alpha\gamma} \prod_{k=1}^{\alpha/\gamma} \{\alpha - p + (p-k)\gamma\}, \quad \alpha, \gamma > 1.$$

Proof. From (74) and the gamma function additive identity, i.e. $\Gamma(x+1) = x\Gamma(x)$, $x \in \mathbb{R}_+^*$, relation (7) holds

3.5 Kullback–Leibler Divergence

As far as the information "discrimination" or "distance" is concerned between two $\mathcal{N}_\gamma$ r.v., the Kullback–Leibler (K–L) measure of information divergence (also known as relative entropy) is evaluated. Recall the K–L divergence $\mathrm{D}_{\mathrm{KL}}(X, Y)$ defined in Sect. 1. Specifically, for two multivariate γ-order normally distributed r.v. with the same mean and shape, i.e. $X_i \in \mathcal{N}_\gamma(\mu_i, \sigma_i^2\mathbb{I}_p)$, $i = 1, 2$, with $\mu_1 = \mu_2$, the K–L divergence of X_1 over X_2 is given by, [16],

$$\mathrm{D}_{\mathrm{KL}}(X_1, X_2) = p \log \tfrac{\sigma_2}{\sigma_1} - p\left(\tfrac{\gamma-1}{\gamma}\right)\left[1 - \left(\tfrac{\sigma_1}{\sigma_2}\right)^{\frac{\gamma}{\gamma-1}}\right], \tag{76}$$

while for $\mu_1 \neq \mu_2$ and $\gamma = 2$,

$$\mathrm{D_{KL}}(X_1, X_2) = \frac{p}{2}\left[\left(\log\frac{\sigma_2^2}{\sigma_1^2}\right) - 1 + \frac{\sigma_1^2}{\sigma_2^2} + \frac{\|\mu_1-\mu_0\|^2}{p\sigma_2^2}\right].$$

Moreover, from (76), the K–L divergence between two uniformly distributed r.v. $U_1, U_2 \in \mathscr{U}^p(\mu, \sigma_i^2\mathbb{I}_p)$, $i = 1, 2$, is given by,

$$\mathrm{D_{KL}}(U_1, U_2) = \lim_{\gamma\to 1^+} \mathrm{D_{KL}}(X_1, X_2) = \begin{cases} p\log\frac{\sigma_2}{\sigma_1}, & \sigma_1 > \sigma_2, \\ +\infty, & \sigma_1 < \sigma_2, \end{cases}$$

while the K–L divergence between two Laplace distributed r.v. $L_1, L_2 \in \mathscr{L}^p(\mu, \sigma_i^2\mathbb{I}_p)$, $i = 1, 2$, is given by

$$\mathrm{D_{KL}}(L_1, L_2) = \lim_{\gamma\to+\pm\infty} \mathrm{D_{KL}}(X_1, X_2) = p\left(\log\frac{\sigma_2}{\sigma_1} - 1 + \frac{\sigma_1}{\sigma_2}\right).$$

We have already discussed all the well-known entropy type measures and new generalized results have been obtained. We now approach the notion of complexity from a new generalized point of view, as discussed below.

4 Complexity and the Generalized Gaussian

The entropy of a continuous system is defined over a random variable X as the expected value of the *information content*, say $I(X)$, of X, i.e. $\mathrm{H}(X) := \mathrm{E}[I(X)]$. For the usual Shannon entropy (or differential entropy) case, the information content $I(X) = \log f_X$ is adopted, where f_X is the p.d.f. of the r.v. X.

In principle, the entropy can be considered as a measure of the "disorder" of a system. However in applied sciences, the normalized Shannon entropy $\mathrm{H}^* = \mathrm{H}/\max\mathrm{H}$ is usually considered as a measure of "disorder" because H^* is independent of all various states that the system can adopt, [24]. Respectively, the quantity $\Omega = 1 - \mathrm{H}^*$ is considered as a measure of "order". For the estimation of "disorder," information measures play a fundamental role on describing the inner-state or the complexity of a system, see [27] among others. We believe that concepts are useful in Cryptography.

A quantitative measure of complexity with the simplest possible expression is considered to be the "order–disorder" product $\mathrm{K}^{\omega,h}$ given by

$$\mathrm{K}^{\omega,h} = \Omega^\omega \mathrm{H}^{*h} = \mathrm{H}^{*h}(1-\mathrm{H}^*)^\omega = \Omega^\omega(1-\Omega)^h, \quad \omega, h \in \mathbb{R}_+. \tag{77}$$

This is usually called as *simple complexity with "order" power ω and "disorder" power h*.

The above measure $\mathrm{K}^{\omega,h}$, $\omega, h \geq 1$, satisfies the three basic rules of complexity measures. Specifically, we distinguish the following cases:

Rule 1. Vanishing "order" power, $\omega = 0$. Then $\mathrm{K}^{\omega,h} = (\mathrm{H}^*)^h$, i.e. $\mathrm{K}^{0,h}$ is an increasing function of the system's "disorder" H.

Rule 2. Non-vanishing "order" and "disorder" powers, $\omega, h > 0$. Then for "absolute-ordered" or "absolute-disordered" systems the complexity vanishes. Moreover, it adopts a maximum value (with respect to H^*) for an intermediate state $\mathrm{H}^* = h/(\omega + h)$ or $\Omega = \omega/(\omega + h)$, with $\max_{\mathrm{H}^*}\{\mathrm{K}^{\omega,h}\} = h^h \omega^\omega (\omega + h)^{\omega+h}$. In other words the "absolute-complexity" systems are such that their "order" and "disorder" are "balanced," hence $\mathrm{H}^* = h/(\omega + h)$.

Rule 3. Vanishing "disorder" power, $h = 0$. Then $\mathrm{K}^{\omega,h} = \Omega^\omega$, i.e. $\mathrm{K}^{\omega,0}$ is an increasing function of the system's "order" Ω.

The Shiner–Davison–Landsberg (SDL) measure of complexity $\mathrm{K}_{\mathrm{SDL}}$ is an important measure in bio-sciences that satisfies the second rule as it is defined by, [29],

$$\mathrm{K}_{\mathrm{SDL}} = 4\mathrm{K}^{1,1} = 4\mathrm{H}^*(1 - \mathrm{H}^*) = 4\Omega(1 - \Omega). \tag{78}$$

It is important to mention that all the systems with the same degree of "disorder" have the same degree of SDL complexity. Moreover, SDL complexity vanishes for all systems in an equilibrium state and therefore it cannot distinguish between systems with major structural and organizing differences, see also [7, 27].

Now, consider the evaluation of the SDL complexity in a system where its various states are described by a wide range of distributions, such as the univariate γ-ordered Normal distributions. In such a case we may consider the normalized Shannon entropy $\mathrm{H}^*(X_\gamma) := \mathrm{H}(X_\gamma)/\mathrm{H}(Z)$ where $X_\gamma \sim \mathcal{N}_\gamma(\mu, \sigma^2)$ as in (22), and we let $Z \sim \mathcal{N}(\mu, \sigma_Z^2)$ with $\sigma_Z^2 = \mathrm{Var}\, Z$. That is, we adopt for the maximum entropy, with respect to $X_\gamma \sim \mathcal{N}_\gamma$, the Shannon entropy of a normally distributed Z with its variance σ_Z^2 being equal to the variance of X_γ. This is due to the fact that the Normal distribution (included also into the $\mathcal{N}_\gamma(\mu, \sigma^2)$ family for $\gamma = 2$) provides the maximum entropy of every distribution (here $\mathcal{N}_\gamma$) for equally given variances, i.e. $\sigma_Z^2 = \mathrm{Var}\, Z = \mathrm{Var}\, X_\gamma$. Hence, $\max_\gamma\{\mathrm{H}(X_\gamma)\} = \mathrm{H}(X_2) = \mathrm{H}(Z)$.

The use of the above normalized Shannon entropy defines a complexity measure that "characterizes" the family of the γ-ordered Normal distributions as it is obtained in the following Theorem, [17].

Theorem 9. *The SDL complexity of a random variable $X_\gamma \sim \mathcal{N}_\gamma(\mu, \sigma^2)$ is given by*

$$\mathrm{K}_{\mathrm{SDL}}(X_\gamma) = 8\frac{\log\left\{2\sigma\left(\frac{\gamma e}{\gamma-1}\right)^{\frac{\gamma-1}{\gamma}}\Gamma\left(\frac{\gamma-1}{\gamma}+1\right)\right\}}{\log^2\left\{2\pi e\sigma^2\left(\frac{\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}}\frac{\Gamma\left(3\frac{\gamma-1}{\gamma}\right)}{\Gamma\left(\frac{\gamma-1}{\gamma}\right)}\right\}}\log\left\{\frac{\pi}{2}e^{\frac{2-\gamma}{\gamma}}\left(\frac{\gamma}{\gamma-1}\right)^2\frac{\Gamma\left(3\frac{\gamma-1}{\gamma}\right)}{\Gamma^3\left(\frac{\gamma-1}{\gamma}\right)}\right\},$$

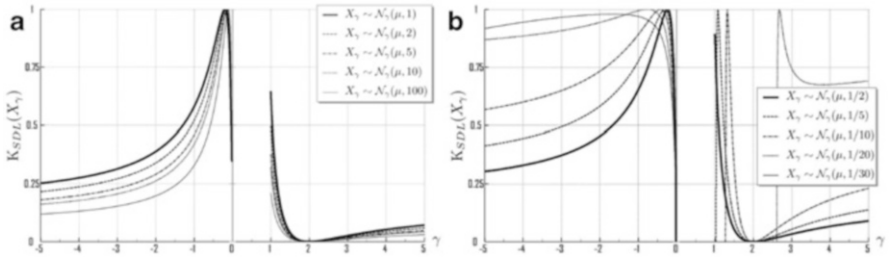


Fig. 4 Graphs of the SDL complexity $\mathrm{KSDL}(X_\gamma)$ along γ, with $X_\gamma \sim \mathcal{N}_\gamma(\mu, \sigma^2)$, for various σ^2 values. (**a**) corresponds to $\sigma \geq 1$ while (**b**) to $\sigma^2 < 1$ values

which vanishes (giving the "absolute-order" or "absolute-disorder" state of a system) for: (a) the normally distributed r.v. X_2, and (b) for scale parameters

$$\sigma = \tfrac{1}{2}\left(\tfrac{\gamma-1}{\gamma e}\right)^{\frac{\gamma-1}{\gamma}}\left[\Gamma\left(\tfrac{\gamma-1}{\gamma}+1\right)\right]^{-1}.$$

Figure 4 illustrates the behavior of the SDL complexity $\mathrm{K}_{\mathrm{SDL}}(X_\gamma)$ with $X_\gamma \sim \mathcal{N}_\gamma(\mu, \sigma^2)$ for various scale parameters σ^2. Notice that, for $\sigma^2 > 1$, depicted in sub-figure (a), the negative-ordered Normals close to 0, i.e. close to the degenerate Dirac distribution (recall Theorem 3), provide the "absolute-complexity" state, i.e. $\mathrm{K}_{\mathrm{SDL}}(X_\gamma) = 1$, of a system, in which their different states described from the γ-ordered Normal distributions. The sub-figure (a) is obtained for $\sigma^2 \geq 1$, while (b) for $\sigma^2 < 1$. Notice, in sub-figure (b), that among all the positive-ordered random variables $X_\gamma \sim \mathcal{N}_{\gamma\geq 0}(\mu, \sigma^2)$ with $\sigma^2 < 1$, the uniformly distributed ones $\gamma = 1$ provide the maximum (but not the absolute) 2-SDL complexity measure.

5 Discussion

In this paper we have provided a concise presentation of a class of generalized Fisher's entropy type information measures, as well as entropy measures, that extend the usual Shannon entropy, such as the α-Shannon entropy and the Rényi entropy. A number of results were stated and proved, and the well-known results were just special cases. These extensions were based on an extra parameter. In the generalized Normal distribution the extra shape parameter γ adjusts fat, or not, tails, while the extra parameter α of the generalized Fisher's entropy type information, or of the generalized entropy, adjusts "optimistic" information measures to better levels. Under this line of thought we approached other entropy type measures as special cases. We believe that these generalizations need further investigation using real data in Cryptography and in other fields. Therefore, these measures were applied on γ-order normally distributed random variables (an exponential-power generalization of the usual Normal distribution) and discussed. A study on a certain form of complexity is also discussed for such random variables.

References

1. Bauer, L.F.: Kryptologie, Methoden and Maximen. Springer, London (1994)
2. Blachman, N.M.: The convolution inequality for entropy powers. IEEE Trans. Inf. Theory **11**(2), 267–271 (1965)
3. Carlen, E.A.: Superadditivity of Fisher's information and logarithmic Sobolev inequalities. J. Funct. Anal. **101**, 194–211 (1991)
4. Cotsiolis, A., Tavoularis, N.K.: On logarithmic Sobolev inequalities for higher order fractional derivatives. C.R. Acad. Sci. Paris Ser. I **340**, 205–208 (2005)
5. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken (2006)
6. Del Pino, M., Dolbeault, J., Gentil, I.: Nonlinear diffusions, hypercontractivity and the optimal L^p-Euclidean logarithmic Sobolev inequality. J. Math. Anal. Appl. **293**(2), 375–388 (2004)
7. Feldman, D.P., Crutchfield, J.P.: Measures of statistical complexity: why? Phys. Lett. A **3**, 244–252 (1988)
8. Ferentinos, K., Papaioannou, T.: New parametric measures of information. Inf. Control **51**, 193–208 (1981)
9. Fisher, R.A.: On the mathematical foundation of theoretical statistics. Philos. Trans. R. Soc. Lond. A **222**, 309–368 (1922)
10. Gómez, E., Gómez–Villegas, M.A., Marin, J.M.: A multivariate generalization of the power exponential family of distributions. Commun. Stat. Theory Methods **27**(3), 589–600 (1998)
11. Goodman, I.R., Kotz, S.: Multivariate θ-generalized normal distributions. J. Multivar. Anal. **3**, 204–219 (1973)
12. Gradshteyn, I.S., Ryzhik, I.M.: Table of Integrals, Series, and Products. Elsevier, Amsterdam (2007)
13. Gross, L.: Logarithm Sobolev inequalities. Am. J. Math. **97**(761), 1061–1083 (1975)
14. Katzan, H. Jr.: The Standard Data Encryption Algorithm. Petrocelli Books, Princeton, NJ (1977)
15. Kitsos, C.P., Tavoularis, N.K.: Logarithmic Sobolev inequalities for information measures. IEEE Trans. Inf. Theory **55**(6), 2554–2561 (2009)
16. Kitsos, C.P., Toulias, T.L.: New information measures for the generalized normal distribution. Information **1**, 13–27 (2010)
17. Kitsos, C.P., Toulias, T.L.: An entropy type measure of complexity. In: Proceedings of COMPSTAT 2012, pp. 403–415 (2012)
18. Kitsos, C.P., Toulias, T.L.: Bounds for the generalized entropy-type information measure. J. Commun. Comput. **9**(1), 56–64 (2012)
19. Kitsos, C.P., Toulias, T.L.: Inequalities for the Fisher's information measures. In: Rassias, T.M. (ed.) Handbook of Functional Equations: Functional Inequalities, Springer Optimization and Its Applications, vol. 95, pp. 281–313. Springer, New York (2014)
20. Kitsos, C.P., Toulias, T.L., Trandafir, C.P.: On the multivariate γ-ordered normal distribution. Far East J. Theor. Stat. **38**(1), 49–73 (2012)
21. Kotz, S.: Multivariate distribution at a cross-road. In: Patil, G.P., Kotz, S., Ord, J.F. (eds.) Statistical Distributions in Scientific Work, vol. 1, pp. 247–270. D. Reidel, Dordrecht (1975)
22. Nadarajah, S.: The Kotz type distribution with applications. Statistics **37**(4), 341–358 (2003)
23. Nadarajah, S.: A generalized normal distribution. J. Appl. Stat. **32**(7), 685–694 (2005)
24. Piasecki, R., Plastino, A.: Entropic descriptor of a complex behaviour. Phys. A **389**(3), 397–407 (2010)
25. Rényi, A.: On measures of entropy and information. In: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 547–561. University of California Press, Berkeley (1961)
26. Rényi, A.: Probability Theory. North-Holland (Ser. Appl. Math. Mech.), Amsterdam (1970)
27. Rosso, O.A., Martin, M.T., Plastino, A.: Brain electrical activity analysis using wavelet-based informational tools (II): Tsallis non-extensivity and complexity measures. Phys. A **320**, 497–511 (2003)

28. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
29. Shiner, J.S., Davison, M., Landsberg, P.T.: Simple measure for complexity. Phys. Rev. E **59**(2), 1459–1464 (1999)
30. Sobolev, S.: On a theorem of functional analysis. AMS Transl. Ser. 2 (English Translation) **34**, 39–68 (1963)
31. Stam, A.J.: Some inequalities satisfied by the quantities of information of Fisher and Shannon. Inf. Control **2**, 255–269 (1959)
32. Stinson, D.R.: Cryptography: Theory and Practice, 3rd edn. CRC Press, Boca Raton (2006)
33. Vajda, I.: $\mathscr{X}^2$-divergence and generalized Fisher's information. In: Transactions of the 6th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, pp. 873–886 (1973)
34. Weissler, F.B.: Logarithmic Sobolev inequalities for the heat-diffusion semigroup. Trans. Am. Math. Soc. **237**, 255–269 (1963)

Maximal and Variational Principles in Vector Spaces

Mihai Turinici

Abstract In Sect. 1, a separable type extension of Ekeland's variational principle (J Math Anal Appl 47:324–353, 1974) is given, in the realm of ordered convergence spaces. The connections with a related statement in Khanh (Bull Acad Pol Sci (Math) 37:33–39, 1989) are then discussed. In Sect. 2, the Brezis–Browder ordering principle (Adv Math 21:355–364, 1976) is used to establish a lot of maximality results in triangular structures due to Pasicki (Nonlinear Anal 74:5678–5684, 2011). Finally, in Sect. 3, some technical aspects of the variational principle due to Bao and Mordukhovich (Control Cyb 36:531–562, 2007) are being analyzed. Further, an extension of this result is proposed, by means of a pseudometric maximal principle in Turinici (Note Mat 28:33–41, 2008).

Keywords: Metric space • Completeness • Ekeland variational principle • Vector space • Convex cone • Separable property • Convergence structure • Ordering • Vector half-metric • Maximal/minimal element • Dependent choice principle • Triangular map • Strong almost completeness • Transitive brezis-browder principle • Domination property • Level-set mapping • Properness • Admissible point

1 Vector EVP on Separable Ordered Convergence Spaces

1.1 Introduction

Let X be a nonempty set and $d : X \times X \to R_+ := [0, \infty[$ be a *metric* over it (in the usual sense); the couple (X, d) will be then referred to as a *metric space*. Further, let $\varphi : X \to R \cup \{\infty\}$ be a *regular* function; i.e.,

(a01) φ is inf-proper ($\mathrm{Dom}(\varphi) \neq \emptyset$ and $\varphi_* := \inf[\varphi(M)] > -\infty$)

(a02) φ is d-lsc on X ($\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \overset{d}{\longrightarrow} x$).

M. Turinici (✉)
"A. Myller" Mathematical Seminar, "A. I. Cuza" University, 700506 Iaşi, Romania
e-mail: mturi@uaic.ro

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_23

The following 1974 statement in Ekeland [22] (referred to as *Ekeland's variational principle*; in short: EVP) is our starting point.

Theorem 1. *Let the precise conditions hold; and (X, d) be complete. Then, for each $u \in \mathrm{Dom}(\varphi)$ there exists $v = v(u) \in \mathrm{Dom}(\varphi)$, with*

$$d(u, v) \leq \varphi(u) - \varphi(v) \text{ (hence } \varphi(u) \geq \varphi(v)) \tag{1}$$

$$d(v, x) > \varphi(v) - \varphi(x), \quad \text{for all } x \in X \setminus \{v\} \tag{2}$$

$$(\forall \varepsilon > 0)\text{: } d(u, v) \leq \varepsilon, \text{ whenever } \varphi(u) \leq \varphi^* + \varepsilon. \tag{3}$$

Concerning the basic theoretical aspects involved here, we stress that, with respect to the Brøndsted (quasi-) order [11]

(a03) $(x, y \in X)$: $x \leq y$ iff $d(x, y) + \varphi(y) \leq \varphi(x)$,

the point $v \in X$ appearing in (2) is *maximal*; so that, (EVP) is nothing but a variant of the Zorn–Bourbaki maximal statement, under the way proposed by Brezis–Browder's ordering principle [10] (in short: BB); hence, (EVP) is deductible from (BB). Concerning the reverse inclusion, note that (BB) is obtainable from the Dependent Choice Principle (in short: DC) due to Bernays [7] and Tarski [51]; wherefrom, (DC) $\Longrightarrow$ (BB) $\Longrightarrow$ (EVP). On the other hand, (EVP) $\Longrightarrow$ (DC); see Brunner [12] and Turinici [62] for details. Summing up, both (BB) and (EVP) are equivalent with (DC); hence, mutually equivalent.

Passing to the practical perspectives of this principle, note that (EVP) found some basic applications to control and optimization, generalized differential calculus, critical point theory, and global analysis; we refer to the 1979 paper by Ekeland [23] for a survey of these. So, it cannot be surprising that, soon after its formulation, many extensions of (EVP) were proposed. For example, the *(pseudo-) metrical* one consists in conditions imposed upon the ambient metric $d(.,.)$ being relaxed. The basic result in this direction, due to Tătaru [52], is essentially founded on Ekeland type techniques; subsequent extensions of it were obtained by Kada, Suzuki, and Takahashi [34]. Since all these are obtainable from (DC), it follows by the above that a deduction of them from (EVP) is possible as well; see Turinici [59] for details. On the other hand, a *functional* extension of (EVP) was carried out in Bao and Khanh [4], as a refinement of some related methods due to Zhong [68]; note that, as precise in Turinici [61], it is nothing but a variant of (EVP). Finally, the *dimensional* way of extension refers to the ambient space (R) of $\varphi(X)$ being substituted by a (topological or not) vector space; an account of these results—including the ones in Chen et al. [15, 16]—is to be found in the 2003 monograph by Goepfert et al. [26, Chap. 3]. It is worth noting that the scalarization type method used there allows us reducing most *sequential* statements in the area to (DC) (hence, ultimately, to (EVP)); see Turinici [55] for details. Unfortunately, this device cannot cover the 1989 variational principle in Khanh [38]; but, for *higher order* versions of (DC) taken as in Wolk [66], this must be possible. To clarify our assertion, the *natural*

setting of Khanh's result is to be re-analyzed. As we shall see, this especially refers to the lattice structure of co-domain space being removed; we refer to Sect. 1.3 for details. Further, some local versions of the obtained facts are indicated in Sect. 1.4. All preliminary concepts and auxiliary statements for getting the results in question are described in Sect. 1.2. Some other aspects will be delineated elsewhere.

1.2 Preliminaries

(A) Let Y be a (real) vector space; and K be a *(convex) cone* of it

$$\alpha K + \beta K \subseteq K, \text{ for } \alpha, \beta \in R_+;$$

supposed to be *pointed* ($K \cap (-K) = \{0\}$). The relation $(\leq_K)$ over Y introduced as

(b01) $(x, y \in Y)$: $x \leq_K y$ iff $y - x \in K$

is reflexive transitive, antisymmetric; hence, an *order*; moreover, it is *compatible* with the linear operations on Y:

$$x \leq_K y \Longrightarrow x + z \leq_K y + z,\ \lambda x \leq_K \lambda y,\quad \forall z \in Y,\ \forall \lambda \in R_+.$$

For simplicity, we shall denote this relation as $(\leq)$. Note that the relation $(\geq)$ over Y introduced as

$(x, y \in Y)$: $x \geq y$ iff $y \leq x$

is again an order; referred to as the *dual* of $(\leq)$. Finally, let $(<)$ stand for the *strict order* attached to $(\leq)$

$(x, y \in Y)$: $x < y$ iff $x \leq y$ and $x \neq y$;

it is *irreflexive* ($x < x$ is false, $\forall x \in X$) and *transitive* ($x < y$ and $y < z$ imply $x < z$), as it can be directly seen. Finally, let $(>)$ stand for its *dual*

$(x, y \in Y)$: $x > y$ iff $y < x$ (or, equivalently: $x > y$ iff $x \geq y$ and $x \neq y$);

clearly, it is a strict order too.

As a rule, the operational concepts to be used are being constructed with the aid of the dual order $(\geq)$ and its attached dual strict order $(>)$; but these may be also viewed as emerging from the initial order $(\leq)$ and its attached strict order $(<)$.

(I) Let $N := \{0, 1, \ldots\}$ stand for the class of *natural* numbers; and $(\leq)$ denote the *standard order* on it, introduced as:

$m \leq n$ iff $m + p = n$, for some $p \in N$.

[Note that $(N, \leq)$ is well ordered; hence, all the more, totally ordered.] Any mapping $x : N \to Y$ will be referred to as a *sequence*; and written as $(x(n); n \geq 0)$ or $(x_n; n \geq 0)$; moreover—when no confusion can arise—we further simplify this notation as $(x(n))$ or (x_n), respectively. By a *subsequence* of (the sequence) $(x_n; n \geq 0)$ we shall mean any sequence $(y_n = x_{i(n)}; n \geq 0)$, where

$(i(n); n \geq 0)$ is strictly ascending (hence: $i(n) \to \infty$ as $n \to \infty$).

(II) Let V be a nonempty part of Y. We say that V is $(\geq)$-*bounded-above* (i.e.: *bounded below*) by $z \in Y$ when $V \geq z$ (i.e.: $x \geq z$, $\forall x \in V$); the class of all such elements will be denoted $\mathrm{ubd}_{(\geq)}(V)(=\mathrm{lbd}(V))$. Further, denote

$$\mathrm{first}_{(\geq)}(V) = \{z \in V; z \geq V\}\ (=\mathrm{last}(V))$$
$$\mathrm{last}_{(\geq)}(V) = \{z \in V; V \geq z\}\ (=\mathrm{first}(V)).$$

These sets appear as singletons, whenever they are nonempty; in this case, their uniquely determined member (in V) will be called a $(\geq)$-*first-element* (i.e.: *last* element) and $(\geq)$-*last-element* (i.e.: *first* element) of V, respectively. Finally, call the point $z \in \mathrm{ubd}_{(\geq)}(V)(=\mathrm{lbd}(V))$, a $(\geq)$-*supremum* (i.e.: *infimum*) of V, when $[V \geq w \Longrightarrow z \geq w]$. The class $\sup_{(\geq)}(V) = \inf(V)$ of all these is either empty or a singleton, $\{z\}$; in this case, we write $\{z\} = \sup_{(\geq)}(V) = \inf(V)$ as $z = \sup_{(\geq)}(V) = \inf(V)$. The corresponding form of these conventions with $(\leq)$ in place of $(\geq)$ is to be introduced in a dual manner.

(III) We say that $L \subseteq Y$ is a $(\geq)$-*super-chain* (in short: $(\geq)$-*schain*) of Y provided each nonempty part of L has a $(\geq)$-first element; the class of all these will be denoted as $\mathrm{schain}_{(\geq)}(Y)$. Given two objects $P, Q \in \mathrm{schain}_{(\geq)}(Y)$, define

(b02) $P \sqsupseteq Q$ (referred to as: Q is $(\geq)$-*cofinal* in P), provided
$(P \supseteq Q)$ and $(\forall x \in P, \exists y \in Q$, with $x \geq y)$.

This relation is reflexive, transitive, and antisymmetric—hence, an ordering—in $\mathrm{schain}_{(\geq)}(Y)$; since the verification is immediate, we omit the details.

Now, call the (nonempty) $P \in \mathrm{schain}_{(\geq)}(Y)$, $(\geq)$-*chain-separable*, when

$$P \sqsupseteq \{x_n; n \geq 0\}, \text{ for some } (\geq)\text{-ascending sequence } (x_n; n \geq 0);$$

or, equivalently (passing to the initial order)

(b03) $P \sqsupseteq \{x_n; n \geq 0\}$, for some descending sequence $(x_n; n \geq 0)$.

[Note that, the definition is consistent, in the sense of $Q := \{x_n; n \geq 0\}$ being an element of $\mathrm{schain}_{(\geq)}(Y)$]. Clearly, any $P \in \mathrm{schain}_{(\geq)}(Y)$ with $\mathrm{last}_{(\geq)}(P) \neq \emptyset$ fulfills such a requirement; just take $(x_n = \mathrm{last}_{(\geq)}(P); n \geq 0)$. So, the question to be posed is that of discussing the underlying condition when $\mathrm{last}_{(\geq)}(P) = \emptyset$. In this case, we claim that the $(\geq)$-ascending (i.e.: descending) property of our sequence $(x_n; n \geq 0)$ may be taken in a strict sense. For, if $(x_n; n \geq 0)$ fulfills

$$(\exists k \geq 0)\colon x_n = x_k, \text{ for all } n \geq k \text{ (or, equivalently: for all } n > k)$$

then $x_k = \mathrm{last}_{(\geq)}(P)$; contradiction. Hence, necessarily,

$$\text{for each } k \geq 0, \text{ there exists } n > k, \text{ such that } x_k > x_n;$$

wherefrom, the descending sequence $(x_n; n \geq 0)$ must admit a strictly descending subsequence $(y_n; n \geq 0)$; and our claim follows via $\{x_n; n \geq 0\} \sqsupseteq \{y_n; n \geq 0\}$. Hence, $P \in \mathrm{schain}_{(\geq)}(Y)$ with $\mathrm{last}_{(\geq)}(P) = \emptyset$ is $(\geq)$-chain-separable iff

(b04) $P \sqsupseteq \{x_n; n \geq 0\}$, for some strictly descending sequence $(x_n; n \geq 0)$.

It remains now to determine sufficient global conditions for such a property. Let us say that the nonempty subset $C \subseteq Y$ is $(\geq)$-*chain*, provided

for each $x, y \in U$: either $x \geq y$ or $y \geq x$;

this will be also referred to as: C is $(\geq)$-*totally ordered*. The dual concept of $(\leq)$-*chain* (or: $(\leq)$-*totally ordered set*) is identical with the above one; so, we may talk about a *chain* (or: *totally ordered set*) in this case. Further, let us say that Y is $(\geq)$-*complete* when (cf. Peressini [46, Chap. 1, Sect. 1.7])

(D=chain, $\mathrm{ubd}_{(\geq)}(D) \neq \emptyset$) $\Longrightarrow$ $\sup_{(\geq)}(D) \neq \emptyset$;

or, equivalently (passing to the initial order)

(b05) (D=totally ordered, bounded from below) $\Longrightarrow$ $\inf(D)$ exists.

Likewise, call Y, $(\geq)$-*separable* if (cf. Peressini [46, Chap. 1, Sect. 5.18])

(G=chain, $\sup_{(\geq)}(G) \neq \emptyset$) $\Longrightarrow$ there exists a sequence $(x_n; n \geq 0)$ in G such that $\sup_{(\geq)}(G) = \sup_{(\geq)}(\{x_n; n \geq 0\})$;

or, equivalently (passing to the initial order)

(b06) (G=totally ordered, $\inf(G) \neq \emptyset$) $\Longrightarrow$ there exists a sequence $(x_n; n \geq 0)$ in G such that $\inf(G) = \inf(\{x_n; n \geq 0\})$.

Putting these together, we therefore get the following practical statement.

Proposition 1. *Suppose that Y is $(\geq)$-complete and $(\geq)$-separable. Then, each (nonempty) bounded from below part $P \in \mathrm{schain}_{(\geq)}(Y)$ is $(\geq)$-chain-separable.*

Proof. Let the nonempty $P \in \mathrm{schain}_{(\geq)}(Y)$ be bounded from below. By a previous remark, the case of $\mathrm{last}_{(\geq)}(P) \neq \emptyset$ is clear; so, there is no loss in generality if one assumes that $\mathrm{last}_{(\geq)}(P) = \emptyset$. In particular, P is totally ordered ($x, y \in P \Longrightarrow x \leq y$ or $y \leq x$); so, as Y is $(\geq)$-complete, $\inf(P)$ exists. Combining with Y being $(\geq)$-separable, one derives

there is a sequence $(x_n; n \geq 0)$ in P such that $\inf(P) = \inf(\{x_n; n \geq 0\})$.

We now claim that this last property is equivalent with

(b07) there must be a descending sequence $(y_n; n \geq 0)$ in P such that $\inf(P) = \inf(\{y_n; n \geq 0\})$.

In fact, let us construct the sequence (in P)

$$y_n := \inf\{x_0, \ldots, x_n\} = \min\{x_0, \ldots, x_n\},\ n \geq 0.$$

As $(y_n; n \geq 0)$ consists of elements taken from the original sequence $(x_n; n \geq 0)$, we have (by the very definition of infimum)

$$y_n \geq \inf(\{x_n; n \geq 0\}),\ \text{for all } n;$$

whence $D := \{y_n; n \geq 0\}$ is bounded from below. Further, as $(y_n; n \geq 0)$ is descending, D is totally ordered; wherefrom (as Y is $(\geq)$-complete)

$$(\exists)\ \inf(D) = \inf(\{y_n; n \geq 0\}) \geq \inf(\{x_n; n \geq 0\}).$$

Finally, as ($x_n \geq y_n$, for all n), we derive

$$\inf(\{x_n; n \geq 0\}) \geq \inf(\{y_n; n \geq 0\});$$

whence (combining with the above) $\inf(\{y_n; n \geq 0\}) = \inf(\{x_n; n \geq 0\})$. This, along with a previous relation involving $(x_n; n \geq 0)$, gives the desired fact.

Finally, note that a slight extension of the above result is to be reached when the $(\geq)$-complete and $(\geq)$-separable conditions upon Y would be formulated with respect to $(\geq)$-schains (in place of chains). However, for the developments below, the provided version of this criterion will suffice.

1.3 Main Results

With this information, we may return to the questions of introductory part.

(A) Let Y be a real vector space; and K be some pointed convex cone of Y; the associated order $(\leq_K)$ will be denoted as $(\leq)$, for simplicity. Assume that

(c01) Y is $(\geq)$-complete and $(\geq)$-separable (see above).

(B) Let X be a nonempty set. Denote by $\mathscr{S}(X)$ the class of all sequences (x_n) in X. By a (sequential) *convergence structure* on X we mean, as in Kasahara [36], any part $(\to)$ of $\mathscr{S}(X) \times X$ with the properties

(cs-1) $(x_n = x, \forall n \geq 0)$ implies $((x_n); x) \in (\to)$
(cs-2) $((x_n); x) \in (\to)$ implies $((y_n); x) \in (\to)$,
for each subsequence (y_n) of (x_n).

In this case, $((x_n); x) \in (\to)$ writes $x_n \to x$; and reads: x is the *limit* of (x_n). The set of all such x is denoted $\lim(x_n)$; when it is nonempty, we say that (x_n) is *convergent* (modulo $(\to)$).

(C) Let $(X, \to)$ be a convergence structure. According to Khanh [38], any map $d : X \times X \to K$ with

(vhm-1) (reflexive sufficient) $x = y \Longleftrightarrow d(x, y) = 0$
(vhm-2) (triangular) $d(x, z) \leq d(x, y) + d(y, z), \ \forall x, y, z \in X$

will be called a *vector half-metric* on X. Fix in the following such an object; as well as some function $\varphi : X \to Y$.

Given a sequence (x_n) in X, call it *d-strongly asymptotic* (in short: *d-strasy*), if

(str) $\sum_{i=0}^{m} d(x_i, x_{i+1}) \leq k$, for all $m > 0$ and some $k \in K$.

The following condition is to be considered

(c02) X is *φ-descending $(d, \to)$-complete*: each d-strasy sequence (x_n) in X with $(\varphi(x_n))$=descending, converges (modulo $(\to)$).

Given another function $\psi : X \to Y$, call it *φ-descending $(\to)$-lsc*, in case

(lsc-d) $(\psi(x_n) \le t, \forall n)$, $x_n \to x \in X$ and $(\varphi(x_n))$=descending imply $\psi(x) \le t$.

In particular, this holds whenever ψ is *(standard)* $(\to)$*-lsc*, in the sense

(lsc-st) $(\psi(x_n) \le t, \forall n)$ and $x_n \to x \in X$ imply $\psi(x) \le t$.

(D) Let $(\succeq)$ be an ordering on X. Remember that $z \in X$ is $(\succeq)$*-maximal*, provided

$X(x, \succeq) = \{z\}$; i.e.: $z \succeq w \in X$ implies $z = w$;

the class of all these will be denoted as $\max(X, \succeq)$. The following maximal principle is basic for our developments.

Proposition 2. *Suppose that one of the following conditions holds:*

(zb-1) X is $(\succeq)$-chain-inductive: each $(\succeq)$-chain is $(\succeq)$-bounded-above
(zb-2) X is $(\succeq)$-schain-inductive: each $(\succeq)$-schain is $(\succeq)$-bounded-above.

Then, for each $u \in X$, there exists a $(\succeq)$-maximal $v \in X$ with $u \succeq v$.

The first variant of this statement (based on chains) belongs to Zorn [71]; see also Kelley [37, Chap. 0]. The second (refined) variant of the same (based on $(\succeq)$-schains) is due to Bourbaki [9]. So, it is natural that Proposition 2 be referred to as *Zorn–Bourbaki maximal principle*. For equivalent versions of this one we refer to Moore [41, Chap. 4, Sect. 4.4] and the references therein.

Our main result in this exposition is

Theorem 2. *Let the data (Y, K), $(X, \to)$ be taken according to the general conditions above. Further, let the vector half-metric $d(.,.)$ and the function $\varphi : X \to Y$ be such that X is φ-descending $(d, \to)$-complete, and*

(c03) $\varphi(X)$ is bounded below: $\varphi(X) \subseteq \tilde{y} + K$, for some $\tilde{y} \in Y$
(c04) $t \mapsto d(x, t) + \varphi(t)$ is φ-descending $(\to)$-lsc, for each $x \in X$.

Then, for each $u \in X$, there exists $v \in X$ with

$$d(u, v) \le \varphi(u) - \varphi(v) \ \ (\textit{hence } \varphi(u) \ge \varphi(v)) \tag{4}$$

$$d(v, t) \le \varphi(v) - \varphi(t) \textit{ is impossible, for each } t \in X \setminus \{v\}. \tag{5}$$

Proof. Define the relation over X

(c05) $x \succeq y$ iff $d(x, y) \le \varphi(x) - \varphi(y)$.

Clearly, $(\succeq)$ is reflexive, transitive, and antisymmetric; hence, it is an order on X. We show that $(X, \succeq)$ fulfills conditions of the Zorn–Bourbaki maximal principle; and, from this, all is clear. There are two steps to be passed.

Step 1. We show that, under these conditions, X is $(\succeq)$-schain-inductive (see above). Let A be some $(\succeq)$-schain in X; that is,

for each (nonempty) $A_1 \subseteq A$, there exists $a_1 \in A_1$, such that $a_1 \succeq A_1$.

If $\text{last}_{(\succeq)}(A) \neq \emptyset$, A is $(\succeq)$-bounded above (by this element); so, without loss, one may assume that

(c06) $\text{last}_{(\succeq)}(A) = \emptyset$: for each $u \in A$ there exists $v \in A$ such that $u \succ v$.

By the very definition above, A is totally ordered (modulo $(\succeq)$); so that (as $d(.,.)$ is reflexive sufficient and $(\leq)$ is an order)

$$(x, y \in A): \quad x \succ y \text{ iff } \varphi(x) > \varphi(y). \tag{6}$$

(Here, $(\succ)$ and $(>)$ are the strict orders attached to $(\succeq)$ and $(\geq)$, respectively.) In other words, φ is an order isomorphism between $(A, \succeq)$ and $(E := \varphi(A), \geq)$; so, necessarily, $\text{last}_{(\geq)}(E) = \text{first}(E) = \emptyset$. In addition, (as $E \subseteq \varphi(X)$), E is bounded from below by (c03); hence (as Y is $(\geq)$-complete),

$$\inf(E) \text{ exists in } Y; \text{ and (as first}(E) = \emptyset), E > \inf(E).$$

Combining with Y being $(\geq)$-separable, one derives (from a previous auxiliary fact) that E is $(\geq)$-chain-separable; i.e.,

there exists a sequence (x_n) in A such that $(\varphi(x_n))$ is descending in E and $\inf(\{\varphi(x_n); n \geq 0\}) = \inf(E)$.

Note that the former of these properties tells us (via (6) above) that (x_n) is ascending (modulo $(\succeq)$) in A

$$n < m \Longrightarrow (0 \leq) d(x_n, x_m) \leq \varphi(x_n) - \varphi(x_m). \tag{7}$$

And, from the latter one, $\{\varphi(x_n); n \geq 0\}$ is $(\geq)$-cofinal in E; wherefrom, $\{x_n; n \geq 0\}$ is $(\succeq)$-cofinal in A. By (7), we derive (via (c03)), for all $n > 0$,

$$\sum_{i=0}^{n} d(x_i, x_{i+1}) \leq \varphi(x_0) - \varphi(x_{n+1}) \leq \varphi(x_0) - \tilde{y}.$$

The sequence (x_n) is therefore d-strasy; so (as X is φ-descending $(\to)$-complete), $x_n \to x^*$ for some $x^* \in X$. Now, again by (7), one gets for each n,

$$d(x_n, y_j) + \varphi(y_j) \leq \varphi(x_n), \quad \forall j;$$

where, for simplicity, we denoted $(y_j := x_{n+j}; j \geq 0)$. Passing to limit upon j and taking (c04) into account one derives, for all n,

$$d(x_n, x^*) + \varphi(x^*) \leq \varphi(x_n); \quad \text{that is}: \; x_n \succeq x^*;$$

so that x^* is an upper bound (modulo $(\succeq)$) of (x_n); wherefrom (by the $(\succeq)$-cofinality of this sequence), x^* is an upper bound (modulo $(\succeq)$) of A.

Step 2. By the Zorn–Bourbaki maximal principle, it follows that, for each $u \in X$ there exists $v \in X$ with

(c1) $u \succeq v$, **(c2)** v is $(\succeq)$-maximal in X ($v \succeq x$ is impossible, $\forall x \in X \setminus \{v\}$).

This shows that v is just the desired element; and the proof is complete.

As a direct consequence of this, the following fixed point result is available.

Theorem 3. *Let the conditions of Theorem 2 be fulfilled and the multivalued map $T : X \to 2^X$ be such that*

(c07) for each $x \in X$ there exists $y \in Tx$ with $d(x, y) \leq \varphi(x) - \varphi(y)$.

Then, for each $u \in X$, there exists $v \in X$ with the properties (4) and (5), fulfilling

$$v \in Tv \text{ (i.e.: } v \text{ is fixed under } T). \tag{8}$$

Proof. By (c07), the point $v \in X$ given by Theorem 2 has the property

$$d(v, t) \leq \varphi(v) - \varphi(t) \text{ (i.e., } v \succeq t), \text{ for some } t \in Tv.$$

This, along with (5), yields $v = t \in Tv$; wherefrom, all is clear.

In particular, when $Y = R$ and $d(.,.)$ is a (standard) metric on X, Theorem 2 is nothing else than (EVP); and Theorem 3 is just the Caristi–Kirk fixed point result [13]. Further aspects may be found in Nemeth [43].

1.4 Local Versions

In the following, a local version of Theorem 2 is given, so as to compare it with the last part of Theorem 1.

(A) Let Y be a real vector space; and K be some pointed convex cone of Y; the associated order $(\leq_K)$ will be denoted as $(\leq)$, for simplicity.

Let $H \subseteq Y$ be a nonempty subset. Remember that the nonempty part $C \subseteq H$ is $(\geq)$-*chain*, provided

for each $x, y \in C$: either $x \geq y$ or $y \geq x$;

this will be also referred to as: C is $(\geq)$-*totally ordered*. The dual concept of $(\leq)$-*chain* (or: $(\leq)$-*totally ordered set*) is identical with the above one; so, we may talk about a *chain* (or: *totally ordered set*) in this case. Let chain(H) stand for the class of all (nonempty) chains in H. As before, we may introduce a maximality concept over (chain$(H), \subseteq$), as follows: call $C \in$ chain(H), *maximal (modulo* $(\subseteq)$*)*, when

(d01) $C \subseteq E \in$ chain(H) implies $C = E$;

the class of all such elements (if any) will be denoted as max(chain$(H), \subseteq$). A basic existence result about the class in question is deductible from the Zorn–Bourbaki

maximal principle. Namely, the following Hausdorff–Kuratowski maximal principle is available (cf. Schechter [47, Chap. 6, Sect. 6.20]):

Proposition 3. *The structure* $(\mathrm{chain}(H),\ (\subseteq))$ *is inductive; i.e.,*

for each totally ordered part $\mathscr{D}$ *of* $\mathrm{chain}(H)$*, we have* $G := \cup\mathscr{D} \in \mathrm{chain}(H)$*; (wherefrom,* $\mathscr{D}$ *is bounded above (modulo* $(\subseteq)$*) by* G *in* $\mathrm{chain}(H)$*).*

Hence (by the Zorn–Bourbaki maximal principle) the following property holds: for each $A \in \mathrm{chain}(H)$*, there exists some* $U \in \mathrm{chain}(H)$*, with*
(i) $A \subseteq U$, **(ii)** $U \subseteq V \in \mathrm{chain}(H) \Longrightarrow U = V$ *(i.e.:* U *is* $(\subseteq)$*-maximal).*

This result allows us to get a useful representation of $(\geq)$-maximal (i.e.: minimal) elements in $(H, \leq)$. Precisely, call $z \in H$, $(\geq)$-*maximal* (i.e.: *minimal*), if

(d02) $w \in H, z \geq w$ implies $z = w$.

the class of all these will be denoted as $\max(H, \geq)$ (resp., $\min(H, \leq)$).

Proposition 4. *Let the above conventions be in use. Then,* $z \in H$ *is* $(\geq)$*-maximal (i.e.: minimal), if and only if it may be written as*

$$z = \mathrm{first}(U)\text{, for some } (\subseteq)\text{-maximal } U \in \mathrm{chain}(H).$$

The verification is immediate; we do not give details.

(B) So far, the obtained facts are valid in a general context relative to (Y, K). But, for the next developments, further regularity conditions upon these data must be imposed. Namely, assume that

(d03) Y is $(\geq)$-complete:
if $D \subseteq Y$ is totally ordered and bounded from below, then $\inf(D)$ exists.

Let H be a nonempty bounded from below part of Y. From the above completeness assumption, it follows that

$\inf(U)$ exists, for each [$(\subseteq)$-maximal or not] $U \in \mathrm{chain}(H)$.

Denote for simplicity

$$\mathrm{amin}(H, \leq) = \{\inf(U);\ U \in \max(\mathrm{chain}(H), \subseteq)\};$$

each element in this subset will be referred to as an *almost minimal* element of H. By the statement above, it follows that

$$\min(H, \leq) \subseteq \mathrm{amin}(H, \leq) \ \text{ [whenever } H \text{ is bounded below]}; \tag{9}$$

i.e.: each minimal element of H is almost minimal too. The reciprocal is not in general true; to verify this, it will suffice noting that $\min(H, \leq)$ may be sometimes empty, whereas $\mathrm{amin}(H, \leq)$ is always nonempty (under the previous choice of H).

The following "almost" Zorn–Bourbaki principle is now available.

Proposition 5. *Let* Y *be* $(\geq)$*-complete; and the nonempty subset* H *of* Y *be bounded from below. Then, for each* $A \in \mathrm{chain}(H)$ *there exists* $u \in \mathrm{amin}(H, \leq)$*, with* $A \geq u$*.*

Proof. By the Hausdorff–Kuratowski maximal principle, there exists a maximal (modulo $(\subseteq)$) chain U in H with $A \subseteq U$. On the other hand, by the imposed upon H condition, $u := \inf(U)$ exists; and this yields $A \geq u$. The proof is complete.

Having these precise, the following local version of Theorem 2 is available.

Theorem 4. *Let the conditions of Theorem 2 be in force; as well as (d03). Then, for each $u \in X$, there exists $v \in X$ fulfilling (4) and (5). In addition, for the obtained $v \in X$ there exists $q \in \operatorname{amin}(\varphi(X), \leq)$, with $\varphi(v) \geq q$; hence*

$$d(u, v) \leq r, \ \text{when } \varphi(u) \leq q + r. \tag{10}$$

Proof. Let V be a maximal (modulo $(\subseteq)$) chain in $\varphi(X)$, including the chain $E := \{\varphi(u), \varphi(v)\}$. By the imposed conditions, $q := \inf(V)$ exists (in Y); with, in addition, $\varphi(u) \geq \varphi(v) \geq q$. This, along with $d(u, v) \leq \varphi(u) - \varphi(v)$, gives (10) and concludes the argument.

The obtained result may be compared with a similar one due to Khanh [38]. However, we must say that, under these requirements, $\min(\varphi(X), \leq)$ may be empty; so, the quoted result is not in general retainable. On the contrary, $\operatorname{amin}(\varphi(X), \leq)$ is nonempty; i.e., this conclusion is retainable under the precise conditions. Note finally that a corresponding variant of Theorem 2 may be formulated in the setting of (10); we do not give details. Further aspects may be found in Bao and Mordukhovich [5]; see also Turinici [54].

2 Maximality Principles in Triangular Structures

2.1 Introduction

Let X be a nonempty set; and $d : X \times X \to R_+ := [0, \infty[$ be a *metric* over it (in the usual sense); the couple (X, d) will be then referred to as a *metric space*. Further, let $\varphi : X \to R$ be a function with

(a01) φ is d-lsc on X: $\liminf_n \varphi(x_n) \geq \varphi(x)$ whenever $x_n \xrightarrow{d} x$
(a02) φ is bounded from below on X: $\inf\{\varphi(x); x \in X\} > -\infty$.

The following 1979 statement in Ekeland [23] (referred to as Ekeland's variational principle; in short: EVP) is well known.

Theorem 5. *Let the function $\varphi : X \to R$ be d-lsc and bounded from below on X. In addition, let (X, d) be complete. Then, for each $u \in X$, there exists $v = v(u) \in X$, with the properties*

$$d(u, v) \leq \varphi(u) - \varphi(v) \ (\text{hence } \varphi(u) \geq \varphi(v)) \tag{11}$$

$$d(v,x) > \varphi(v) - \varphi(x), \quad \textit{for all } x \in X \setminus \{v\}. \tag{12}$$

This principle found some basic applications to control and optimization, generalized differential calculus, critical point theory, and global analysis; we refer to Hyers et al. [30, Chap. 5] for a survey of these. So, it cannot be surprising that, soon after, many extensions of (EVP) were proposed. For example, the *abstract* (*order*) one starts from the fact that, with respect to the Brøndsted order [11]

(a03) $(x, y \in X)$: $x \leq y$ iff $d(x,y) \leq \varphi(x) - \varphi(y)$

the point $v \in X$ appearing in (12) is *maximal*; so that, Theorem 5 is nothing but a variant of the Zorn Maximal Principle (cf. Bourbaki [9]), in the way described by the Brezis–Browder ordering principle [10] (in short: BB). The (*pseudo*) *metrical* one consists in conditions imposed to the ambient metric over X being relaxed; a basic result of this type may be found in the 1996 paper by Kada, Suzuki, and Takahashi [34]. Further, we must add to the above list the 1997 "functional" extension of (EVP) obtained by Zhong [68]; this, essentially, consists in the variational conclusion (12) being relaxed as

$$b(d(a,v))d(v,x) > \varphi(v) - \varphi(x), \quad \text{for each } x \in X \setminus \{v\}; \tag{13}$$

where a is an element of X, and $b : R_+ \to R_+$ is a *normal* function:

(a04) b is decreasing, $b(R_+) \subseteq R_+^0 :=]0, \infty[$ and $\int_0^\infty b(\tau)d\tau = \infty$;

cf. Suzuki [48] and Turinici [65]. Finally, the *dimensional* way of extension refers to the ambient space (R) of $\varphi(X)$ being substituted by a (topological or not) vector space. A pioneering work in this direction is the 1983 Pareto efficiency statement due to Isac [31]. Further, the topological vector space realm is discussed in the papers by Nemeth [43] and Turinici [54]. A "product" type version of these results is to be found in Goepfert, Tammer, and Zălinescu [25]; see also Isac [32].

Now, the natural question to be posed is that of these extensions being or not effective. Some partial (negative) answers were stated in Turinici [59]; see also Bao and Khanh [4]. According to these, the metrical and (sequential type) dimensional extensions of (EVP) are obtainable from either (EVP) or (BB), via straightforward techniques. Concerning the question of (BB) (and its subsequent extensions) being reducible to (EVP), the basic tool for solving it is the Dependent Choice Principle (in short: DC) due, independently, to Bernays [7] and Tarski [51]. Precisely, note that, by the very Ekeland's argument, (DC) $\Longrightarrow$ (EVP); moreover, as shown in Brunner [12], (EVP) $\Longrightarrow$ (DC). Hence, any maximal/variational result—(MP) say—with (DC) $\Longrightarrow$ (MP) $\Longrightarrow$ (EVP) is logically equivalent with both (DC) and (EVP); see Turinici [62] for details. In particular, this is the case with many extensions of (EVP) and/or (BB). But (cf. Turinici [63]), the conclusion is also true for the "smooth" extension of (EVP) due to Borwein and Preiss [8] and the related contributions in Li and Shi [39]; see also Bejancu [6].

Recently, some new maximal principles in the area were obtained by Pasicki [45], for *triangular* maps. By the remarks above, we may ask whether these enter as

well in this reduction scheme. It is our aim in the present exposition to show that the results in question are obtainable from certain "transitive" type ordering principles like in Turinici [56]; whence, they are deductible from (DC). Moreover, these results include both (EVP) and (BB); so, by the above, they are nothing else than equivalent versions of (EVP) or (BB). The obtained conclusion may have a theoretical impact upon such statements; but, from the equilibrium points perspective, these may be useful tools; cf. Turinici [65].

2.2 Preliminaries

Throughout this exposition, the axiomatic system in use is Zermelo-Fraenkel's (abbreviated: ZF), as described by Cohen [17, Chap. 2]. The notations and basic facts to be considered in this system are standard. Some important ones are described below.

(A) Let X be a nonempty set. By a *relation* over X, we mean any nonempty part $\mathscr{R} \subseteq X \times X$. For simplicity, we sometimes write $(x, y) \in \mathscr{R}$ as $x\mathscr{R}y$. Note that $\mathscr{R}$ may be regarded as a mapping between X and 2^X (=the class of all subsets in X). In fact, denote for $x \in X$:

$$X(x, \mathscr{R}) = \{y \in X; x\mathscr{R}y\} \text{ (the } \textit{section} \text{ of } \mathscr{R} \text{ through } x);$$

then, the desired mapping representation is $[\mathscr{R}(x) = X(x, \mathscr{R}),\ x \in X]$. A basic example of such object is

$$\mathscr{I} = \{(x, x); x \in X\} \text{ [the } \textit{identical relation} \text{ over } X].$$

Given the relations $\mathscr{R}$, $\mathscr{S}$ over X, define their *product* $\mathscr{R} \circ \mathscr{S}$ as

$$(x, z) \in \mathscr{R} \circ \mathscr{S}, \text{ if there exists } y \in X \text{ with } (x, y) \in \mathscr{R},\ (y, z) \in \mathscr{S}.$$

Also, for each relation $\mathscr{R}$ in X, denote

$$\mathscr{R}^{-1} = \{(x, y) \in X \times X; (y, x) \in \mathscr{R}\} \text{ (the } \textit{inverse} \text{ of } \mathscr{R}).$$

Finally, given the relations $\mathscr{R}$ and $\mathscr{S}$ on X, let us say that $\mathscr{R}$ is *coarser* than $\mathscr{S}$ (or, equivalently: $\mathscr{S}$ is *finer* than $\mathscr{R}$), provided

$$\mathscr{R} \subseteq \mathscr{S}; \text{ i.e.: } x\mathscr{R}y \text{ implies } x\mathscr{S}y.$$

Given a relation $\mathscr{R}$ on X, the following properties are to be discussed here:

(P1) $\mathscr{R}$ is reflexive: $\mathscr{I} \subseteq \mathscr{R}$
(P2) $\mathscr{R}$ is irreflexive: $\mathscr{R} \cap \mathscr{I} = \emptyset$
(P3) $\mathscr{R}$ is transitive: $\mathscr{R} \circ \mathscr{R} \subseteq \mathscr{R}$
(P4) $\mathscr{R}$ is symmetric: $\mathscr{R}^{-1} = \mathscr{R}$
(P5) $\mathscr{R}$ is antisymmetric: $\mathscr{R}^{-1} \cap \mathscr{R} \subseteq \mathscr{I}$.

This yields the classes of relations to be used; the following ones are important for our developments:

(C0) $\mathscr{R}$ is *trivial* (i.e.: $\mathscr{R} = X \times X$)

(C1) $\mathscr{R}$ is a *(partial) order* (reflexive, transitive, antisymmetric)
(C2) $\mathscr{R}$ is a *strict order* (irreflexive and transitive)
(C3) $\mathscr{R}$ is a *quasi-order* (reflexive and transitive)
(C4) $\mathscr{R}$ is an *equivalence* (reflexive, transitive, symmetric).

A basic ordered structure is $(N, \leq)$; here, $N = \{0, 1, \ldots\}$ is the set of natural numbers and (the partial order) $(\leq)$ is defined as

$m \leq n$ iff $m + p = n$, for some $p \in N$.

In fact, $(N, \leq)$ is well ordered; i.e.: any (nonempty) subset of N has a first element. By a *sequence* in X, we mean any mapping $x : N \to X$. For simplicity reasons, it will be useful to denote it as $(x(n); n \geq 0)$, or $(x_n; n \geq 0)$; moreover, when no confusion can arise, we further simplify this notation as $(x(n))$ or (x_n), respectively. Also, any sequence $(y_n := x_{i(n)}; n \geq 0)$ with

$(i(n); n \geq 0)$ is strictly ascending [hence, $i(n) \to \infty$ as $n \to \infty$]

will be referred to as a *subsequence* of $(x_n; n \geq 0)$.

(B) Remember that, an outstanding part of (ZF) is the *Axiom of Choice* (abbreviated: AC), which, in a convenient manner, may be written as

(AC) For each nonempty set X, there exists a (selective) function $f : (2)^X \to X$, with $f(Y) \in Y$, for each $Y \in (2)^X$.

(Here, $(2)^X$ stands for the class of all nonempty elements in 2^X). There are many logical equivalents of (AC); see, for instance, Moore [41, Appendix 2]. A basic group of these refers to (partially) ordered structures. Some preliminaries are needed. Let $(X, \leq)$ be a (partially) ordered set. By a $(\leq)$-*chain* in X, we mean any part $C \in (2)^X$, with

C is totally ordered (modulo $(\leq)$):
for each $x, y \in C$, we have either $x \leq y$ or $y \leq x$.

The family of all $(\leq)$-chains in X will be denoted as $\mathrm{chain}(X, \leq)$; it may be viewed as a partially ordered structure, with respect to the usual inclusion $(\subseteq)$. Given the nonempty part Y of X, let us say that $u \in X$ is an *upper bound* of it, provided

$Y \leq u$; (in the sense: $y \leq u$, $\forall y \in Y$);

the class of all these will be denoted as $\mathrm{ubd}(Y)$. Call $(X, \leq)$, *inductive* provided:

$\mathrm{ubd}(C)$ is nonempty, for each $(\leq)$-chain C in X;

note that $(\mathrm{chain}(X, \leq), \subseteq)$ is endowed with such a property. Finally, call the point $z \in X$, *maximal (modulo* $(\leq)$*)*, provided

$X(z, \leq) = \{z\}$; or, equivalently: $X(z, <) = \emptyset$.

(Here, $(<)$ is the *strict order* attached to $(\leq)$, as

$x < y$ iff $x \leq y$ and $x \neq y$.

As precise, this means that $(<)$ is irreflexive and transitive; we do not give details.) Denote the class of all such elements as $\max(X, \leq)$. We then say that $(X, \leq)$ is a *Zorn ordered structure*, provided

> for each $x \in X$, there exists a maximal (modulo $(\leq)$) element $z \in X$, with the property $x \leq z$.

Returning to the general setting, we stress that a basic equivalent form of (AC) is the *Zorn Maximal Principle* (in short: ZMP), expressed as

> (ZMP) If the (partially) ordered structure $(X, \leq)$ is inductive, then, necessarily, $(X, \leq)$ is Zorn ordered.

(For an outline of proof, we refer to Bourbaki [9].) As precise, the ordered structure $(\text{chain}(X, \leq), \subseteq)$ is inductive. This, via (ZMP), yields the so-called *Hausdorff–Kuratowski Maximal Principle*:

> (HKMP) The partially ordered structure $(\text{chain}(X, \leq), \subseteq)$ is a Zorn one: for each $(\leq)$-chain C in X, there exists a maximal (modulo $(\subseteq)$) $(\leq)$-chain D in X, with $C \subseteq D$;

or, in a simplified way: any $(\leq)$-chain is included in a maximal $(\leq)$-chain. The converse implication ((HKMP) $\Longrightarrow$ (ZMP)) is also true (cf. Kelley [37, Chap. 0]); hence, these maximal principles are equivalent.

The following variant of (HKMP) is to be noted. Let (∇) be a transitive relation over X. Call the (nonempty) subset C of X, ∇-*chain*, provided

> C is totally ordered (modulo (∇)): for each $x, y \in C$, we have either $x \nabla y$ or $y \nabla x$;

the class of all these will be denoted as $\text{chain}(X, \nabla)$. Clearly, $(\text{chain}(X, \nabla), \subseteq)$ is inductive; however, we must stress that—unlike the partially ordered case—the singleton $C = \{a\}$ (where $a \in X$) need not be an element of it, unless $a \nabla a$. Having these precise, let us consider the transitive type version of the Hausdorff–Kuratowski Maximal Principle

> (HKMP-t) The (partially) ordered structure $(\text{chain}(X, \nabla), \subseteq)$ is a Zorn one: for each ∇-chain C in X, there exists a maximal (modulo $(\subseteq)$) ∇-chain D in X, with $C \subseteq D$.

Note that (by the inductive property above), (ZMP) $\Longrightarrow$ (HKMP-t); hence (by the precise equivalence relation) (HKMP) $\Longrightarrow$ (HKMP-t). Moreover, (HKMP-t) $\Longrightarrow$ (HKMP); so that, (HKMP-t) is equivalent with both (ZMP) and (HKMP).

Sometimes, when the ambient set X is endowed with denumerable type structures, the existence of maximal elements may be determined by using a weaker form of (AC), called: *Dependent Choice Principle* (in short: DC). Call the relation $\mathscr{R}$ over X, *proper* when

> $(X(x, \mathscr{R}) =) \mathscr{R}(x)$ is nonempty, for each $x \in X$.

Note that, in this case, $\mathscr{R}$ is to be viewed as a mapping between X and $(2)^X$; the couple $(X, \mathscr{R})$ will be then referred to as a *proper relational structure*. Further, given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a; \mathscr{R})$-*iterative*, provided $[x_0 = a; x_{n+1} \in \mathscr{R}(x_n), \forall n]$.

Proposition 6. *Let the relational structure $(X, \mathscr{R})$ be proper. Then, for each $a \in X$ there is at least an $(a, \mathscr{R})$-iterative sequence in X.*

This principle—proposed, independently, by Bernays [7] and Tarski [51]—is deductible from (AC), but not conversely; cf. Wolk [66]. Moreover, by the developments in Moskhovakis [42, Chap. 8], and Schechter [47, Chap. 6], the *reduced system* (ZF-AC+DC) is comprehensive enough so as to cover the "usual" mathematics; see also Moore [41, Appendix 2].

(D) Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of relations on X. Given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a; (\mathscr{R}_n; n \geq 0))$-*iterative* provided $[x_0 = a, x_{n+1} \in \mathscr{R}_n(x_n), \forall n]$. The following *Diagonal Dependent Choice Principle* (in short: DDC) is also taken into consideration.

Proposition 7. *Let $(\mathscr{R}_n; n \geq 0)$ be a sequence of proper relations on X. Then, for each $a \in X$, there exists at least one $(a; (\mathscr{R}_n; n \geq 0))$-iterative sequence in X.*

Clearly, (DDC) includes (DC), to which it reduces when $(\mathscr{R}_n; n \geq 0)$ is constant. The reciprocal of this is also true. In fact, letting the premises of (DDC) hold, put $P = N \times X$; and let $\mathscr{S}$ be the relation over P introduced as

$$\mathscr{S}(i, x) = \{i + 1\} \times \mathscr{R}_i(x), \ \ (i, x) \in P.$$

It will suffice applying (DC) to $(P, \mathscr{S})$ and $b := (0, a) \in P$ to get the conclusion in the statement; we do not give details.

Summing up, (DDC) is provable in (ZF-AC+DC). This is valid as well for its variant, referred to as: the *Selected Dependent Choice Principle* (in short: SDC).

Proposition 8. *Let the map $F : N \to (2)^X$ and the relation $\mathscr{R}$ over X fulfill*

$$(\forall n \in N): \mathscr{R}(x) \cap F(n + 1) \neq \emptyset, \ \ \forall x \in F(n) \ \ [F \text{ is } \mathscr{R}\text{-chainable}].$$

Then, for each $a \in F(0)$ there exists a sequence $(x(n); n \geq 0)$ in X with

$$x(0) = a; \ x(n) \in F(n), \ \forall n; \ x(n) \mathscr{R} x(n + 1), \ \forall n.$$

As before, (SDC) $\Longrightarrow$ (DC) ($\Longleftrightarrow$ (DDC)); just take $F(n) = X$, $n \geq 0$. But, the reciprocal is also true, in the sense: (DDC) $\Longrightarrow$ (SDC). This follows from

Proof (Proposition 8). Let the premises of (SDC) be admitted. Define a sequence of relations $(\mathscr{R}_n; n \geq 0)$ over X as: for each $n \geq 0$,

$$\mathscr{R}_n(x) = \mathscr{R}(x) \cap F(n + 1), \ \text{if } x \in F(n); \mathscr{R}_n(x) = \{x\}, \text{if } x \in X \setminus F(n).$$

Clearly, $\mathscr{R}_n$ is proper, for all $n \geq 0$. So, by (DDC), it follows that, for the starting $a \in F(0)$, there exists an $(a; (\mathscr{R}_n; n \geq 0))$-iterative sequence $(x(n); n \geq 0)$ in X. Combining with the very definition of $(\mathscr{R}_n; n \geq 0)$ yields the desired conclusion.

In particular, when $\mathscr{R} = X \times X$, F is $\mathscr{R}$-chainable. The corresponding variant of (SDC) is just the Denumerable Axiom of Choice (in short: (AC-N)):

Proposition 9. *Let $F : N \to (2)^X$ be a function. Then, for each $a \in F(0)$ there exists a function $f : N \to X$ with $f(0) = a$ and $f(n) \in F(n)$, $\forall n \geq 0$.*

Remark 1. As a consequence of the above facts, (DC) $\Longrightarrow$ (AC-N) in (ZF-AC). A direct verification of this is obtainable by taking $P = N \times X$ and introducing the relation over it:

$$\mathscr{R}(n, x) = \{n + 1\} \times F(n + 1), n \geq 0, x \in X;$$

we do not give details. The reciprocal of the written inclusion is not true; see Moskhovakis [42, Chap. 8, Sect. 8.25] for details.

2.3 Pasicki Approach

Let X be a nonempty set; and $f : X \times X \to R$ be a *triangular* map; i.e.,

(c01) $f(x, z) \leq f(x, y) + f(y, z)$, $\forall x, y, z \in X$;

then, (X, f) will be called a *triangular structure*. By this very definition,

$$f(x, x) \leq 2f(x, x) \text{ (hence, } 0 \leq f(x, x)), \forall x \in X; \tag{14}$$

$$(0 \leq) f(x, x) \leq f(x, y) + f(y, x), \ \forall x, y \in X. \tag{15}$$

Let $(\preceq)$ be the relation over X introduced as

(c02) $(x, y \in X)$: $y \preceq x$ iff $f(y, x) \leq 0$.

From the triangular property, $(\preceq)$ appears as *transitive* $[z \preceq y, y \preceq x \Longrightarrow z \preceq x]$; but, in general, it is neither reflexive nor irreflexive. In fact, for each $x \in X$, we have (by definition and a previous property)

$$x \preceq x \text{ iff } f(x, x) \leq 0 \text{ (hence, } f(x, x) = 0).$$

This shows us that

$$x \preceq x \Longleftrightarrow f(x, x) = 0; \neg(x \preceq x) \Longleftrightarrow f(x, x) > 0;$$

hence the claim. Let also $(\succeq)$ stand for the *dual* of $(\preceq)$; i.e.,

$(x, y \in X)$: $x \succeq y$ if and only if $y \preceq x$ (i.e.: $f(y, x) \leq 0$).

This relation has the same properties as the original one, $(\preceq)$; we do not give details.

Now, call $z \in X$, $(\preceq)$-*minimal* (or, equivalently: $(\succeq)$-*maximal*), provided

(c03) $X(z,\succeq)\setminus\{z\}$ is empty; or, equivalently:
$f(x,z)>0$, for all $x\in X\setminus\{z\}$.

Let also $Q:=$ [$(\preceq)$-minimal] stand for the corresponding logical property; i.e.

$Q(z)\Longleftrightarrow[f(x,z)>0$, for all $x\in X\setminus\{z\}]$
$\neg Q(z)\Longleftrightarrow[\exists x\in X\setminus\{z\}: f(x,z)\le 0]$ (hence, $X(z,\succeq)\setminus\{z\}\neq\emptyset$).

The existence of such elements is to be studied under the regularity setting below:

(reg-1) Given $u\in X$, call the triangular structure (X,f), *u-bounded* when

$h(.)=f(.,u)$ is bounded below: $\inf\{h(x);x\in X\}>-\infty$.

If this holds for each $u\in X$, we say that (X,f) is *globally bounded*.

(reg-2) We say that $(x_n;n\ge 0)$ is *f-almost-convergent* towards $x\in X$ (written as $x_n\overset{ff}{\longrightarrow}x$) when $\liminf_n f(x,x_n)\le 0$. In this case, x will be called an *f-almost-limit* of (x_n); if such elements exist, we say that (x_n) is *f-almost-convergent*.

(reg-3) We say that $(x_n;n\ge 0)$ is *f-strong-Cauchy* when:

$\forall\varepsilon>0,\exists n(\varepsilon): n(\varepsilon)<n<m\Longrightarrow -\varepsilon<f(x_m,x_n)\le 0$.

In this case, (X,f) will be termed *strongly almost complete* (in short: *stralm complete*) when each f-strong-Cauchy sequence in X is f-almost-convergent.

The following result in Pasicki [45] is our starting point. Denote, for $x_0\in X$ and $A\subseteq X$ with $x_0\in A$,

$m(A,f;x_0):=\{u\in A; f(u,x_0)\le f(x,x_0),\forall x\in A\}$.

Further, let $P(.)$ be a logical property concerning elements of X; we shall term it *Q-weaker* (where $Q(.)$ is taken as before), provided

(c04) $Q(x)\Longrightarrow P(x)$; or, equivalently: $\neg P(x)\Longrightarrow\neg Q(x)$.

Theorem 6. *Assume that $x_0\in X$ is such that (X,f) is x_0-bounded, and*

(c05) each maximal chain A containing x_0, for which $m(A,f;x_0)\neq\emptyset$, has a unique smallest element.

Further, let (X,f) be stralm complete and the logical property $P(.)$ (involving elements of X) be Q-weaker. Then, one of the alternatives below holds:

(i) *$P(x_0)$ is true*
(ii) *there exists a $(\preceq)$-minimal $z\in X$ with $z\preceq x_0$, such that $P(z)$ holds.*

The proof runs as follows. Assume that

(c06) $\neg P(x_0)$ is true; hence (see above), $X(x_0,\succeq)\setminus\{x_0\}$ is nonempty.

Part 1. By the (transitive form of) Hausdorff–Kuratowski Maximal Principle, there exists a maximal $(\preceq)$-chain $A\subseteq\{x_0\}\cup X(x_0,\succeq)$, containing x_0. Denote $\gamma:=\inf\{f(y,x_0);y\in A\}$; clearly, $-\infty<\gamma\le 0$, in view of (X,f) being x_0-bounded and [from (c06)]

$$f(u,x_0)\le 0,\text{ for each } u\in X(x_0,\succeq)\setminus\{x_0\}.$$

Assume that $f(y, x_0) > \gamma$, for all $y \in A$. Then, there exists a sequence (y_n) in A such that $(f(y_n, x_0))$ is descending and convergent towards γ. As a consequence, (y_n) is f-strong-Cauchy; so, by our completeness assumption, there exists $y \in X$ with $y_n \xrightarrow{ff} y$. This element belongs to $m(A, f; x_0)$ (whence, $m(A, f; x_0)$ is nonempty); so, by (c05), y is the only smallest element of $m(A, f; x_0)$ (and of A).

Part 2. Suppose there exists $z \in X \setminus \{y\}$ such that $z \preceq y$. It results that $z \neq y$ is another smallest element of A; in contradiction to (c05).

Part 3. By replacing y with z we obtain $P(z), f(z, x_0) = \gamma \leq 0$ and $0 < f(x, z)$, $\forall x \in X \setminus \{z\}$. Clearly, z is a minimal element in X, as A is a maximal chain.

Technically speaking, Theorem 6 is equivalent with its particular version based upon P identified with Q; referred to as: Theorem 6 (Q-version). In fact, suppose that Theorem 6 (Q-version) holds; and let the logical property $P(.)$ (involving elements of X) be Q-weaker; i.e.: $Q(x) \Longrightarrow P(x)$, $\forall x \in X$. Two alternatives are possible:

(i) $Q(x_0)$ is true; **(ii)** $Q(x_0)$ is false.

In the former case, $P(x_0)$ is true; and we are done. In the latter case, by the argument above, there exists $z \in X$ with $z \preceq x_0$, fulfilling

z is $(\preceq)$-minimal (so that, by definition, $Q(z)$ is true);

and this, along with $Q(z) \Longrightarrow P(z)$, tells us that $P(z)$ is also true; hence the claim.

Note that, as long as (KMP-t) (equivalent, as above said, with (AC)) is involved here, the obtained result is deductible in the complete axiomatic system (ZF). But, all arguments appearing in this proof are sequential in nature; so, we may expect that Theorem 6 is deductible in the reduced system (ZF-AC+DC). It is our aim in the following to show that, ultimately, this is possible by the recursion to a certain "transitive" form of Brezis–Browder ordering principle [10] established in Turinici [56]. Further aspects occasioned by these developments are also discussed.

2.4 Transitive Brezis–Browder Principles

Let X be some nonempty set. Take a *quasi-order* (i.e.: reflexive and transitive relation) $(\preceq)$ over it; as well as a function $g : X \to R$. Call the point $z \in X$, $(\preceq, g)$-*maximal* when:

(d01) $w \in X$ and $z \preceq w$ imply $g(z) = g(w)$;

i.e.: g is constant on $X(z, \preceq)$. A basic result about existence of such points is the 1976 Brezis–Browder ordering principle [10] (in short: BB).

Theorem 7. *Assume that*

(d02) $(X, \preceq)$ *is sequentially inductive:*
each ascending sequence has an upper bound (modulo $(\preceq)$*)*
(d03) g *is* $(\preceq)$*-decreasing* $(x \preceq y \Longrightarrow g(x) \geq g(y))$
(d04) g *is bounded from below* $(\inf(g(X)) > -\infty)$.

Then, each $u \in X$ is (BB)-admissible (modulo $(\preceq, g)$), in the sense: there exists a $(\preceq, g)$-maximal $v \in X$ with $u \preceq v$.

Proof. Define the function $b : X \to R$ as: $b(v) := \inf[g(X(v, \preceq))]$, $v \in X$. The convention is meaningful, via (d04); in addition, $b(.)$ is increasing and

$$g(v) \geq b(v), \text{ for all } v \in X. \tag{16}$$

Moreover, as g=decreasing, one gets a characterization like

$$v \text{ is } (\preceq, g)\text{-maximal iff } g(v) = b(v). \tag{17}$$

Now, assume by contradiction that the conclusion in this statement is false; i.e. [in combination with (17)] there must be some $u \in X$ such that:

(d05) for each $v \in X_u := X(u, \preceq)$, one has $g(v) > b(v)$.

Consequently (for all such v), $g(v) > (1/2)(g(v) + b(v)) > b(v)$; hence

$$v \preceq w \text{ and } (1/2)(g(v) + b(v)) > g(w), \tag{18}$$

for at least one w (belonging to X_u). The relation $\mathscr{R}$ over X_u introduced via (18) fulfills $X_u(v, \mathscr{R}) \neq \emptyset$, for all $v \in X_u$. So, by the Dependent Choice Principle, there must be a sequence (u_n) in X_u with $u_0 = u$ and

$$u_n \preceq u_{n+1}, (1/2)(g(u_n) + b(u_n)) > g(u_{n+1}), \text{ for all } n. \tag{19}$$

We have thus constructed an ascending sequence (u_n) in X_u for which the real sequence $(g(u_n))$ is (from (d04) and (16)) strictly descending and bounded below; hence $\lambda := \lim_n g(u_n)$ exists in R. By (d02), (u_n) is bounded from above in X: there exists $v \in X$ such that $u_n \preceq v$, for all n. From the properties of (g, b), we have

$$g(u_n) \geq g(v), \text{ and } g(v) \geq b(v) \geq b(u_n), \forall n.$$

The former of these relations gives $\lambda \geq g(v)$ (passing to limit as $n \to \infty$). On the other hand, the latter of these relations yields (via (19))

$$(1/2)(g(u_n) + b(v)) > g(u_{n+1}), \text{ for all } n \geq 0.$$

Passing to limit as $n \to \infty$ gives $(g(v) \geq) b(v) \geq \lambda$; so, combining with the preceding one, $g(v) = b(v) (= \lambda)$, contradiction. Hence, (d05) cannot be accepted; and the conclusion follows.

Note that, by the argument above, (DC) $\Longrightarrow$ (BB) in (ZF-AC). For a slightly different proof, we refer to Cârjă, Necula, and Vrabie [14, Chap. 2, Sect. 2.1]. Further metrical extensions of (BB) may be found in Turinici [55].

(A) In the following, a transitive type variant of this principle is to be stated. Let (∇) be a (nonempty) relation over X; assumed to be *transitive* $[x\nabla y, y\nabla z \Longrightarrow x\nabla z]$. The associated relation $(\preceq)$ on X introduced as

(d06) $x \preceq y$ iff either $x = y$ or $x\nabla y$

is reflexive and transitive; hence, a quasi-order on X; moreover, the following assertion is true

$$x\nabla y \text{ and } y \preceq z \text{ imply } x\nabla z. \tag{20}$$

Further, take a function $g : X \to R$; and consider the following condition:

(d07) g is ∇-decreasing: $x\nabla y \Longrightarrow g(x) \geq g(y)$.

Note that, from the very definition of the induced quasi-order (see above), one has the generic relation

$$g \text{ is } \nabla\text{-decreasing} \Longleftrightarrow g \text{ is } (\preceq)\text{-decreasing}. \tag{21}$$

Finally, call the point $z \in X$, (∇, g)-*maximal*, provided

(d08) $w \in X$ and $z\nabla w$ imply $g(z) = g(w)$.

As before, the generic relation holds

$$z \text{ is } (\nabla, g)\text{-maximal} \Longleftrightarrow z \text{ is } (\preceq, g)\text{-maximal}. \tag{22}$$

As a consequence of this, maximality results involving the transitive relation (∇) are deductible from Brezis–Browder's principle involving its associated quasi-order $(\preceq)$. A key moment of this approach is that of the sequential inductivity condition for $(X, \preceq)$ being assured. It would be useful to have expressed this requirement in terms of the initial transitive relation. This necessitates a few conventions and auxiliary facts. Call the sequence (x_n), *ascending* (modulo (∇)) when

$x_n\nabla x_{n+1}, \forall n$ (or, equivalently: $x_n\nabla x_m$, if $n < m$).

Clearly, the generic (sequential) relation holds

$$\text{ascending (modulo } (\nabla)) \Longrightarrow \text{ascending (modulo } (\preceq)).$$

The reciprocal is not in general true. For example, given $a \in X$, the constant sequence $(x_n = a; n \geq 0)$ is ascending (modulo $(\preceq)$); but not ascending (modulo (∇)), as long as $a\nabla a$ is false. Further, given the sequence (x_n) in X, let us say that $u \in X$ is an *upper bound* (modulo (∇)) of it, provided

(d09) $x_n\nabla u, \forall n$ (written as: $(x_n)\nabla u$).

If such elements u exist, we say that (x_n) is *bounded above* (modulo (∇)). As before, the relation below is clear

$$(\forall u):\ [(x_n)\nabla u] \Longrightarrow [(x_n) \preceq u];$$

the converse is not in general valid. Finally, let us consider the condition

(d10) (X, ∇) is sequentially inductive:
each ascending sequence has an upper bound (modulo (∇)).

The following auxiliary statement is useful for us.

Lemma 1. *The generic implication is true*

$$(X, \nabla) \textit{ is sequentially inductive} \Longrightarrow (X, \preceq) \textit{ is sequentially inductive.} \tag{23}$$

Proof. Let (x_n) be an ascending (modulo $(\preceq)$) sequence in M

$$x_n \preceq x_{n+1}, \forall n \ \text{ (hence } x_n \preceq x_m \text{ whenever } n \leq m).$$

If this sequence is stationary beyond a certain rank

$$\exists k \text{ such that: } \forall n > k \text{ one has } x_n = x_k$$

we are done; because $(x_n) \preceq u$, where $u = x_k$. Otherwise,

$$\forall p, \exists q > p \text{ such that } x_p \neq x_q \ \text{ (hence } x_p \nabla x_q).$$

Consequently, a subsequence $(y_n = x_{i(n)})$ of (x_n) may be constructed with the property of being ascending (modulo (∇)); wherefrom (by the admitted hypothesis) $(y_n)\nabla t$ (hence, $(y_n) \preceq t$), for some $t \in X$. But then, $(x_n) \preceq t$; hence the claim.

Remark 2. Formally, the subsequence construction above requires the Denumerable Axiom of Choice (AC-N). However, since its ambient set is N, this may be avoided. In fact, define the couple of functions

(d11) $\Phi(p) = \{q > p; x_p \nabla x_q\}, \varphi(p) = \min \Phi(p), p \in N.$

Note that, in these conventions, no choice arguments are needed; moreover,

$$p < \varphi(p),\ \forall p \in N;\ \text{(i.e.: } \varphi \text{ is strictly progressive).}$$

Then, the strictly ascending rank sequence to be used here is

$$i(0) = 0, i(n + 1) = \varphi(i(n)), n \geq 0;$$

and this proves our assertion.

We may now give an appropriate answer to the posed question. Given $u \in X$, call it *(BB)-admissible* (modulo (∇, g)), if there exists a (∇, g)-maximal $v \in X$ with $u \preceq v$. Note that, when

$X(u, \nabla) = \emptyset$ (i.e.: u is (trivially) (∇, g)-maximal),

this property is fulfilled; so, the verification is required for those $u \in X$ with $X(u, \nabla) \neq \emptyset$; referred to as: u is ∇-*starting*.

The following "transitive" form of (BB) (in short: (BB-t)) is now available.

Theorem 8. *Let the transitive relation* (∇) *and the function* $g : X \to R$ *be such that* (X, ∇) *is sequentially inductive and* g *is* ∇*-decreasing, bounded from below. Then, each* ∇*-starting* $u \in X$ *is (BB)-admissible (modulo* (∇, g)*), in the sense: there exists a* ∇*-maximal* $v \in X$ *with* $u \nabla v$.

Proof. By the preceding auxiliary fact, $(X, \preceq)$ is sequentially inductive; and (cf. a previous remark), g is $(\preceq)$-decreasing; hence, (BB) applies to $(X, \preceq; g)$. From its conclusion we have that, given the ∇-starting point $u \in X$, we have: for the arbitrary fixed element $u_1 \in X(u, \nabla)$, there exists another one, $v \in X$ such that

(i) $u_1 \preceq v$; **(ii)** v is $(\preceq, g)$-maximal.

The latter of these yields v is (∇, g)-maximal (see above). And the former one gives $u \nabla v$, if one takes a simple observation involving the couple $(\nabla, \preceq)$ into account. The proof is complete.

By this very argument, (BB) $\Longrightarrow$ (BB-t). The reciprocal inclusion ((BB-t) $\Longrightarrow$ (BB)) is also true; just note that, given the quasi-order $(\leq)$ on X, we have

$(\leq)$ is transitive, and any $u \in X$ is $(\leq)$-starting.

Hence, summing up, (BB) $\Longleftrightarrow$ (BB-t). Note that, a further extension of (BB-t) in terms of general (amorphous) relations is possible; see Turinici [56] for details.

2.5 Main Results

Let X be a nonempty set; and $f : X \times X \to R$ be a *triangular* map; i.e.,

$f(x, z) \leq f(x, y) + f(y, z)$, $\forall x, y, z \in X$;

the couple (X, f) will be then referred to as a *triangular structure*. Note that, by the very definition above,

$$f(x, x) \leq 2f(x, x) \text{ (hence, } 0 \leq f(x, x)), \forall x \in X; \tag{24}$$

this will be useful in the sequel.

(A) Let (∇) denote the transitive relation attached to $f(.,.)$

$(x, y \in X)$: $x \nabla y$ iff $f(y, x) \leq 0$;

and $(\preceq)$ stand for the associate quasi-order

$x \preceq y$ iff either $x = y$ or $x \nabla y$.

In the following, we are interested to apply (BB-t) to these data. This will necessitate a lot of new concepts and auxiliary facts.

Lemma 2. *Let $a, b \in X$ be arbitrary fixed. Then,*

$$g(.) := f(a,.) \text{ is } \nabla\text{-increasing } (x \nabla y \Longrightarrow g(x) \le g(y))$$

$$h(.) := f(.,b) \text{ is } \nabla\text{-decreasing } (x \nabla y \Longrightarrow h(x) \ge h(y)).$$

Proof. Let $x, y \in X$ be such that $x \nabla y$; i.e.: $f(y,x) \le 0$. By the triangular property,

$$f(a,x) \le f(a,y) + f(y,x) \le f(a,y),$$
$$f(y,b) \le f(y,x) + f(x,b) \le f(x,b);$$

and conclusion follows.

(B) Remember that, given $u \in X$, the triangular structure (X,f) is called *u-bounded*, when

$h(.) := f(.,u)$ is bounded below: $\inf\{h(x); x \in X\} > -\infty$.

If this holds for at least one $u \in X$, we say that (X,f) is *locally bounded*; and, if this holds for each $u \in X$, we say that (X,f) is *globally bounded*. Clearly, each globally bounded triangular structure is locally bounded too. But, the reciprocal is also true; as results from

Lemma 3. *Each locally bounded triangular structure is globally bounded too. Hence, for each triangular structure, (X,f),*

$$\text{locally bounded} \Longleftrightarrow \text{globally bounded.} \tag{25}$$

Proof. Assume that (X,f) is u-bounded; and let $v \in X$ be arbitrary fixed. By the triangular inequality,

$$f(x,u) - f(v,u) \le f(x,v) \le f(x,u) + f(u,v), \ \forall x \in X;$$

and, from this, we derive that (X,f) is v-bounded.

(C) Call the sequence $(x_n; n \ge 0)$, ∇-*ascending*, when

$n < m$ implies $x_n \nabla x_m$ (i.e.: $f(x_m, x_n) \le 0$);

the class of all such objects will be denoted $\mathrm{asc}(X, \nabla)$. Let us introduce a convergence structure on this class as follows: given the sequence $(x_n; n \ge 0)$ in $\mathrm{asc}(X, \nabla)$ and the point $x \in X$, put

(e01) $x_n \xrightarrow{f} x$ iff $\lim_n f(x, x_n) \le 0$;

referred to as: $(x_n; n \geq 0)$, *f-converges* towards x. Note that, by a previous auxiliary fact, the real sequence $(f(x, x_n); n \geq 0)$ is ascending; hence, $\lim_n f(x, x_n)$ exists (in the generalized sense). This tells us that the convention above is meaningful; and reads: $x \in X$ is a *f-limit* of the sequence $(x_n; n \geq 0)$ in $\mathrm{asc}(X, \nabla)$. The class of all such elements will be denoted as $f - \lim_n(x_n)$; when it is nonempty, we say that $(x_n; n \geq 0)$ is *f-convergent*.

The following property of a f-convergent sequence will be useful for us.

Lemma 4. *Let the ∇-ascending sequence $(x_n; n \geq 0)$ in X and the point $x \in X$ be such that $x_n \xrightarrow{f} x$. Then, $(x_n)\nabla x$; i.e.: $x_n \nabla x$, for all n.*

Proof. By a previous remark, the real sequence $(f(x, x_n); n \geq 0)$ is ascending. This yields, for each n,

$$f(x, x_n) \leq \lim_n f(x, x_n) \leq 0; \text{ (i.e.: } x_n \nabla x);$$

and, from this, we are done.

(D) Further, let us say that the sequence $(x_n; n \geq 0)$ in $\mathrm{asc}(X, \nabla)$ is *f-strong-Cauchy* provided (see above)

(e02) $\forall \varepsilon > 0, \exists n(\varepsilon)$, such that: $[n(\varepsilon) < n < m \Longrightarrow -\varepsilon < f(x_m, x_n) \leq 0]$.

Note that, by the lack of symmetry for the triangular map $f(.,.)$, an f-convergent sequence in $\mathrm{asc}(X, \nabla)$ need not be f-strong-Cauchy. However, we say that the triangular structure (X, f) is *∇-complete*, when each ∇-ascending f-strong-Cauchy sequence in X is f-convergent.

The following auxiliary fact will be used in the sequel.

Lemma 5. *Assume that (X, f) is globally bounded and ∇-complete. Then,*

(X, ∇) is sequentially inductive: for each ∇-ascending sequence $(x_n; n \geq 0)$ in X, there exists $x \in X$ with $x_n \nabla x$, $\forall n$.

Proof. Fix some $u \in X$; and put $h(.) = f(., u)$. Further, let the sequence $(x_n; n \geq 0)$ in X be ∇-ascending. By the triangular property,

$$h(x_m) - h(x_n) \leq f(x_m, x_n) \leq 0, \text{ if } n < m.$$

The (real) sequence $(h(x_n); n \geq 0)$ is descending and bounded from below; hence, a convergent one; and this (combined with the above) tells us that $(x_n; n \geq 0)$ is a f-strong-Cauchy ∇-ascending sequence in X. As (X, f) is ∇-complete, there must be some $x \in X$ with $x_n \xrightarrow{f} x$. Taking a previous auxiliary fact into account yields $x_n \nabla x$, $\forall n$; and our conclusion follows.

(B) Given $u \in X$, let us again denote $h(.) := f(., u)$. Remember that u is called *(BB)-admissible* (modulo (∇, h)), when there exists a (∇, h)-maximal element $v \in X$ with $u \preceq v$. Note that, if $X(u, \nabla) = \emptyset$, this condition is fulfilled, with $v = u$; so, the verification of the underlying property is required for those $u \in X$ with $X(u, \nabla) \neq \emptyset$, referred to as: u is *∇-starting*.

Concerning the sufficient conditions for this, it is worth stressing that, by the triangular context, these may be expressed in an "absolute" way (not depending on starting point $u \in X$). Define a new relation $(<)$ over X as

(e03) $x < y$ iff $f(y, x) < 0$.

Clearly, $(<)$ is, by (24), *irreflexive* [$x < x$ is false, $\forall x \in X$] and (via f=triangular) *transitive* [$x < y, y < z \Longrightarrow x < z$]; hence, it is a *strict order* on X. Let also $(\leq)$ stand for the associated relation

(e04) $x \leq y$ iff either $x = y$ or $x < y$;

it is an order (i.e.: antisymmetric quasi-order) as it can be directly seen.

Concerning the connections with our initial relation (∇), we have (by definition)

$$x < y \Longrightarrow x \nabla y; \text{ hence, } x \leq y \Longrightarrow x \preceq y. \tag{26}$$

The converse relation is not in general true. For example, if $x, y \in X$ with $x \neq y$ are such that $f(y, x) = 0$, then $x \nabla y$; but, evidently, $x < y$ is false.

Now, call $v \in X$, $(<)$-*maximal* if

(e05) $X(v, <) = \emptyset$; or, equivalently: $X(v, \leq) = \{v\}$.

By the second relation above, v is also referred to as $(\leq)$-*maximal*; because this is the usual concept of maximality, as in Bourbaki [9]. Denote by $Q := (<, \max)$ the corresponding logical property; i.e.

$Q(v) \Longleftrightarrow (f(x, v) \geq 0$, for all $x \in X)$
$\neg Q(v) \Longleftrightarrow (f(x, v) < 0$ (hence, $v < x$), for some $x \in X)$.

Lemma 6. *Let $u \in X$ be fixed and $v \in X$ be (∇, h)-maximal, where $h(.) := f(., u)$. Then,*

$$v \text{ is } (<)\text{-maximal; hence, } f(x, v) \geq 0, \ \forall x \in X \tag{27}$$

$$x \in X, \ v \nabla x \Longrightarrow f(x, v) = 0. \tag{28}$$

Proof. **(i)** Assume by contradiction that $v < x$ (i.e.: $f(x, v) < 0$) for some $x \in X$. From a previous implication, we then have $v \nabla x$; so that (by maximality), $h(v) = h(x)$. Combining the triangular property, gives

$$0 = h(x) - h(v) \leq f(x, v) < 0;$$

contradiction. This proves our claim.

(ii) Let $x \in X$ be such that $v \nabla x$. By definition (and the preceding relation)

$$0 = h(x) - h(v) \leq f(x, v) \leq 0;$$

whence $f(x, v) = 0$.

The first main result of the present exposition is

Theorem 9. *Assume that (X,f) is globally bounded and ∇-complete. Then, for each ∇-starting point $u \in X$ there exists another point $v \in X$ with the properties (27)+(28), such that*

$$u\nabla v \text{ (hence, } f(v,u) \leq 0). \tag{29}$$

Proof. Denote for simplicity $h(.) := f(.,u)$. We claim that (BB-t) applies to $(X,\nabla;h)$ and $u \in X$. Firstly, by a preceding auxiliary fact, $h(.)$ is ∇-decreasing on X. Secondly (by the globally bounded property), $h(.)$ is bounded from below on X. Finally, again by the regularity conditions upon (X,f), it results (see above) that (X,∇) is sequentially inductive; hence, the claim. From (BB-t) it follows that, for the ∇-starting $u \in X$ there exists a (∇,h)-maximal $v \in X$ with $u\nabla v$. This, along with a previous fact, gives all conclusions we need.

As a direct consequence of this, we derive the second main result of this exposition. Namely, let $P(.)$ be a logical property relative to elements of X; we call it *Q-weaker*, if

(e06) $(\forall z \in X)$: $Q(z) \Longrightarrow P(z)$.

Theorem 10. *Assume that (X,f) is globally bounded and ∇-complete. Further, let $P(.)$ be a logical property relative to elements of X; supposed to be Q-weaker. Then, for each ∇-starting $u \in X$ there exists some $v \in X$ in such a way that (27)–(29) hold, as well as*

$$Q(v) \text{ is true; hence, } P(v) \text{ is true.} \tag{30}$$

For the moment, Theorem 9 $\Longrightarrow$ Theorem 10; moreover, the latter of these may be viewed as a simplified form of the Pasicki statement. The reciprocal implication is also true; just take the logical property $P(.)$ as identical with $Q(.)$. Hence, these two results are equivalent to each other. This also tells us that the introduction of logical properties like before in Theorem 9 does not produce any generalizing effect upon it. But, from a practical perspective, this may be useful; we do not give details.

2.6 Converse Question

The developments above tell us that the main result (subsumed to Theorem 9) is reducible to (BB-t); hence, ultimately, to (BB). So, we may ask whether the converse question is available too. It is our aim in the following to provide a positive answer to this. Some preliminary facts are in order. Remember that (see above) (DC) $\Longrightarrow$ (BB) and (BB) $\Longleftrightarrow$ (BB-t). On the other hand, (BB) $\Longrightarrow$ (EVP) (cf. Turinici [62]); and, finally (by the developments in Brunner [12]) (EVP) $\Longrightarrow$ (DC); hence, all these principles are equivalent to each other. As a consequence,

(I) any maximal/variational principle (MP) with (DC) $\Longrightarrow$ (MP) $\Longrightarrow$ (EVP) is equivalent with both (DC) and (BB)

(II) any maximal/variational principle (MP) with (BB) $\Longrightarrow$ (MP) $\Longrightarrow$ (EVP) is equivalent with both (BB) and (EVP).

For example, the first conclusion is applicable to many extensions of (BB), like the ones in Altman [1] and Szaz [49]; see also Du [21]. On the other hand, the second conclusion is applicable to the vector variational principle in Goepfert, Tammer, and Zălinescu [25]; see, for instance, Turinici [55]. It also works for the results in Sect. 2.5; but, with a slightly modified version of (EVP), described as follows.

Let (X,d) be a complete metric space; and $\varphi : X \to R$ be a function. The following "generic" version of (EVP) (in short: (EVP-g)) is to be considered.

Theorem 11. *Assume that φ is d-lsc and bounded from below on X. In addition, let (X,d) be complete. Then, there exists at least one $v \in X$, with*

$$d(v,x) > \varphi(v) - \varphi(x), \quad \text{for all } x \in X \setminus \{v\}. \tag{31}$$

Clearly, (EVP) $\Longrightarrow$ (EVP-g) in a trivial way. But, the remarkable fact to be added is that the converse implication is also true:

$$\text{(EVP-g)} \Longrightarrow \text{(EVP); hence, (EVP-g)} \Longleftrightarrow \text{(EVP)}. \tag{32}$$

In fact, assume that the premises of (EVP) hold. Let $(\leq)$ stand for the Brøndsted order [11] attached to φ

$$(x,y \in X)\text{: } x \leq y \text{ iff } d(x,y) \leq \varphi(x) - \varphi(y);$$

and fix some $u \in X$; note that (by the d-lsc property of φ), $X_u := X(u, \leq)$ is d-closed (hence, (X_u, d) is complete). It will suffice applying (EVP-g) to (X_u, d) and φ (restricted to X_u) to deduce all desired conclusions in (EVP); see Bao and Khanh [4] for details.

(B) In the following, we shall give the announced answer concerning the logical equivalence question. Remember that (BB) $\Longrightarrow$ (BB-t) $\Longrightarrow$ Theorem 9. So, to close the circle, it will suffice proving that Theorem 9 $\Longrightarrow$ (EVP); or, equivalently (see above) Theorem 9 $\Longrightarrow$ (EVP-g).

Proposition 10. *Under these notations, Theorem 9 $\Longrightarrow$ (EVP-g). So (combining with the above) Theorem 9 is equivalent with both (BB) and (EVP).*

Proof. Let the premises of (EVP-g) be admitted. Fix $\lambda > 1$; and define a map $f : X \times X \to R$ as

$$f(x,y) = \lambda(\varphi(x) - \varphi(y)) + d(x,y), x, y \in X;$$

clearly, f is triangular. Let (∇) denote the relation over X

$$\text{(f01)} \quad x \nabla y \text{ iff } f(y,x) \leq 0; \text{ i.e.: } d(x,y) \leq \lambda(\varphi(x) - \varphi(y));$$

note that, as a direct consequence of this, (∇) is reflexive, transitive, antisymmetric—hence, a (partial) order—on X. Further, let $(<)$ stand for the strict order on X:

(f02) $x < y$ iff $f(y,x) < 0$; i.e.: $d(x,y) < \lambda(\varphi(x) - \varphi(y))$.

We claim that Theorem 9 applies to (X,f) and $(\nabla, <)$. This will be done in three steps, as follows.

Step 1. Fix in the following $u \in X$; and put

$$h(x) = f(x,u)(= \lambda(\varphi(x) - \varphi(u)) + d(x,u)), x \in X.$$

As φ is bounded from below, it immediately follows that so is $h(.)$; this, along with a previous auxiliary fact, tells us that (X,f) is globally bounded.

Step 2. Let $(x_n; n \geq 0)$ be an ascending (modulo (∇)) sequence in X; i.e.,

(f03) $d(x_n, x_m) \leq \lambda(\varphi(x_n) - \varphi(x_m))$, if $n < m$.

The real sequence $(\varphi(x_n); n \geq 0)$ is descending and bounded from below; hence, a Cauchy one. Combining with the working hypothesis, one gets that $(x_n; n \geq 0)$ is a Cauchy sequence in X; so, by completeness, $x_n \overset{d}{\longrightarrow} x$ for some $x \in X$. As φ is d-lsc, this gives $\varphi(x_n) \geq \varphi(x)$, for all n. Replacing in the working hypothesis gives

$$d(x_n, x_m) \leq \lambda(\varphi(x_n) - \varphi(x)), \text{ if } n < m;$$

so that, passing to limit as $m \to \infty$,

$$d(x_n, x) \leq \lambda(\varphi(x_n) - \varphi(x)), \text{ for all } n;$$

which may be also written as

$$f(x, x_n) \leq 0 \text{ (i.e.: } x_n \nabla x), \text{ for all } n.$$

On the other hand, by a previous remark, the real sequence $(f(x,x_n); n \geq 0)$ is ascending; hence, $\lim_n f(x,x_n)$ exists (in the general sense). Combining with the above relation gives $\lim_n f(x,x_n) \leq 0$; whence $x_n \overset{f}{\longrightarrow} x$. This shows that each ∇-ascending sequence in X is f-convergent; hence, in particular, (X,f) is complete.

Step 3. As (∇) is (partial) order, it results that each $u \in X$ is ∇-starting. By Theorem 9 it follows that, for the (fixed) element $u \in X$ there exists an element $v \in X(u, \nabla)$ such that

$$v \text{ is } (<)\text{-maximal: } d(v,x) \geq \lambda(\varphi(v) - \varphi(x)), \forall x \in X. \tag{33}$$

We claim that v is our desired element for (31). Assume not; i.e.,

(f04) $(0 <) d(v,y) \leq \varphi(v) - \varphi(y)$, for some $y \in X \setminus \{v\}$.

This yields $\varphi(v) - \varphi(y) > 0$; so that, combining with (33), one gets (via $\lambda > 1$)

$$d(v, y) \geq \lambda(\varphi(v) - \varphi(y)) > \varphi(v) - \varphi(y);$$

in contradiction to the working assumption. This ends the argument.

Summing up, the results in Sect. 2.5 are nothing else than equivalent versions of (BB) and/or (EVP). Further aspects may be found in Zhu et al. [70]; see also Turinici [65].

3 GTZ Maximal Principles in Topological Vector Spaces

3.1 Introduction

Let $(Y, \mathscr{T})$ be a (real) separated *locally convex space*; and K be some *(convex) cone* of it:

$\alpha K + \beta K \subseteq K$, for all $\alpha, \beta \in R_+ := [0, \infty[$.

The relation $(\leq_K)$ on Y

$(y_1, y_2 \in Y)$: $y_1 \leq_K y_2$ if and only if $y_2 - y_1 \in K$

is reflexive and transitive; hence, a *quasi-order*; when K is understood, we simply denote it as $(\leq)$. Further, take a metric space (X, d); and let $F : X \to 2^Y$ be a multivalued map from X to Y (identified with its *graph* in $X \times Y$), fulfilling

(a01) F is *proper* ($\mathrm{Dom}(F) := \{x \in X; F(x) \neq \emptyset\}$ is nonempty).

Finally, pick some $k^0 \in K$; and let $(\preceq)$ be the quasi-order on $X \times Y$, introduced as

$(x_1, y_1) \preceq (x_2, y_2)$ if and only if $k^0 d(x_1, x_2) \leq y_1 - y_2$.

For both practical and theoretical reasons, it would be useful to determine sufficient conditions under which $(F, \preceq)$ has points with certain maximal properties. The basic result in this area obtained by Goepfert, Tammer, and Zălinescu [25] (in short: GTZ), deals with convex cones K and points $k^0 \in K$ taken as

(a02) k^0 is *K-admissible*: $k^0 \in K \setminus (-\mathrm{cl}(K))$.

[Here, "cl" is the *closure operator* on Y]. Precisely, assume that

(a03) F is *bounded below*: $F(X) \subseteq \tilde{y} + K$, for some $\tilde{y} \in Y$
(a04) F is $(\preceq)$*-semi-closed*:
if $((x_n, y_n)) \subseteq F$ is $(\preceq)$-ascending and $x_n \stackrel{d}{\longrightarrow} x$, then $x \in \mathrm{Dom}(F)$
and there exists $y \in F(x)$ such that $(x_n, y_n) \preceq (x, y)$, for all n.

Theorem 12. *Let the above conditions be in force. In addition, let (X, d) be complete. Then, for each $(x_0, y_0) \in F$ there exists $(\bar{x}, \bar{y}) \in F$ with*

$$(x_0, y_0) \preceq (\bar{x}, \bar{y}) \ [hence\ y_0 \geq \bar{y}] \tag{34}$$

$$(\bar{x}, \bar{y}) \preceq (x', y') \in F \ implies \ \bar{x} = x'. \tag{35}$$

This result includes Ekeland's variational principle [23] (in short: EVP). Concerning the reciprocal inclusion, note that (GTZ) is deductible from the Brezis–Browder ordering principle [10] (in short: BB); see, for instance, Turinici [55]. On the other hand (as established in Turinici [64]), (BB) is deductible from the Dependent Choices Principle (in short: DC) due to Bernays [7] and Tarski [51]. Finally, by the result in Brunner [12], (EVP) includes (DC). Summing up, we have the inclusion chain between these statements

$$(\mathrm{DC}) \Longrightarrow (\mathrm{BB}) \Longrightarrow (\mathrm{EVP}) \Longrightarrow (\mathrm{DC}).$$

This in particular says that, any variational/maximal principle (VP) with (BB) $\Longrightarrow$ (VP) $\Longrightarrow$ (EVP) is equivalent with both (BB) and (EVP). For example, (GTZ) belongs to this "logical" interval between (BB) and (EVP); hence, it supports such a conclusion. It follows that genuine extensions of (BB) and/or (EVP) must be not deductible from (DC). A basic example in the area is the one due to Zhu and Li [69]; we do not give details.

Note that, due to semi-complete form of (35), (GTZ) is not an authentic Zorn maximal principle. To get such a conclusion, an extra condition must be added. Given the subset V of Y, call $v \in V$, *minimal* (modulo K) when

$$w \in V,\ v \geq w \text{ imply } v \leq w;$$

the class of all these will be denoted $\min(V; K)$; or, simply, $\min(V)$ (if K is understood). Now, the announced condition writes:

(a05) F has the *domination property*:
$\forall x \in \mathrm{Dom}(F)$, $\forall z \in F(x)$, $\exists \bar{z} \in \min(F(z))$, such that $z \geq \bar{z}$.

Theorem 13. *Suppose that (in addition), F has the domination property. Then, for each $(x_0, y_0) \in F$ there exists $(\bar{x}, \bar{y}) \in F$ such that*

$$(x_0, y_0) \preceq (\bar{x}, \bar{y}), \quad \bar{y} \in \min(F(\bar{x})) \tag{36}$$

$$(\bar{x}, \bar{y}) \preceq (x', y') \in F \Longrightarrow \bar{x} = x', \bar{y} \leq y'. \tag{37}$$

The proof is immediate, by the remarks above; see also Sect. 3.4 for details. Some other aspects were discussed in Hamel and Tammer [29].

The basic assumption of both these results is (a04). For example, it holds when

(a06) K has closed lower sections: $K \cap (v - k^0 R_+)$ is closed, $\forall v \in K$

(a07) F is sub-monotone: for each sequence $((x_n, y_n))$ in F with $x_n \overset{d}{\longrightarrow} x$ and (y_n)=descending, there exists $y \in F(x)$ such that $y_n \geq y$, $\forall n$;

see the quoted papers for details. This continues to hold for transfinite versions of Theorem 12; cf. Turinici [60]. It is our aim in the following to establish (in Sects. 3.4 and 3.5) that the key condition (a04) is still appropriate for a recent variational principle due to Bao and Mordukhovich [5] (discussed in Sect. 3.3). The basic tool of our developments is a lot Brezis–Browder maximality principles, analyzed in Sect. 3.2. Note that the proposed techniques are applicable as well to some other vector type variational results in the area; we do not give details.

3.2 Brezis–Browder Principles

Let M be a nonempty set. Take a *quasi-order* $(\leq)$ (i.e.: reflexive and transitive relation) over it, as well as a function $\psi : M \to R_+$. Call the point $z \in M$, $(\leq, \psi)$-*maximal* when it satisfies

$$w \in M \text{ and } z \leq w \text{ imply } \psi(z) = \psi(w).$$

A basic result about existence of such points is the 1976 Brezis–Browder ordering principle [10] (in short: BB).

Proposition 11. *Suppose that*

(b01) $(M, \leq)$ *is sequentially inductive:*
each ascending sequence has an upper bound (modulo $(\leq)$*)*
(b02) ψ *is* $(\leq)$*-decreasing* $(x \leq y \Longrightarrow \psi(x) \geq \psi(y))$.

Then, for each $u \in M$ *there exists a* $(\leq, \psi)$*-maximal* $v \in M$*, with* $u \leq v$.

This statement, including the well-known Ekeland's variational principle [23] (in short: EVP), found some useful applications to convex and nonconvex analysis (cf. the above references). So, it cannot be surprising that many extensions of Proposition 11 were proposed; see, for instance, Altman [1], Anisiu [2], or Bae et al. [3]. The obtained results are interesting from a technical viewpoint. However, we must emphasize that, in all concrete situations when a maximality principle of this type is to be applied, a substitution of it by the Brezis–Browder's is always possible. This (cf. Bao and Khanh [4]) raises the question of to what extent are these enlargements of (BB) effective. As already precise, the answer is negative for most of these; hence, in particular, for the 1990 ordering principle in Kang and Park [35]. But, from a practical viewpoint, these are interesting tools in the area; so, a discussion of them is always welcomed. It is our aim in the following to list some basic maximal statements of this type, for a practical use.

Let M be some nonempty set; and $(\leq)$, some quasi-order on it. Further, let $x \mapsto \varphi(x)$ stand for a function between M and $R_+ \cup \{\infty\} = [0, \infty]$.

Proposition 12. *Assume (b01) and (b02) are true, as well as*

(b03) $((M, \leq)$ *is almost regular (modulo* $\varphi))$
$\forall x \in M, \forall \varepsilon > 0, \exists y = y(x, \varepsilon) \geq x : \varphi(y) \leq \varepsilon.$

Then, for each $u \in M$ there exists $v \in M$ with $u \leq v$ and $\varphi(v) = 0$ (whence, necessarily, v is $(\leq, \varphi)$-maximal).

Proof. From (b03), there must be some $z \geq u$ with $\varphi(z) < \infty$. Clearly, (b01)-(b02) apply to $M(z, \leq) := \{x \in M; z \leq x\}$ and $(\leq, \varphi)$. So, for the starting point $z \in M(z, \leq)$ there exists $v \in M(z, \leq)$ with

(max-1) $z \leq v$ (hence $u \leq v$)
(max-2) v is $(\leq, \varphi)$-maximal in $M(z, \leq)$
(hence: $t \in M(z, \leq)$, $v \leq t$ imply $\varphi(v) = \varphi(t)$).

Suppose by contradiction that $\gamma := \varphi(v) > 0$; and fix some β in $]0, \gamma[$. Again via (b03), there must be $y = y(v, \beta) \geq v$ (hence $y \in M(z, \leq)$), fulfilling $\varphi(y) \leq \beta < \gamma (= \varphi(v))$. This cannot be in agreement with the second conclusion above. Hence, $\varphi(v) = 0$; and we are done.

Clearly, Proposition 12 is a logical consequence of (BB). But, the converse inclusion is also true; to verify it, we need some conventions. By a (generalized) *pseudometric* over M, we shall mean any map $d : M \times M \to R_+ \cup \{\infty\}$. Suppose that we introduced such an object, with

d is *reflexive* $[d(x, x) = 0, \forall x \in M]$.

Call the point $z \in M$, $(\leq, d)$-*maximal*, if:

$u, v \in M$ and $z \leq u \leq v$ imply $d(u, v) = 0$.

Note that, if (in addition)

d is *sufficient*: $d(x, y) = 0 \Longrightarrow x = y$,

the $(\leq, d)$-maximal property becomes:

$w \in M, z \leq w \Longrightarrow z = w$ (referred to as: z is *strongly* $(\leq)$-*maximal*).

So, existence results involving such points may be viewed as "metrical" versions of the Zorn maximality principle (cf. Moore [41, Chap. 4, Sect. 4]). To get sufficient conditions for these, one may proceed as below. Let (x_n) be an ascending sequence in M. The d-Cauchy property for it is introduced in the usual way

$\forall \varepsilon > 0, \exists h(\varepsilon)$ such that $h(\varepsilon) \leq p \leq q \Longrightarrow d(x_p, x_q) \leq \varepsilon$
(or, equivalently: $d(x_m, x_n) \to 0$, as $m, n \to \infty, m \leq n$).

Also, call (x_n), *d-asymptotic* when

$\forall \varepsilon > 0, \exists k(\varepsilon)$ such that $k(\varepsilon) \leq p \Longrightarrow d(x_p, x_{p+1}) \leq \varepsilon$
(or, equivalently: $d(x_n, x_{n+1}) \to 0$, as $n \to \infty$).

Clearly, each (ascending) d-Cauchy sequence is d-asymptotic too. The reverse implication is also true when all such sequences are involved; i.e., the global conditions below are equivalent

(b04) each ascending sequence is d-Cauchy
(b05) each ascending sequence is d-asymptotic.

By definition, either of these will be referred to as $(M, \leq)$ is *regular* (modulo d). Note that this property implies its relaxed version

(b06) $((M, \leq)$ is *weakly regular* (modulo d))
$\forall x \in M, \forall \varepsilon > 0, \exists y = y(x, \varepsilon) \geq x: y \leq u \leq v \Longrightarrow d(u, v) \leq \varepsilon.$

The following ordering principle is then available (cf. Kang and Park [35]):

Proposition 13. *Assume that $(M, \leq)$ is sequentially inductive and weakly regular (modulo d). Then, for each $u \in M$ there exists a $(\leq, d)$-maximal $v \in M$ with $u \leq v$.*

Proof. Let us introduce the function (from M to $R_+ \cup \{\infty\}$)

$$\varphi_d(x) = \sup\{d(u, v); x \leq u \leq v\}, \ x \in M.$$

Clearly, φ_d is $(\leq)$-decreasing; moreover (as $(M, \leq)$ is weakly regular (modulo d)), the quasi-ordered set $(M, \leq)$ is almost regular (modulo φ_d). Hence, Proposition 12 is applicable to M and $(\leq, \varphi_d)$. This, added to

$\varphi_d(z) = 0$ if and only if z is $(\leq, d)$-maximal

gives the desired conclusion.

As a direct consequence of this, we get the maximality principle in Turinici [53] (see also Conserva and Rizzo [18]):

Proposition 14. *Assume that $(M, \leq)$ is sequentially inductive and regular (modulo d). Then, conclusion of Proposition 13 is holding.*

So far, Proposition 14 is a logical consequence of Proposition 11. The reciprocal of this is also true, by simply taking

$$d(x, y) = |\psi(x) - \psi(y)|, \ x, y \in M \text{ (where } \psi \text{ is the above one).}$$

We therefore established the inclusional chain

$$\text{Prop 11} \Longrightarrow \text{Prop 12} \Longrightarrow \text{Prop 13} \Longrightarrow \text{Prop 14} \Longrightarrow \text{Prop 11}.$$

Hence, all these ordering principles are nothing but logical equivalents of the Brezis–Browder's [10] (Proposition 11). (This also includes the related statement in Tătaru [52], which extends the one in Dancs et al. [20]). Further aspects may be found in Hamel [28, Chap. 4]; see also Hyers et al. [30, Chap. 5].

Now, a close examination of the argument in Proposition 13 shows that if the sequential inductivity condition is imposed *φ_d-asymptotically* (i.e.: to sequences (x_n) with $\varphi_d(x_n) \to 0$) its conclusion is still retainable. So, it is natural to ask whether this remark is applicable to Proposition 12 as well (with φ in place of φ_d). A positive answer to this may be given under the lines below. Let again M be some nonempty set. Take a quasi-order $(\leq)$ over it, as well as a function $\varphi : M \to R_+ \cup \{\infty\}$. The following counterpart of Proposition 12 is now available.

Proposition 15. *Assume that (b02) and (b03) are true, as well as*

(b07) $(M, \leq)$ *is sequentially inductive (modulo* φ*): each ascending sequence* (x_n) *with* $\varphi(x_n) \to 0$ *has an upper bound (modulo* $(\leq)$*).*

Then, for each $u \in M$ *there exists* $v \in M$ *with* $u \leq v$ *and* $\varphi(v) = 0$ *(hence, necessarily,* v *is* $(\leq, \varphi)$*-maximal).*

Proof. Starting from (b03), it is not hard to construct—via (DC)—an ascending (modulo $(\leq)$) sequence (u_n) in M, with ($u \leq u_0$ and)

$$\varphi(u_n) \leq 2^{-n}, \forall n \text{ (hence } \varphi(u_n) \to 0).$$

Let v stand for an upper bound (modulo $(\leq)$) of this sequence (assured by (b07)). This element has all properties we need.

Now, (b01) is a particular case of (b07). This tells us that Proposition 12 (hence Proposition 11 as well) is a particular case of Proposition 15. The reciprocal question (Prop 12 $\Longrightarrow$ Prop 15) is also true; because Proposition 15 is deductible from (DC). Further aspects may be found in Liu [40]; see also Jinag and Cho [33].

A basic particular case of these facts corresponds to the construction we already exposed. Precisely, let $d : M \times M \to R_+ \cup \{\infty\}$ be a reflexive (generalized) pseudometric (over M); and $\varphi_d : M \to R_+ \cup \{\infty\}$, its associated function (see above). Clearly, (b02) holds in this context; and the almost regularity (modulo φ_d) condition (b03) is just the one in (b06). Putting these together, it results the following maximality statement involving these data.

Proposition 16. *Assume that* $(M, \leq)$ *is sequentially inductive (modulo* φ_d*) and weakly regular (modulo* d*). Then, for each* $u \in M$ *there exists a* $(\leq, d)$*-maximal* $v \in M$ *with* $u \leq v$.

Clearly, the sequential inductivity (modulo φ_d) condition holds under (b01); wherefrom, Proposition 16 includes Proposition 13. As before, the reciprocal inclusion is also retainable; we do not give details.

Now, the pseudometric setting above is also appropriate for discussing the sequential inductivity (modulo φ_d) condition. This will necessitate some conventions. Denote by $\mathscr{S}(M)$ the class of all sequences (x_n) in M. By a (sequential) *convergence structure* on M we mean, as in Kasahara [36], any part $\mathscr{C}$ of $\mathscr{S}(M) \times M$ with the properties

(sc-1) $(x_n = x, \forall n \geq 0) \Longrightarrow ((x_n); x) \in \mathscr{C}$
(sc-2) $((x_n); x) \in \mathscr{C} \Longrightarrow ((y_n); x) \in \mathscr{C}$, for each subsequence (y_n) of (x_n).

In this case, $((x_n); x) \in \mathscr{C}$ will be denoted $x_n \overset{\mathscr{C}}{\longrightarrow} x$; and referred to as: x is the $\mathscr{C}$*-limit* of (x_n); when such elements exist, we say that (x_n) is $\mathscr{C}$*-convergent*. Assume that we fixed such an object, and let $(\leq, d)$ be taken as before. Call the subset Z of M, $(\leq)$*-closed* (modulo $\mathscr{C}$) when the $\mathscr{C}$-limit of each ascending sequence in Z is an element of it. Further, let us say that $(\leq)$ is *self-closed* (modulo $\mathscr{C}$) when $M(x, \leq)$ is $(\leq)$-closed (modulo $\mathscr{C}$), for each $x \in M$; or, equivalently: the $\mathscr{C}$-limit of each ascending sequence is an upper bound of it. Finally, term

the (reflexive) pseudometric d, $(\leq)$-*complete* (modulo $\mathscr{C}$) when each ascending d-Cauchy sequence in M is $\mathscr{C}$-convergent.

We may now give an appropriate answer to the posed question.

Proposition 17. *Suppose that* $(\leq)$ *is self-closed (modulo* $\mathscr{C}$*),* d *is* $(\leq)$*-complete (modulo* $\mathscr{C}$*), and* $(M, \leq)$ *is weakly regular (modulo* d*). Then, conclusions of Proposition 16 are retainable.*

Proof. We claim that, under the accepted conditions, Proposition 16 is applicable to $(M, \leq; d)$; precisely, that $(M, \leq)$ is sequentially inductive (modulo φ_d). Let (x_n) be an ascending sequence with $\varphi_d(x_n) \to 0$. In particular, (x_n) is an ascending d-Cauchy sequence; so that (by the $(\leq)$-completeness (modulo $\mathscr{C}$) of d), $x_n \overset{\mathscr{C}}{\longrightarrow} y$ for some $y \in M$. Combining with the self-closeness (modulo $\mathscr{C}$) of $(\leq)$ yields $x_n \leq y$, for all n; and this proves the claim.

Now, a good choice for our convergence structure is $\mathscr{C} = (\overset{d}{\longrightarrow})$, introduced as:

$$x_n \overset{d}{\longrightarrow} x \text{ whenever } d(x_n, x) \to 0 \text{ as } n \to \infty;$$

and called the *convergence* structure attached to d. For, if (in addition)

$$d \text{ is } \textit{triangular } [d(x, z) \leq d(x, y) + d(y, z), \forall x, y, z \in M],$$

Proposition 17 includes the statement by Kang and Park [35], which, in turn, includes the maximality principle by Granas and Horvath [27]. (Note, incidentally, that all applications (based on Proposition 17) discussed by these authors may be also handled via Ekeland's variational principle [23]; we do not give details.) Further aspects of structural nature may be found in Gajek and Zagrodny [24]; see also Brunner [12] and Turinici [57]. Some applications of the obtained facts to (metrical) surjectivity theory may be found in Park and Yie [44].

3.3 Bao–Mordukhovich Approach

In the following, we shall discuss the variational principle due to Bao and Mordukhovich [5, Theorem 1]. This will be done along the notations we just introduced, so as to get a better comparison with the results of introductory part.

Let $(Z, \mathscr{T})$ be a (real) linear topological space. [For the moment, a regularity assumption like

(c01) $(Z, \mathscr{T})$ is Hausdorff separated

is not assumed by the authors. However, as we shall see, this condition is indispensable to the present discussion].

Assume that K is a (convex) cone of Z, with

(c02) K is pointed ($K \cap (-K) = \{0\}$) and closed ($K = \mathrm{cl}(K)$).

Its associated relation $(\preceq_K)$ is therefore an order on Z; we shall denote it as $(\preceq)$, for simplicity. Further, take a metric space (X, d), as well as a proper multivalued map $F : X \to 2^Z$ from X to Z (identified with its *graph* in $X \times Z$). Finally, take some point $k^0 \in K \setminus \{0\}$; and let $(\preceq)$ stand for the order on $X \times Z$

$$(x_1, z_1) \preceq (x_2, z_2) \text{ if and only if } k^0 d(x_1, x_2) \leq z_1 - z_2.$$

As before, we want to determine sufficient conditions under which $(F, \preceq)$ has maximal points. Some preliminaries are needed. Remember that (in our context), given a subset V of Z, we say that $v \in V$ is a *minimal point*, provided $V(v, \geq) = \{v\}$; the class of all these will be denoted $\min(V)$. Further, define the *level-set* multivalued map $\mathscr{L} : Z \to 2^X$ attached to F as

$$\mathscr{L}(v) = \{x \in X; z \leq v, \text{ for some } z \in F(x)\}, v \in Z.$$

The underlying conditions may now be written as

(c03) (F is *quasi-bounded below*)
$F(X) \subseteq M + K$, for some closed bounded part $M \subseteq Z$
(c04) (F is *min-closed*)
$\min(F(x))$ is (nonempty) closed, for all $x \in \mathrm{Dom}(F)$
(c05) (F has the *domination property*)
$\forall x \in \mathrm{Dom}(F), \forall z \in F(x), \exists \bar{z} \in \min(F(x))$, such that $z \geq \bar{z}$.
(c06) (F is *level-closed*)
$\mathscr{L}(z)$ is closed, for each $z \in Z$.

Having these precise, we are in position to state the announced variational principle due to the quoted authors.

Theorem 14. *Suppose (under (c02)) that conditions (c03)–(c06) hold. In addition, let (X, d) be complete. Then, for each $(x_0, z_0) \in F$ there exists $(\bar{x}, \bar{z}) \in F$, with*

$$(x_0, z_0) \preceq (\bar{x}, \bar{z}), \quad \bar{z} \in \min(F(\bar{x})) \tag{38}$$

$$(\bar{x}, \bar{z}) \preceq (x', z') \in F \Longrightarrow \bar{x} = x', \bar{z} = z'. \tag{39}$$

[As a matter of fact, the quoted statement has also an extra conclusion involving the associated minimizers; but this is not essential for us].

In the following, we shall expose the proposed authors' reasoning for establishing their result. At the same time, some comments involving different stages of this argument will be inserted, so as to make precise the extra conditions under which authors' conclusions are retainable.

Proof (Theorem 14). There are several steps to be passed.

Step 1. Define the set-valued map $T : X \times Z \to 2^X$ as

$$T(x, z) = \{y \in X; (x, z) \preceq (y, v), \text{ for some } v \in F(y)\}.$$

Note that, by this very definition,

$$y \in T(x, z) \text{ iff } k^0 d(x, y) \leq z - v, \text{ for some } v \in F(y);$$

and this gives the following characterization of the mapping $T(.,.)$ above:

$$T(x,z) = \{y \in X; y \in \mathscr{L}(z - k^0 d(x,y))\},\ (x,z) \in X \times Z.$$

The basic properties of this map are

(pro-1) $T(x,z) \neq \emptyset$, whenever $(x,z) \in F$.

In fact, let $(x,z) \in F$ be arbitrary fixed; hence, $z \in F(x)$. As $(x,z) \preceq (x,z)$, we derive that $x \in T(x,z)$; and the assertion follows.

(pro-2) $T(x,z)$ is closed, for each $(x,z) \in F$.

Remark 3. This property seems to be not in general true, under the min-closed and level-closed conditions. To reach this conclusion, two possible strategies may be adopted. Let $(x,z) \in F$ be arbitrary fixed; hence (see above), $z \in F(x)$.

Strat-1. Suppose that $(Z, \mathscr{T})$ is Hausdorff separable and the min-closed condition is to be substituted by (the stronger condition)

(c07) (F is *min-compact*)
$\min(F(x))$ is (nonempty) compact, for all $x \in \mathrm{Dom}(F)$.

Let $(y_n; n \geq 0)$ be a sequence in $T(x,z)$; i.e., there must be a sequence $(w_n; n \geq 0)$ in Z, with

$$w_n \in F(y_n),\ k^0 d(x,y_n) \leq z - w_n, \text{ for all } n \geq 0.$$

Further, assume that $y_n \xrightarrow{d} y$ (i.e.: $d(y_n,y) \to 0$) as $n \to \infty$, for some $y \in X$. Let $\varepsilon > 0$ be arbitrary fixed. From this convergence property, there exists some rank $n(\varepsilon) \geq 0$, such that

$$d(y_n,y) \leq \varepsilon, \text{ for all } n \geq n(\varepsilon).$$

Combined with the preceding relations yields (by the triangular inequality),

$$k^0 d(x,y) \leq k^0 d(x,y_n) + k^0 d(y_n,y) \leq z - w_n + k^0 \varepsilon, \text{ for all } n \geq n(\varepsilon);$$

wherefrom (for the same ranks)

$$w_n \leq z - k^0 d(x,y) + k^0 \varepsilon; \text{ i.e.: } y_n \in \mathscr{L}(z - k^0 d(x,y) + k^0 \varepsilon).$$

As F is level-closed, this yields, for each $\varepsilon > 0$,

$$y \in \mathscr{L}(z - k^0 d(x,y) + k^0 \varepsilon); \text{ i.e.: } v \leq z - k^0 d(x,y) + k^0 \varepsilon, \text{ for some } v \in F(y).$$

Combining with the closeness of K, the domination property, and the closeness of $\min(F(y))$ (deductible from the compactness of the same and $(Z, \mathscr{T})$=Hausdorff separated), it follows that

$$G_\varepsilon := \min(F(y)) \cap [z - k^0 d(x,y) + k^0 \varepsilon - K] \text{ is nonempty closed, } \forall \varepsilon > 0.$$

The family $(G_\varepsilon; \varepsilon > 0)$ of (nonempty) closed subsets in the compact set $\min(F(y))$ has the finite intersection property. Hence, by a well-known characterization of compactness (cf. Kelley [37, Chap. 5]), we must have

$$G := \cap\{G_\varepsilon; \varepsilon > 0\} \text{ is nonempty closed in } \min(F(y)).$$

Now, by the closeness of K, each element $w \in G$ fulfills

$$w \in \min(F(y)) \subseteq F(y),\ w \leq z - k^0 d(x, y) \text{ (whence, } (x, z) \preceq (y, w));$$

which tells us that $y \in T(x, z)$.

Strat-2. Suppose that the level-closed condition is to be substituted by the stronger condition

(c08) (F is graph level-closed) $\mathscr{L}$ is closed in $Z \times X$:
$z_n \to z$, $x_n \xrightarrow{d} x$, $[x_n \in \mathscr{L}(z_n), \forall n]$ imply $x \in \mathscr{L}(z)$.

Let $(y_n; n \geq 0)$ be a sequence in $T(x, z)$; i.e., by the previous representation,

$$y_n \in \mathscr{L}(z - k^0 d(x, y_n)),\ \text{for all } n \geq 0. \tag{40}$$

Further, assume that, $y_n \xrightarrow{d} y$ (i.e.: $d(y_n, y) \to 0$) as $n \to \infty$, for some $y \in X$. Passing to limit as $n \to \infty$ in (40) above gives $y \in \mathscr{L}(z - k^0 d(x, y))$, which, by the same characterization, is just $y \in T(x, z)$.

Note that, from a technical viewpoint, this graph level-closed condition is comparable with the $(\preceq)$-semi-closed condition upon F. In fact, let $(x_n, z_n); n \geq 0)$ be $(\preceq)$-ascending in F and $x_n \xrightarrow{d} x$, for some $x \in X$. By definition, we have

$$k^0 d(x_p, x_{p+m}) \leq z_p - z_{p+m},\quad \forall p \geq 0, \forall m \geq 1;$$

and this in turn yields

$$x_{p+m} \in \mathscr{L}(z_p - k^0 d(x_p, x_{p+m})),\quad \forall p \geq 0, \forall m \geq 1.$$

From the (graph) closeness of $\mathscr{L}$, we have $x \in \mathscr{L}(z_p - k^0 d(x_p, x))$; hence, in particular, $x \in \mathrm{Dom}(F)$; and, moreover,

$$(\forall p \geq 0)\text{: } (x_p, z_p) \preceq (x, w_p), \text{ for some } w_p \in \min(F(x)).$$

This shows that, if

(c09) $\min(F(x))$ is a singleton, for each $x \in \mathrm{Dom}(F)$,

the $(\preceq)$-semi-closed condition follows; and the corresponding version of Theorem 14 is a particular case of Theorem 12 above, in the locally convex setting.

(pro-3) The sets $T(x,z)$ are uniformly bounded for all $(x,z) \in F$
In fact, as F is quasi-bounded below,

$$F(X) \subseteq M + K, \text{ for some closed bounded part } M \subseteq Z.$$

Then, evidently,

$$T(x,z) \subseteq \{y \in X; k^0 d(x,y) \in z - M - K\},$$

and our claim follows.

(pro-4) The following inclusion holds, for each $(x,z) \in F$,

$$T(y,v) \subseteq T(x,z), \text{ for all } y \in T(x,z) \text{ and } v \in F(y) \text{ with } (x,z) \preceq (y,v).$$

In fact, let $a \in T(y,v)$ be arbitrary fixed; hence,

$$(y,v) \preceq (a,b), \text{ for some } b \in F(a).$$

This, by the choice of our data, yields $(x,z) \preceq (a,b)$; whence, $a \in T(x,z)$; and our assertion is proved.

Step 2. Let us iteratively construct a sequence of pairs $((x_i,z_i); i \geq 0)$ in F (starting from (x_0,z_0) in the statement) as: having the i-th iteration (x_i,z_i), we select the next one (x_{i+1},z_{i+1}), according to

$$x_{i+1} \in T(x_i,z_i), \tag{41}$$

$$d(x_i,x_{i+1}) \geq \sup\{d(x_i,x); x \in T(x_i,z_i)\} - 1/(i+1), \tag{42}$$

$$z_{i+1} \in F(x_{i+1}), (x_i,z_i) \preceq (x_{i+1},z_{i+1}). \tag{43}$$

Note that, by the above properties, this iterative procedure is well defined. Summing up the inequalities in (43) gives (as F is quasi-bounded below)

$$k^0(\sum_{i=0}^{m} d(x_i,x_{i+1})) \in z_0 - z_{m+1} - K \subseteq z_0 - M - K, \ \forall m > 0.$$

Combining with the hypotheses about (K,M), we obtain

$$\sum_{i=0}^{\infty} d(x_i,x_{i+1}) < \infty, \ k^0(\sum_{i=0}^{\infty} d(x_i,x_{i+1})) \in z_0 - M - K. \tag{44}$$

On the other hand, from

$$\text{diam}T(x_{i+1},z_{i+1}) \leq \text{diam}T(x_i,z_i), \ \ \forall i$$

and the relation (obtained via (42) above)

$$\text{diam}T(x_i, z_i) \leq 2[d(x_i, x_{i+1}) + 1/(i+1)], \ \forall i,$$

one derives $\text{diam}T(x_i, z_i) \to 0$ as $i \to \infty$. As (X, d) is complete, we conclude that

$$\cap \{T(x_i, z_i); i \geq 0\} = \{\bar{x}\}, \quad \text{for some } \bar{x} \in X. \tag{45}$$

Remark 4. Relation (44) seems to hold only when

$P - K$ is closed, where $P = z_0 - M$.

Sufficient conditions for this are to be determined by the Dieudonné closeness criterion; see in this direction Zălinescu [67, Chap. 1, Sect. 1.1].

Remark 5. Relation (45) holds only if

$T(x_i, z_i)$ is closed, for all $i \geq 0$.

But, as precise, this property is not in general valid under the min-closed and level-closed conditions upon F. However, when one of the conditions below holds

(i) $(Z, \mathscr{T})$ is Hausdorff separable, and the min-closed condition is to be substituted by the (stronger) min-compact condition
(ii) the level-closed condition is to be substituted by (the stronger) graph level-closed condition,

the underlying property is retainable.

Step 3. Now, given the iterative process $((x_i, z_i); i \geq 0)$ [constructed by means of (41)–(43)], define the set sequence

$$R(x_i, z_i) = \{z \in \min(F(\bar{x})); (x_i, z_i) \preceq (\bar{x}, z)\}, \ i \geq 0.$$

The following properties hold:

(prop-1) $R(x_i, z_i)$ is nonempty closed, for each $i \geq 0$.

In fact, by (45), one has, for each $i \geq 0$,

$$k^0 d(x_i, \bar{x}) \leq z_i - \tilde{z}, \quad \text{for some } \tilde{z} \in F(\bar{x}).$$

By the domination condition, there exists $z \in \min(F(\bar{x}))$ with $\tilde{z} \geq z$. Taking the previous relation into account gives $z \in R(x_i, z_i)$; i.e., $R(x_i, z_i) \neq \emptyset$. On the other hand, by this very definition,

$$R(x_i, z_i) = \min(F(\bar{x})) \cap [z_i - k^0 d(x_i, \bar{x}) - K].$$

This, along with the closeness of (the nonempty subsets) $\min(F(x))$ and K, yields the closeness property of $R(x_i, z_i)$.

(prop-2) the sequence $(R(x_i, z_i); i \geq 0)$ is descending:

$$R(x_i, z_i) \supseteq R(x_{i+1}, z_{i+1}), \quad \text{for all } i \geq 0.$$

To verify this, pick any $z \in R(x_{i+1}, z_{i+1})$; hence, $(x_{i+1}, z_{i+1}) \preceq (\bar{x}, z)$. Combining with $(x_i, z_i) \preceq (x_{i+1}, z_{i+1})$, gives $(x_i, z_i) \preceq (\bar{x}, z)$; wherefrom, $z \in R(x_i, z_i)$.

Step 4. It follows from the above properties that

$$\emptyset \neq \cap\{R(x_i, z_i); k \geq 0\} \subseteq F(\bar{x}). \tag{46}$$

Remark 6. The nonemptiness of this intersection is not in general available under the min-closed condition upon F. But, if we suppose that

> $(Z, \mathscr{T})$ is Hausdorff separable and the min-closed condition is to be substituted by the min-compact condition,

this happens. In fact, the family $(H_i := R(x_i, z_i); i \geq 0)$ of (nonempty) closed subsets in the compact set $\min(F(\bar{x}))$ has the finite intersection property. Hence, by a well-known characterization of compactness (see above) we must have

$$H := \cap\{H_i; i \geq 0\} \text{ is nonempty closed in } \min(F(\bar{x}));$$

and our assertion follows.

Step 5. Take an arbitrary point $\bar{z}$ of this intersection. The pair $(\bar{x}, \bar{z})$ has all properties we want.

3.4 Main Result

As a conclusion of the remarks above, the arguments in Theorem 14—developed under the lines in Bao and Mordukhovich [5]—are (essentially) retainable when (in addition), the following "combined" regularity condition holds

> (h-comp) $(Z, \mathscr{T})$ is Hausdorff separable, and the min-closed condition is to be substituted by the (stronger) min-compact condition (upon F).

Clearly, the second property holds in the univalued case (modulo F); but, for the multivalued case, an assumption like this is a little stringent. Moreover, the imposed condition (in combination with the remaining ones) makes Theorem 14 be reducible to a corresponding variant of Theorem 12 involving quasi-bounded from below maps. It is our aim in the following to clarify this assertion, by means of the Brezis–Browder techniques we just developed. But, for the moment, the non-topological vector case will be considered.

Let Y be a (real) vector space. Take a (convex) cone K of it; and let $(\leq_K)$ stand for the associated quasi-order; also denoted as $(\leq)$, when K is understood. Further,

let (X, d) be a metric space; and take a proper multivalued map $F : X \to 2^Y$ from X to Y (identified with its graph in $X \times Y$). Finally, take some $k^0 \in K$; and let $(\preceq)$ stand for the quasi-order on $X \times Y$:

$$(x_1, y_1) \preceq (x_2, y_2) \text{ if and only if } k^0 d(x_1, x_2) \leq y_1 - y_2.$$

We are interested to find sufficient conditions (extending those of introductory section) under which $(F, \preceq)$ should have points with certain maximal properties. The basic assumption to be used is again the one of introductory parts; i.e.,

(d01) (F is $(\preceq)$-*semi-closed*):
if $((x_n, y_n)) \subseteq F$ is $(\preceq)$-ascending and $x_n \xrightarrow{d} x$, then $x \in \mathrm{Dom}(F)$
and there exists $y \in F(x)$ such that $(x_n, y_n) \preceq (x, y)$, for all n.

For the remaining ones, we need some conventions. Given the nonempty subset V of Y, let us say that $(y, k) \in V \times K$ is *singular*, provided

$$\text{for each } n \in N, \text{ there exists } y_n \in V, \text{ such that } nk \leq y - y_n;$$

the family of all these couples (y, k) will be denoted $\mathrm{Sing}(V; K)$. Put also

$$\mathrm{Sing}(V) = \{k \in K; (y, k) \in \mathrm{Sing}(V; K), \text{ for at least one } y \in V\};$$

each element of it will be referred to as V-*singular*. Note that the class of all such $k \in K$ is always nonempty; because $0 \in \mathrm{Sing}(V)$.

We are now in position to state our first main result.

Theorem 15. *Suppose that F is proper, $(\preceq)$-semi-closed, and*

(d02) k^0 is (K, F)-admissible: $k^0 \in K \setminus \mathrm{Sing}(F(X))$.

In addition, let (X, d) be complete. Then, for each $(x_0, y_0) \in F$ there exists $(\bar{x}, \bar{y}) \in F$ such that (34)+(35) hold.

Proof. Let $e(.,.)$ stand for the semi-metric on $X \times Y$

$$e((x_1, y_1), (x_2, y_2)) = d(x_1, x_2), (x_1, y_1), (x_2, y_2) \in X \times Y.$$

We show that Proposition 14 (see also Turinici [58]) is applicable to $(F; \preceq; e)$. Let $((x_n, y_n); n \geq 0)$ be an ascending (modulo $(\preceq)$) sequence in F:

(d03) $k^0 d(x_n, x_m) \leq y_n - y_m$, if $n \leq m$.

(I) We show that $(x_n; n \geq 0)$ is a d-Cauchy sequence in $\mathrm{Dom}(F)$; or, equivalently: $((x_n, y_n); n \geq 0)$ is an e-Cauchy sequence in F. Note that, as $d(.,.)$ is a metric, the underlying property may be written as

$$\forall \varepsilon > 0, \exists k = (\varepsilon): k < n \Longrightarrow d(x_k, x_n) < \varepsilon.$$

Assume by contradiction that this is not true; then, for some $\varepsilon > 0$, one has

(d04) for each n, there exists $m > n$, with $d(x_n, x_m) \geq \varepsilon$.

(Clearly, without loss one may assume that $0 < \varepsilon < 1$). Inductively, we get a subsequence $(u_n := x_{i(n)}; n \geq 0)$ of $(x_n; n \geq 0)$, with

$$d(u_n, u_{n+1}) \geq \varepsilon, \quad \text{for all } n \geq 0. \tag{47}$$

This yields, for the corresponding subsequence $(v_n := y_{i(n)}; n \geq 0)$ of $(y_n; n \geq 0)$, an evaluation like

$$k^0 \varepsilon \leq k^0 d(u_n, u_{n+1}) \leq v_n - v_{n+1}, \quad \text{for all } n \geq 0;$$

wherefrom (adding the first q inequalities)

$$(q\varepsilon)k^0 \leq v_0 - v_q, \quad \text{for all } q \geq 0. \tag{48}$$

Define the function $Q : N \to N$, as

$$Q(n) = \inf\{q \in N; n \leq q\varepsilon\}, n \in N; \text{ hence: } n \leq Q(n), \forall n \in N.$$

By the relation above,

$$nk^0 \leq (Q(n)\varepsilon)k^0 \leq v_0 - v_{Q(n)}, \quad \text{for all } n \geq 0; \tag{49}$$

which tells us that k^0 is $F(X)$-singular, contradiction. Hence, our working hypothesis cannot hold; i.e.: $(F, \preceq)$ is regular (modulo e).

(II) As (X, d) is complete, $x_n \xrightarrow{d} x$ as $n \to \infty$, for some $x \in X$. Taking the $(\preceq)$-semi-closed condition into account gives $x \in \mathrm{Dom}(F)$ and $[(x_n, y_n) \preceq (x, y), \forall n]$, for some $y \in F(x)$; this, by the arbitrariness of our sequence, tells us that $(F, \preceq)$ is sequentially inductive.

Summing up, Proposition 14 is indeed applicable to $(F; \preceq; e)$. Hence, by its conclusion, we derive that, for the starting $(x_0, y_0) \in F$, there exists some $(\bar{x}, \bar{y}) \in F$ with the properties (34) and

$$(\bar{x}, \bar{y}) \preceq (x', y') \in F \Longrightarrow e((\bar{x}, \bar{y}), (x', y')) = 0. \tag{50}$$

But, this is nothing else than (35); and the conclusion follows.

A useful completion of this result is obtainable under the lines of introductory part. Precisely, we have

Theorem 16. *Let the conditions in Theorem 15 be fulfilled; and F has the domination property (see above). Then, for each $(x_0, y_0) \in F$, there exists $(\bar{x}, \bar{y}) \in F$ with the properties (36) and (37).*

Proof. By Theorem 15, for the starting $(x_0, y_0) \in F$ there exists $(\tilde{x}, \tilde{y}) \in F$ fulfilling (34) and (35) (with $(\tilde{x}, \tilde{y})$ in place of $(\bar{x}, \bar{y})$). Further, by the domination property, there exists $\bar{y} \in \min(F(\tilde{x}))$ such that $\tilde{y} \geq \bar{y}$; hence, $(\tilde{x}, \tilde{y}) \preceq (\tilde{x}, \bar{y})$. It is now clear that the couple $(\bar{x}, \bar{y}) \in F$ where $\bar{x} := \tilde{x}$, has all properties we need.

Note that, if the domination property is not available, the conclusion (37) above may be written in a form involving gauge functions; see Turinici [55] for details.

3.5 *Particular Aspects*

The obtained results are given in the realm of (general) vector spaces. It will be therefore useful to see whether these extend the ones of introductory part. The answer is affirmative; but, it will necessitate some preliminaries and auxiliary facts.

Let $(Y, \mathscr{T})$ be a (real) topological vector space; note that its linear topology $\mathscr{T}$ is characterized by a fundamental system $\mathscr{B}$ of zero-neighborhoods. Take a (convex) cone K of Y; and denote its associated quasi-order as $(\preceq_K)$; or, simply, $(\preceq)$ [when K is understood]. Further, let (X, d) be a complete metric space; and take a proper multivalued function $F : X \to 2^Y$ from X to Y; identified with its graph in $X \times Y$. Finally, letting $k^0 \in K$ be a fixed element, denote by $(\preceq)$ the quasi-order on $X \times Y$

$$(x_1, y_1) \preceq (x_2, y_2) \text{ if and only if } k^0 d(x_1, x_2) \leq y_1 - y_2.$$

As in the preceding sections, we intend to get conditions under which $(F, \preceq)$ admits maximal elements. The basic one is again the $(\preceq)$-semi-closed condition upon F; for the specific ones we may proceed as follows. Call the (nonempty) part P of Y, *bounded* (modulo $\mathscr{T}$), when (cf. Cristescu [19, Chap. 1, Sect. 2])

(e01) $\forall B \in \mathscr{B}, \exists \alpha = \alpha(B) > 0$, such that $P \subseteq \alpha B$.

Denote

bound(Y)=the class of all bounded parts of Y.

The following characterization of this concept is to be noted (see the above reference for details):

Lemma 7. *The subset P of Y is bounded, if and only if*

(e02) for each sequence $(y_n; n \geq 0)$ in P and each sequence $(\lambda_n; n \geq 0)$ in R with $\lambda_n \to 0$, it is the case that $\lambda_n y_n \to 0$.

As a direct consequence, on gets a lot of basic properties for the class bound(Y). (The proof is almost evident; so, we do not give details).

Lemma 8. *Under these conventions, we have*

(bd-1) $Q \subseteq P \in \text{bound}(Y) \Longrightarrow Q \in \text{bound}(Y)$
(bd-2) $P, Q \in \text{bound}(Y) \Longrightarrow P \cup Q \in \text{bound}(Y)$
(bd-3) P=*finite* $\Longrightarrow P \in \text{bound}(Y)$
(bd-4) $P, Q \in \text{bound}(Y) \Longrightarrow P + Q \in \text{bound}(Y)$
(bd-5) $P \in \text{bound}(Y) \Longrightarrow \lambda P \in \text{bound}(Y)$, $\forall \lambda \in R$.

Now, assume in the following that

(e03) (F is *quasi-bounded below*):
$F(X) \subseteq M + K$, for some $M \in \text{bound}(Y)$.

The following auxiliary fact makes the necessary connections with the results in the preceding part.

Lemma 9. *Suppose that F is quasi-bounded below. Then*

$$\text{Sing}(F(X)) \subseteq -\text{cl}(K). \tag{51}$$

Proof. Let $k \in K$ be some arbitrary fixed point in $\text{Sing}(F(X))$. By definition,

$$Nk \subseteq y - F(X) - K, \text{ for some } y \in F(X).$$

Combining with the quasi-bounded from below condition yields

$$Nk \subseteq y - M - K, \text{ for some } y \in F(X).$$

The subset $V = y - M$ is bounded, from our preceding statement. On the other hand, by definition, there must be a sequence $(v_n; n \geq 1)$ in V such that

$$k \leq (1/n)v_n \text{ (hence, } -k + (1/n)v_n \in K), \forall n \geq 1.$$

Passing to limit as $n \to \infty$ yields (by a previous auxiliary fact) $-k \in \text{cl}(K)$; and we are done.

Now, by simply combining this with Theorem 15, one gets

Theorem 17. *Suppose that F is proper, $(\preceq)$-semi-closed, quasi-bounded below, and $k^0 \in K$ is K-admissible (see above). In addition, let (X, d) be complete. Then, for each $(x_0, y_0) \in F$ there exists $(\bar{x}, \bar{y}) \in F$ such that (34)+(35) hold.*

Proof. From the auxiliary fact above, one derives that k^0 is (K, F)-admissible; and this, along with Theorem 15, ends the argument.

On the other hand, we have (from Theorem 16) the following practical statement.

Theorem 18. *Suppose that F is proper, $(\preceq)$-semi-closed, quasi-bounded below, and $k^0 \in K$ is K-admissible. In addition, let (X, d) be complete and F have the domination property. Then, for each $(x_0, y_0) \in F$, there exists $(\bar{x}, \bar{y}) \in F$, fulfilling the relations (36) and (37).*

In the following, a basic particular case of this last result is to be developed under the lines in Bao and Mordukhovich [5].

Let $(Y, \mathscr{T})$ be a (real) Hausdorff separated topological vector space. Take a closed (convex) cone K of Y; its associated quasi-order will be denoted as $(\leq)$, for simplicity. Further, let (X, d) be a complete metric space; and take a proper multivalued function $F : X \to 2^Y$ from X to Y (identified with its graph in $X \times Y$).

Finally, take some $k^0 \in K \setminus (-K)$; hence (as K is closed) k^0 is K-admissible; and denote by $(\preceq)$ the quasi-order on $X \times Y$

$$(x_1, y_1) \preceq (x_2, y_2) \text{ if and only if } k^0 d(x_1, x_2) \leq y_1 - y_2.$$

As before, we intend to get conditions under which $(F, \preceq)$ admits maximal elements. There are two groups of such conditions

(con-g) The first group (of general conditions) is taken as in the result above:

F is quasi-bounded below, and has the domination property.

(con-s) For the second group (of specific conditions), we introduce a convention. Define the *level-set* multivalued map $\mathscr{L} : Y \to 2^X$ attached to F, as

$$\mathscr{L}(v) = \{x \in X; y \leq v, \text{ for some } y \in F(x)\}, v \in Y.$$

The underlying conditions write, in this case

(e04) (F is *level-closed*) $\mathscr{L}(y)$ is closed in X, for each $y \in Y$
(e05) (F is *min-compact*)
$\min(F(x))$ is (nonempty) compact, for all $x \in \mathrm{Dom}(F)$.

The following auxiliary fact provides the necessary connection with the conditions of our preceding statement.

Lemma 10. *Assume that F has the domination property, and is level-closed, min-compact. Then F is $(\preceq)$-semi-closed (see above).*

Proof. Let the sequence $((x_n, y_n); n \geq 0)$ in F be $(\preceq)$-ascending and $x_n \xrightarrow{d} x$, for some $x \in X$. Denote, for simplicity

$$M_i = \{v \in \min(F(x)); (x_i, y_i) \preceq (x, v)\}, i \geq 0.$$

There are two steps to be passed.

(I) We firstly claim that

$$M_i \text{ is nonempty closed, for each } i \geq 0. \tag{52}$$

The closeness property is clear, in view of

$$M_i = \min(F(x)) \cap [y_i - k^0 d(x_i, x) - K], \ i \geq 0, \tag{53}$$

combined with the closeness of K and the closeness of $\min(F(x))$ (deductible from the compactness of the same and $(Y, \mathscr{T})$=Hausdorff separated). So, all we have to establish is the nonemptiness of each member from the family $(M_i; i \geq 0)$. Fix $i \geq 0$; and note that, by the $(\preceq)$-ascending property, we have

$$k^0 d(x_i, x_n) \leq y_i - y_n, \ \forall n \geq i.$$

Let $\varepsilon > 0$ be arbitrary fixed. From the imposed convergence property, there exists some rank $n(\varepsilon) \geq i$, such that

$$d(x_n, x) \leq \varepsilon, \text{ for all } n \geq n(\varepsilon).$$

Combined with the preceding relations yields (by the triangular inequality)

$$k^0 d(x_i, x) \leq k^0 d(x_i, x_n) + k^0 d(x_n, x) \leq y_i - y_n + k^0 \varepsilon, \text{ for all } n \geq n(\varepsilon);$$

wherefrom (for the same ranks)

$$y_n \leq y_i - k^0 d(x_i, x) + k^0 \varepsilon; \text{ i.e.: } x_n \in \mathscr{L}(y_i - k^0 d(x_i, x) + k^0 \varepsilon).$$

As F is level-closed, this yields, for each $\varepsilon > 0$,

$$\begin{array}{l} x \in \mathscr{L}(y_i - k^0 d(x_i, x) + k^0 \varepsilon); \text{ wherefrom:} \\ v \leq y_i - k^0 d(x_i, x) + k^0 \varepsilon, \text{ for some } v \in F(x); \end{array}$$

hence, in particular, $x \in \mathrm{Dom}(F)$. Combining with the closeness of K, the domination property, and the closeness of $\min(F(x))$ (see above), it follows that

$$G_\varepsilon := \min(F(x)) \cap [y_i - k^0 d(x_i, x) + k^0 \varepsilon - K] \text{ is nonempty closed, } \forall \varepsilon > 0.$$

The family $(G_\varepsilon; \varepsilon > 0)$ of (nonempty) closed subsets in the compact set $\min(F(x))$ has the finite intersection property. Hence, by a well-known characterization of compactness (cf. Kelley [37, Chap. 5]), we must have

$$G := \cap\{G_\varepsilon; \varepsilon > 0\} \text{ is (nonempty) closed in } \min(F(x)).$$

Now, by the closeness of K, each element $v \in G$ fulfills

$$v \in \min(F(x)) \subseteq F(x), \ v \leq y_i - k^0 d(x_i, x) \text{ (whence, } (x_i, y_i) \preceq (x, v));$$

which tells us that $v \in M_i$; whence, M_i is nonempty (for each $i \geq 0$).

(II) We now claim that $(M_i; i \geq 0)$ is a descending sequence of sets; i.e.,

$$M_i \supseteq M_j, \text{ whenever } i \leq j. \tag{54}$$

In fact, let the ranks $i, j \geq 0$ be such that $i \leq j$; and take some $v \in M_j$; hence, by definition $v \in \min(F(x))$ and $(x_j, y_j) \preceq (x, v)$. Combining with $(x_i, y_i) \preceq (x_j, y_j)$, one derives $(x_i, y_i) \preceq (x, v)$; so that, $v \in M_i$; and the claim follows.

Summing up, the family $(M_i; i \geq 0)$ of (nonempty) closed subsets in the compact set $\min(F(x))$ has the finite intersection property. Hence, by the characterization of compactness we just evoked,

$$M := \cap\{M_i; i \geq 0\} \text{ is nonempty closed in } \min(F(x)).$$

Now, evidently, any $v \in M$ fulfills

$$v \in M_i \text{ (that is: } (x_i, y_i) \preceq (x, v)), \text{ for all } i \geq 0.$$

This, along with the arbitrariness of the $(\preceq)$-ascending sequence $((x_n, y_n); n \geq 0)$ and $x \in X$ with $x_n \xrightarrow{d} x$, proves the desired conclusion.

Now, by simply combining this with Theorem 18, one derives the following practical statement.

Theorem 19. *Suppose that F is proper, quasi-bounded below, and $k^0 \in K\backslash(-K)$. In addition, let (X, d) be complete and F be level-closed, min-compact, and having the domination property. Then, for each $(x_0, y_0) \in F$, there exists $(\bar{x}, \bar{y}) \in F$, fulfilling the relations (36) and (37).*

This result may be viewed as a (corrected) simplified version of Theorem 14; due, as above said, to Bao and Mordukhovich [5]. In particular, when F is bounded below, the obtained facts include in a direct way the statements in Tammer and Zălinescu [50]. Some transfinite versions of these may be found in Nemeth [43] and Khanh [38]; see also Turinici [60].

References

1. Altman, M.: A generalization of the Brezis-Browder principle on ordered sets. Nonlinear Anal. **6**, 157–165 (1982)
2. Anisiu, M.C.: On maximality principles related to Ekeland's theorem. In: Seminar Funct. Analysis Numer. Meth. (Faculty of Math. Research Seminars), Preprint No. 1. "Babeş-Bolyai" Univ., Cluj-Napoca (România) (1987)
3. Bae, J.S., Cho, E.W., Yeom, S.H.: A generalization of the Caristi-Kirk fixed point theorem and its applications to mapping theorems. J. Korean Math. Soc. **31**, 29–48 (1994)
4. Bao, T.Q., Khanh, P.Q.: Are several recent generalizations of Ekeland's variational principle more general than the original principle? Acta Math. Vietnam. **28**, 345–350 (2003)
5. Bao, T.Q., Mordukhovich, B.S.: Variational principles for set-valued mappings with applications to multiobjective optimization. Control Cyb. **36**, 531–562 (2007)
6. Bejancu, A.: On the Ekeland and Borwein-Preiss principles in finite dimensions. An. Şt. Univ. "A. I. Cuza" Iaşi (S. I-a, Mat.) **40**, 63–67 (1994)
7. Bernays, P.: A system of axiomatic set theory: part III. Infinity and enumerability analysis. J. Symb. Log. **7**, 65–89 (1942)
8. Borwein, J.M., Preiss, D.: A smooth variational principle with applications to subdifferentiability and to differentiability of convex functions. Trans. Am. Math. Soc. **303**, 517–527 (1987)
9. Bourbaki, N.: Sur le théorème de Zorn. Arch. Math. **2**, 434–437 (1949/1950)
10. Brezis, H., Browder, F.E.: A general principle on ordered sets in nonlinear functional analysis. Adv. Math. **21**, 355–364 (1976)
11. Brøndsted, A.: Fixed points and partial orders. Proc. Am. Math. Soc. **60**, 365–366 (1976)
12. Brunner, N.: Topologische Maximalprinzipien. Z. Math. Logik Grundl. Math. **33**, 135–139 (1987)

13. Caristi, J., Kirk, W.A.: Geometric fixed point theory and inwardness conditions. In: The Geometry of Metric and Linear Spaces (Michigan State Univ., 1974). Lecture Notes in Mathematics, vol. 490, pp. 74–83. Springer, Berlin (1975)
14. Cârjă, O., Necula, M., Vrabie, I.I.: Viability, Invariance and Applications. North Holland Mathematics Studies, vol. 207. Elsevier B. V., Amsterdam (2007)
15. Chen, G.Y., Huang, X.X., Hou, S.H.: General Ekeland's variational principle for set-valued mappings. J. Optim. Theory Appl. **106**, 151–164 (2000)
16. Chen, G.Y., Huang, X.X., Hou, S.H.: General Ekeland's variational principle for set-valued mappings [Errata Corrige]. J. Optim. Theory Appl. **117**, 217–218 (2003)
17. Cohen, P.J.: Set Theory and the Continuum Hypothesis. Benjamin, New York (1966)
18. Conserva, V., Rizzo, S.: Maximal elements in a class of order complete metric subspaces. Math. Jpn. **37**, 515–518 (1992)
19. Cristescu, R.: Topological Vector Spaces. Noordhoff Intl. Publishers, Leyden (1977)
20. Dancs, S., Hegedus, M., Medvegyev, P.: A general ordering and fixed-point principle in complete metric space. Acta Sci. Math. (Szeged) **46**, 381–388 (1983)
21. Du, W.S.: On some nonlinear problems induced by an abstract maximal element principle. J. Math. Anal. Appl. **347**, 391–399 (2008)
22. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)
23. Ekeland, I.: Nonconvex minimization problems. Bull. Am. Math. Soc. (New Ser.) **1**, 443–474 (1979)
24. Gajek, L., Zagrodny, D.: Countably orderable sets and their application in optimization. Optimization **26**, 287–301 (1992)
25. Goepfert, A., Tammer, C., Zălinescu, C.: On the vectorial Ekeland's variational principle and minimal points in product spaces. Nonlinear Anal. **39**, 909–922 (2000)
26. Goepfert, A., Riahi, H., Tammer, C., Zălinescu, C.: Variational Methods in Partially Ordered Spaces. Canadian Mathematical Society Books in Mathematics, vol. 17. Springer, New York (2003)
27. Granas, A., Horvath, C.D.: On the order-theoretic Cantor theorem. Taiwan. J. Math. **4**, 203–213 (2000)
28. Hamel, A.: Variational Principles on Metric and Uniform Spaces. Habilitation Thesis, Martin-Luther University, Halle-Wittenberg (2005)
29. Hamel, A.H., Tammer, C.: Minimal elements for product orders. Optimization **57**, 263–275 (2008)
30. Hyers, D.H., Isac, G., Rassias, T.M.: Topics in Nonlinear Analysis and Applications. World Scientific Publ., Singapore (1997)
31. Isac, G.: Sur l'existence de l'optimum de Pareto. Rivista Mat. Univ. Parma (Serie IV) **9**, 303–325 (1983)
32. Isac, G.: The Ekeland's principle and Pareto ε-efficiency. In: Tamiz, M. (ed.) Multi-Objective Programming and Goal Programming. Lecture Notes in Economics and Mathematical Systems, vol. 432, pp. 148–163. Springer, Berlin (1996)
33. Jinag, G.J., Cho, Y.J.: Cantor order and completeness. Int. J. Pure Appl. Math. **2**, 393–398 (2002)
34. Kada, O., Suzuki, T., Takahashi, W.: Nonconvex minimization theorems and fixed point theorems in complete metric spaces. Math. Jpn. **44**, 381–391 (1996)
35. Kang, B.G., Park, S.: On generalized ordering principles in nonlinear analysis. Nonlinear Anal. **14**, 159–165 (1990)
36. Kasahara, S.: On some generalizations of the Banach contraction theorem. Publ. Res. Inst. Math. Sci. Kyoto Univ. **12**, 427–437 (1976)
37. Kelley, J.L.: General Topology. Springer, Berlin (1975)
38. Khanh, P.Q.: On Caristi-Kirk's theorem and Ekeland's variational principle for Pareto extrema. Bull. Polish Acad. Sci. (Math.) **37**, 33–39 (1989)
39. Li, Y., Shi, S.: A generalization of Ekeland's ε-variational principle and its Borwein-Preiss smooth variant. J. Math. Anal. Appl. **246**, 308–319 (2000)
40. Liu, Z.: Order completeness and stationary points. Rostock Math. Kolloq. **50**, 85–88 (1997)

41. Moore, G.H.: Zermelo's Axiom of Choice: Its Origin, Development and Influence. Springer, New York (1982)
42. Moskhovakis, Y.: Notes on Set Theory. Springer, New York (2006)
43. Nemeth, A.B.: A nonconvex vector minimization problem. Nonlinear Anal. **10**, 669–678 (1986)
44. Park, J.A., Yie, S.: Surjectivity of generalized locally expansive maps. J. Korean Math. Soc. **24**, 179–185 (1987)
45. Pasicki, L.: Transitivity and variational principles. Nonlinear Anal. **74**, 5678–5684 (2011)
46. Peressini, A.L.: Ordered Topological Vector Spaces. Harper and Row Publ., New York (1967)
47. Schechter, E.: Handbook of Analysis and Its Foundation. Academic Press, New York (1997)
48. Suzuki, T.: Generalized distance and existence theorems in complete metric spaces. J. Math. Anal. Appl. **253**, 440–458 (2001)
49. Szaz, A.: An improved Altman type generalization of the Brezis Browder ordering principle. Math. Commun. **12**, 155–161 (2007)
50. Tammer, C., Zălinescu, C.: Vector variational principles for set-valued functions. Optimization **60**, 839–857 (2011)
51. Tarski, A.: Axiomatic and algebraic aspects of two theorems on sums of cardinals. Fundam. Math. **35**, 79–104 (1948)
52. Tătaru, D.: Viscosity solutions of Hamilton-Jacobi equations with unbounded nonlinear terms. J. Math. Anal. Appl. **163**, 345–392 (1992)
53. Turinici, M.: Metric variants of the Brezis-Browder ordering principle. Demonstr. Math. **22**, 213–228 (1989)
54. Turinici, M.: Vector extensions of the variational Ekeland's result. An. Şt. Univ. "A. I. Cuza" Iaşi (S I-a: Mat) **40**, 225–266 (1994)
55. Turinici, M.: Minimal points in product spaces. An. Şt. Univ. "Ovidius" Constanţa (Math.) **10**, 109–122 (2002)
56. Turinici, M.: Relational Brezis-Browder principles. Fixed Point Theory **7**, 111–126 (2006)
57. Turinici, M.: Brezis-Browder principles in separable ordered sets. Libertas Math. **26**, 15–30 (2006)
58. Turinici, M.: Brezis-Browder principle revisited. Note Mat. **28**, 33–41 (2008)
59. Turinici, M.: Variational statements on KST-metric structures. An. Şt. Univ. "Ovidius" Constanţa (Mat.) **17**, 231–246 (2009)
60. Turinici, M.: GTZ principles and cone-valued metrics. Libertas Math. **29**, 17–36 (2009)
61. Turinici, M.: Function variational principles and coercivity over normed spaces. Optimization **59**, 199–222 (2010)
62. Turinici, M.: Brezis-Browder principle and dependent choice. An Şt. Univ. "Al. I. Cuza" Iaşi (Mat.) **57**, 263–277 (2011)
63. Turinici, M.: Smooth variational principles and diagonal dependent choices. Bul. Inst. Polit. Iaşi (Sect. Mat., Mec. Teor., Fiz.) **57**(61), 269–282 (2011)
64. Turinici, M.: Vector maximal principles and dependent choice. Libertas Math. **31**, 35–48 (2011)
65. Turinici, M.: Almost metric versions of Zhong's variational principle. Mat. Vesnik **65**, 519–532 (2013)
66. Wolk, E.S.: On the principle of dependent choices and some forms of Zorn's lemma. Canad. Math. Bull. **26**, 365–367 (1983)
67. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific Publishing, Singapore (2002)
68. Zhong, C.K.: A generalization of Ekeland's variational principle and application to the study of the relation between the weak P. S. condition and coercivity. Nonlinear Anal. **29**, 1421–1431 (1997)
69. Zhu, J., Li, S.J.: Generalization of ordering principles and applications. J. Optim. Theory Appl. **132**, 493–507 (2007)
70. Zhu, J., Zhong, C.K., Cho, Y.J.: Generalized variational principle and vector optimization. J. Optim. Theory Appl. **106**, 201–217 (2000)
71. Zorn, M.: A remark on method in transfinite algebra. Bull. Am. Math. Soc. **41**, 667–670 (1935)

All Functions $g: \mathbb{N} \to \mathbb{N}$ Which have a Single-Fold Diophantine Representation are Dominated by a Limit-Computable Function $f: \mathbb{N} \setminus \{0\} \to \mathbb{N}$ Which is Implemented in *MuPAD* and Whose Computability is an Open Problem

Apoloniusz Tyszka

Abstract Let $E_n = \{x_k = 1,\ x_i + x_j = x_k,\ x_i \cdot x_j = x_k : i,j,k \in \{1,\ldots,n\}\}$. For any integer $n \geq 2214$, we define a system $T \subseteq E_n$ which has a unique integer solution $(a_1,\ldots,a_n)$. We prove that the numbers $a_1,\ldots,a_n$ are positive and $\max(a_1,\ldots,a_n) > 2^{2^n}$. For a positive integer n, let $f(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a unique solution in non-negative integers $x_1,\ldots,x_n$, this solution belongs to $[0,b]^n$. We prove that if a function $g: \mathbb{N} \to \mathbb{N}$ has a single-fold Diophantine representation, then f dominates g. We present a *MuPAD* code which takes as input a positive integer n, performs an infinite loop, returns a non-negative integer on each iteration, and returns $f(n)$ on each sufficiently high iteration.

Keywords: Davis-Putnam-Robinson-Matiyasevich theorem • Diophantine equation with a unique integer solution • Diophantine equation with a unique solution in non-negative integers • Limit-computable function • Single-fold Diophantine representation • Trial-and-error computable function

Let $E_n = \{x_k = 1,\ x_i + x_j = x_k,\ x_i \cdot x_j = x_k : i,j,k \in \{1,\ldots,n\}\}$. The following system

A. Tyszka (✉)
University of Agriculture, Faculty of Production and Power Engineering,
Balicka 116B, 30-149 Kraków, Poland
e-mail: rttyszka@cyf-kr.edu.pl

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_24

$$\begin{cases} x_1 = 1 \\ x_1 + x_1 = x_2 \\ x_2 \cdot x_2 = x_3 \\ x_3 \cdot x_3 = x_4 \\ x_4 \cdot x_4 = x_5 \\ \dots \\ x_{n-1} \cdot x_{n-1} = x_n \end{cases}$$

has a unique complex solution, namely $\left(1, 2, 4, 16, 256, \dots, 2^{2^{n-3}}, 2^{2^{n-2}}\right)$. The following system

$$\begin{cases} x_1 + x_1 = x_2 \\ x_1 \cdot x_1 = x_2 \\ x_2 \cdot x_2 = x_3 \\ x_3 \cdot x_3 = x_4 \\ \dots \\ x_{n-1} \cdot x_{n-1} = x_n \end{cases}$$

has exactly two complex solutions, namely:
$(0, \dots, 0)$ and $\left(2, 4, 16, 256, \dots, 2^{2^{n-2}}, 2^{2^{n-1}}\right)$.

Theorem 1. *For each integer $n \geq 2203$, the following system T*

$$\begin{cases} (T_1) & \forall i \in \{1, \dots, n\}\ x_i \cdot x_i = x_{i+1} \\ (T_2) & x_{n+2} + x_{n+2} = x_{n+3} \\ (T_3) & x_{n+3} + x_{n+3} = x_{n+4} \\ (T_4) & x_{n+4} + x_{n+2} = x_{n+5} \\ (T_5) & x_{n+6} = 1 \\ (T_6) & x_{n+5} + x_{n+6} = x_{n+7} \\ (T_7) & x_{n+7} + x_{n+6} = x_{n+8} \\ (T_8) & x_{n+8} + x_{n+6} = x_1 \\ (T_9) & x_{n+8} \cdot x_{n+8} = x_{n+9} \\ (T_{10}) & x_{n+9} \cdot x_{n+10} = x_{n+11} \\ (T_{11}) & x_{n+11} + x_1 = x_{2204} \end{cases}$$

has a unique integer solution $(a_1, \dots, a_{n+11})$. The numbers $a_1, \dots, a_{n+11}$ are positive and $\max\left(a_1, \dots, a_{n+11}\right) > 2^{2^{n+11}}$.

Proof. Equations (T_2)–(T_7) imply that $x_{n+8} = 5x_{n+2} + 2$. Hence, $x_{n+8} \notin \left\{-1, 0, 1, -2^{2203} + 1\right\}$. The system (T_1) implies that $x_{n+1} = {x_1}^{2^n}$ and ${x_1}^{2^{2203}} = x_{2204}$. By this and equations (T_5) and (T_8)–(T_{11}), we get:

$$\left(x_{n+8}+1\right)^{2^{2203}} = \left(x_{n+8}+x_{n+6}\right)^{2^{2203}} = {x_1}^{2^{2203}} = x_{2204} = x_{n+11} + x_1 =$$

$$(x_{n+9}\cdot x_{n+10}) + x_1 = \left(x_{n+8}^2 \cdot x_{n+10}\right) + (x_{n+8}+x_{n+6}) = x_{n+8}^2 \cdot x_{n+10} + x_{n+8} + 1 \tag{1}$$

Next,

$$\left(x_{n+8}+1\right)^{2^{2203}} = 1 + 2^{2203}\cdot x_{n+8} + x_{n+8}^2 \cdot \sum_{k=2}^{2^{2203}} \binom{2^{2203}}{k} \cdot x_{n+8}^{k-2} \tag{2}$$

Formulae (1) and (2) give:

$$x_{n+8}^2 \cdot \left(x_{n+10} - \sum_{k=2}^{2^{2203}} \binom{2^{2203}}{k} \cdot x_{n+8}^{k-2}\right) = \left(2^{2203}-1\right)\cdot x_{n+8}$$

The number $2^{2203}-1$ is prime [8, pp. 79 and 81] and $x_{n+8} \notin \left\{-1, 0, 1, -2^{2203}+1\right\}$. Hence, $x_{n+8} = 2^{2203}-1$. This proves that exactly one integer tuple $(x_1, \ldots, x_{n+11})$ solves T and the numbers $x_1, \ldots, x_{n+11}$ are positive. Next, $x_1 = x_{n+8} + x_{n+6} = \left(2^{2203}-1\right)+1 = 2^{2203}$, and finally

$$x_{n+1} = {x_1}^{2^n} = \left(2^{2203}\right)^{2^n} > \left(2^{2048}\right)^{2^n} = 2^{2^{n+11}}$$

Explicitly, the whole solution is given by

$$\begin{cases} \forall i \in \{1, \ldots, n+1\}\ a_i = \left(2^{2203}\right)^{2^{i-1}} \\ a_{n+2} = \frac{1}{5}\cdot\left(2^{2203}-3\right) \\ a_{n+3} = \frac{2}{5}\cdot\left(2^{2203}-3\right) \\ a_{n+4} = \frac{4}{5}\cdot\left(2^{2203}-3\right) \\ a_{n+5} = 2^{2203}-3 \\ a_{n+6} = 1 \\ a_{n+7} = 2^{2203}-2 \\ a_{n+8} = 2^{2203}-1 \\ a_{n+9} = \left(2^{2203}-1\right)^2 \\ a_{n+10} = 1 + \sum\limits_{k=2}^{2^{2203}} \binom{2^{2203}}{k} \cdot \left(2^{2203}-1\right)^{k-2} \\ a_{n+11} = \left(2^{2203}\right)^{2^{2203}} - 2^{2203} \end{cases}$$

□

If we replace the equation (T_5) by the system $\forall i \in \{1,\dots,n+11\}\, x_{n+6}\cdot x_i = x_i$, then the system T contains only equations of the form $x_i + x_j = x_k$ or $x_i \cdot x_j = x_k$, and exactly two integer tuples solve T, namely $(0,\dots,0)$ and $(a_1,\dots,a_{n+11})$. Theorem 1 disproves the conjecture in [10], where the author proposed the upper bound $2^{2^{n-1}}$ for positive integer solutions to any system

$$S \subseteq \{x_i + x_j = x_k,\ x_i \cdot x_j = x_k\colon\ i,j,k \in \{1,\dots,n\}\}$$

which has only finitely many solutions in positive integers $x_1,\dots,x_n$. Theorem 1 disproves the conjecture in [11], where the author proposed the upper bound $2^{2^{n-1}}$ for modulus of integer solutions to any system $S \subseteq E_n$ which has only finitely many solutions in integers $x_1,\dots,x_n$. For each integer $n \geq 2$, the following system

$$\begin{cases} \forall i \in \{1,\dots,n\}\ x_i \cdot x_i = x_{i+1} \\ x_{n+2} = 1 \\ x_{n+2} + x_{n+2} = x_{n+3} \\ x_{n+3} + x_{n+3} = x_{n+4} \\ x_{n+4} + x_{n+5} = x_{n+6} \\ x_{n+6} + x_{n+2} = x_1 \\ x_{n+6} \cdot x_{n+6} = x_{n+7} \\ x_{n+8} + x_{n+8} = x_{n+9} \\ x_{n+9} + x_{n+2} = x_{n+10} \\ x_{n+7} \cdot x_{n+10} = x_{n+11} \\ x_{n+11} + x_{n+2} = x_{n+1} \end{cases}$$

has a unique solution $(a_1,\dots,a_{n+11})$ in non-negative integers [1]. The proof of this gives also that $a_{n+1} > 2^{2^{(n+11)-2}}$ for any $n \geq 512$ [1]. The above-described result inspired the author to formulate Theorem 1 and the next Theorem 2.

Theorem 2. *If $n \in \mathbb{N}$ and $2^n - 1$ is prime, then the following system*

$$\begin{cases} \forall i \in \{1,\dots,n\}\ x_i \cdot x_i = x_{i+1} \\ x_{n+2} = 1 \\ x_{n+3} + x_{n+2} = x_{n+4} \\ x_{n+4} + x_{n+2} = x_{n+5} \\ x_{n+5} + x_{n+2} = x_1 \\ x_{n+5} \cdot x_{n+5} = x_{n+6} \\ x_{n+6} \cdot x_{n+7} = x_{n+8} \\ x_{n+8} + x_1 = x_{n+1} \end{cases}$$

has a unique solution $(x_1,\dots,x_{n+8})$ in non-negative integers and $x_{n+1} = (2^n)^{2^n}$.

Proof. The proof is analogous to that of Theorem 1. □

The Davis–Putnam–Robinson–Matiyasevich theorem states that every recursively enumerable set $\mathscr{M} \subseteq \mathbb{N}^n$ has a Diophantine representation, that is

$$(a_1, \ldots, a_n) \in \mathscr{M} \Longleftrightarrow \exists x_1, \ldots, x_m \in \mathbb{N} \;\; W(a_1, \ldots, a_n, x_1, \ldots, x_m) = 0 \qquad (\mathtt{R})$$

for some polynomial W with integer coefficients, see [4]. The polynomial W can be computed, if we know the Turing machine M such that, for all $(a_1, \ldots, a_n) \in \mathbb{N}^n$, M halts on $(a_1, \ldots, a_n)$ if and only if $(a_1, \ldots, a_n) \in \mathscr{M}$, see [4]. The representation (R) is said to be single-fold, if for any $a_1, \ldots, a_n \in \mathbb{N}$ the equation $W(a_1, \ldots, a_n, x_1, \ldots, x_m) = 0$ has at most one solution $(x_1, \ldots, x_m) \in \mathbb{N}^m$. Y. Matiyasevich conjectures that each recursively enumerable set $\mathscr{M} \subseteq \mathbb{N}^n$ has a single-fold Diophantine representation, see [2, pp. 341–342], [5, p. 42], [6, p. 79], and [7, p. 745].

Let us say that a set $\mathscr{M} \subseteq \mathbb{N}^n$ has a bounded Diophantine representation, if there exists a polynomial W with integer coefficients such that

$$(a_1, \ldots, a_n) \in \mathscr{M} \Longleftrightarrow$$

$$\exists x_1, \ldots, x_m \in \{0, \ldots, \max(a_1, \ldots, a_n)\} \;\; W(a_1, \ldots, a_n, x_1, \ldots, x_m) = 0$$

Of course, any bounded Diophantine representation is finite-fold and any subset of $\mathbb{N}$ with a bounded Diophantine representation is computable. A simple diagonal argument shows that there exists a computable subset of $\mathbb{N}$ without any bounded Diophantine representation, see [2, p. 360]. The authors of [2] suggest a possibility that each subset of $\mathbb{N}$ which has a finite-fold Diophantine representation has also a bounded Diophantine representation, see [2, p. 360].

Let ω denote the least infinite cardinal number, and let ω_1 denote the least uncountable cardinal number. Let $\kappa \in \{2, 3, 4, \ldots, \omega, \omega_1\}$. We say that the representation (R) is κ-fold, if for any $a_1, \ldots, a_n \in \mathbb{N}$ the equation $W(a_1, \ldots, a_n, x_1, \ldots, x_m) = 0$ has less than κ solutions $(x_1, \ldots, x_m) \in \mathbb{N}^m$. Of course, 2-fold Diophantine representations are identical to single-fold Diophantine representations. Next, ω-fold Diophantine representations are identical to finite-fold Diophantine representations. Finally, ω_1-fold Diophantine representations are identical to Diophantine representations.

For a positive integer n, let $f_\kappa(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ which has a solution in non-negative integers $x_1, \ldots, x_n$ and which has less than κ solutions in non-negative integers $x_1, \ldots, x_n$, there exists a solution of S in non-negative integers not greater than b. For a positive integer n, let $f(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with a unique solution in non-negative integers $x_1, \ldots, x_n$, this solution belongs to $[0, b]^n$. Obviously, $f = f_2$, $f(1) = 1$, and $f(2) = 2$.

Lemma 1 ([3]). *If $k \in \mathbb{N}$, then the equation $x^2 + 1 = 5^{2k+1} \cdot y^2$ has infinitely many solutions in non-negative integers. The minimal solution is given by*

$$x = \frac{\left(2+\sqrt{5}\right)^{5^k} + \left(2-\sqrt{5}\right)^{5^k}}{2}$$

$$y = \frac{\left(2+\sqrt{5}\right)^{5^k} - \left(2-\sqrt{5}\right)^{5^k}}{2 \cdot \sqrt{5} \cdot 5^k}$$

Theorem 3. *For each positive integer n, the following system*

$$\begin{cases} \forall i \in \{1,\ldots,n\}\ x_i \cdot x_i = x_{i+1} \\ x_1 \cdot x_{n+1} = x_{n+2} \\ x_{n+3} = 1 \\ x_{n+3} + x_{n+3} = x_{n+4} \\ x_{n+4} + x_{n+4} = x_{n+5} \\ x_{n+5} + x_{n+3} = x_1 \\ x_{n+6} \cdot x_{n+6} = x_{n+7} \\ x_{n+8} \cdot x_{n+8} = x_{n+9} \\ x_{n+9} + x_{n+3} = x_{n+10} \\ x_{n+2} \cdot x_{n+7} = x_{n+10} \end{cases}$$

has infinitely many solutions in non-negative integers $x_1,\ldots,x_{n+10}$. *If an integer tuple* $(x_1,\ldots,x_{n+10})$ *solves the system, then*

$$x_{n+10} \geq \left(\frac{\left(2+\sqrt{5}\right)^{5^{2^n-1}} + \left(2-\sqrt{5}\right)^{5^{2^{n-1}}}}{2} \right)^2 + 1$$

Proof. It follows from Lemma 1, because the system equivalently expresses that $x_{n+10} = x_8^2 + 1 = 5^{2 \cdot 2^{n-1}+1} \cdot x_{n+6}^2 + 1$. □

Let *Rng* denote the class of all rings $\boldsymbol{K}$ that extend $\mathbb{Z}$.

Lemma 2 ([10]). *Let* $D(x_1,\ldots,x_p) \in \mathbb{Z}[x_1,\ldots,x_p]$. *Assume that* $\deg(D,x_i) \geq 1$ *for each* $i \in \{1,\ldots,p\}$. *We can compute a positive integer* $n > p$ *and a system* $T \subseteq E_n$ *which satisfies the following two conditions:*

`Condition 1.` *If* $\boldsymbol{K} \in Rng \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$, *then*

$$\forall \tilde{x}_1,\ldots,\tilde{x}_p \in \boldsymbol{K}\ \Big(D(\tilde{x}_1,\ldots,\tilde{x}_p) = 0 \iff$$

$$\exists \tilde{x}_{p+1},\ldots,\tilde{x}_n \in \boldsymbol{K}\ (\tilde{x}_1,\ldots,\tilde{x}_p,\tilde{x}_{p+1},\ldots,\tilde{x}_n)\ \textit{solves}\ T \Big)$$

Condition 2. *If $\boldsymbol{K} \in \mathit{Rng} \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$, then for each $\tilde{x}_1, \ldots, \tilde{x}_p \in \boldsymbol{K}$ with $D(\tilde{x}_1, \ldots, \tilde{x}_p) = 0$, there exists a unique tuple $(\tilde{x}_{p+1}, \ldots, \tilde{x}_n) \in \boldsymbol{K}^{n-p}$ such that the tuple $(\tilde{x}_1, \ldots, \tilde{x}_p, \tilde{x}_{p+1}, \ldots, \tilde{x}_n)$ solves T.*

Conditions 1 and 2 imply that for each $\boldsymbol{K} \in \mathit{Rng} \cup \{\mathbb{N},\ \mathbb{N} \setminus \{0\}\}$, the equation $D(x_1, \ldots, x_p) = 0$ and the system T have the same number of solutions in $\boldsymbol{K}$.

Theorems 2 and 3 provide a heuristic argument that the function f_{ω_1} grows much faster than the function f. The next Theorem 4 for $\kappa = \omega_1$ implies that the function f_{ω_1} is not computable. These facts lead to the conjecture that the function f is computable. By this, Theorem 4 for $\kappa = 2$ is the first step towards disproving Matiyasevich's conjecture on single-fold Diophantine representations.

Theorem 4. *If a function $g\colon \mathbb{N} \to \mathbb{N}$ has a κ-fold Diophantine representation, then there exists a positive integer m such that $g(n) < f_{\kappa}(n)$ for any $n \geq m$.*

Proof. By Lemma 2 for $\boldsymbol{K} = \mathbb{N}$, there is an integer $s \geq 3$ such that for any non-negative integers x_1, x_2,

$$(x_1, x_2) \in g \Longleftrightarrow \exists x_3, \ldots, x_s \in \mathbb{N}\ \ \Phi(x_1, x_2, x_3, \ldots, x_s), \tag{E}$$

where the formula $\Phi(x_1, x_2, x_3, \ldots, x_s)$ is a conjunction of formulae of the forms $x_k = 1$, $x_i + x_j = x_k$, $x_i \cdot x_j = x_k$ $(i, j, k \in \{1, \ldots, s\})$, and for each non-negative integers x_1, x_2 less than κ tuples $(x_3, \ldots, x_s) \in \mathbb{N}^{s-2}$ satisfy $\Phi(x_1, x_2, x_3, \ldots, x_s)$. Let $[\cdot]$ denote the integer part function. For each integer $n \geq 6 + 2s$,

$$n - \left[\frac{n}{2}\right] - 3 - s \geq 6 + 2s - \left[\frac{6 + 2s}{2}\right] - 3 - s \geq 6 + 2s - \frac{6 + 2s}{2} - 3 - s = 0$$

For an integer $n \geq 6 + 2s$, let S_n denote the following system

$$\begin{cases} \text{all equations occurring in } \Phi(x_1, x_2, x_3, \ldots, x_s) \\ n - \left[\frac{n}{2}\right] - 3 - s \ \text{ equations of the form } z_i = 1 \\ t_1 = 1 \\ t_1 + t_1 = t_2 \\ t_2 + t_1 = t_3 \\ \cdots \\ t_{[\frac{n}{2}]-1} + t_1 = t_{[\frac{n}{2}]} \\ t_{[\frac{n}{2}]} + t_{[\frac{n}{2}]} = w \\ w + y = x_1 \\ y + y = y \ (\text{if } n \text{ is even}) \\ y = 1 \ (\text{if } n \text{ is odd}) \\ x_2 + t_1 = u \end{cases}$$

with n variables. The system S_n has less than κ solutions in $\mathbb{N}^n$. By the equivalence (E), S_n is satisfiable over $\mathbb{N}$. If an n-tuple $(x_1, x_2, x_3, \ldots, x_s, \ldots, w, y, u)$ of non-negative integers solves S_n, then by the equivalence (E),

$$x_2 = g(x_1) = g(w + y) = g\left(2 \cdot \left[\frac{n}{2}\right] + y\right) = g(n)$$

Therefore, $u = x_2 + t_1 = g(n) + 1 > g(n)$. This shows that $g(n) < f_\kappa(n)$ for any $n \geq 6 + 2s$. □

Let us fix an integer $\kappa \geq 2$.

For a positive integer n, let $\theta(n)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ with more than $\kappa - 1$ solutions in non-negative integers $x_1, \ldots, x_n$, at least two such solutions belong to $[0, b]^n$.

For a positive integer n and for a non-negative integer m, let $\beta(n, m)$ denote the smallest non-negative integer b such that for each system $S \subseteq E_n$ which has a solution in integers $x_1, \ldots, x_n$ from the range of 0 to m and which has less than κ solutions in integers $x_1, \ldots, x_n$ from the range of 0 to m, there exists a solution that belongs to $[0, b]^n$. The function $\beta : (\mathbb{N} \setminus \{0\}) \times \mathbb{N} \to \mathbb{N}$ is computable.

The following equalities

$$f_\kappa(n) = \beta(n, \max(f_\kappa(n), \theta(n))) = \beta(n, \max(f_\kappa(n), \theta(n)) + 1) =$$

$$\beta(n, \max(f_\kappa(n), \theta(n)) + 2) = \beta(n, \max(f_\kappa(n), \theta(n)) + 3) = \ldots$$

hold for any positive integer n. Therefore, there is an algorithm which takes as input a positive integer n, performs an infinite loop, returns $\beta(n, m-1)$ on the m-th iteration, and returns $f_\kappa(n)$ on each sufficiently high iteration. This proves that the function f_κ is computable in the limit for any integer $\kappa \geq 2$.

Theorem 5. *Let $\kappa = 2$. We claim that the following MuPAD code implements an algorithm which takes as input a positive integer n, performs an infinite loop, returns $\beta(n, m-1)$ on the m-th iteration, and returns $f(n)$ on each sufficiently high iteration.*

```
input("input the value of n",n):
X:=[0]:
while TRUE do
Y:=combinat::cartesianProduct(X $i=1..n):
W:=combinat::cartesianProduct(X $i=1..n):
for s from 1 to nops(Y) do
for t from 1 to nops(Y) do
m:=0:
for i from 1 to n do
if Y[s][i]=1 and Y[t][i]<>1 then m:=1 end_if:
for j from i to n do
for k from 1 to n do
if Y[s][i]+Y[s][j]=Y[s][k] and Y[t][i]+Y[t][j]<>Y[t][k]
then m:=1 end_if:
if Y[s][i]*Y[s][j]=Y[s][k] and Y[t][i]*Y[t][j]<>Y[t][k]
```

```
then m:=1 end_if:
end_for:
end_for:
end_for:
if m=0 and s<>t then
W:=listlib::setDifference(W,[Y[s]]) end_if:
end_for:
end_for:
print(max(max(W[z][u] $u=1..n) $z=1..nops(W))):
X:=append(X,nops(X)):
end_while:
```

Proof. Let us say that a tuple $y = (y_1, \ldots, y_n) \in \mathbb{N}^n$ is a *duplicate* of a tuple $x = (x_1, \ldots, x_n) \in \mathbb{N}^n$, if

$$(\forall i \in \{1, \ldots, n\}\ (x_i = 1 \Longrightarrow y_i = 1)) \wedge$$

$$(\forall i, j, k \in \{1, \ldots, n\}\ (x_i + x_j = x_k \Longrightarrow y_i + y_j = y_k)) \wedge$$

$$(\forall i, j, k \in \{1, \ldots, n\}\ (x_i \cdot x_j = x_k \Longrightarrow y_i \cdot y_j = y_k))$$

For a positive integer n and for a non-negative integer m, $\beta(n, m)$ equals the smallest non-negative integer b such that the box $[0, b]^n$ contains all tuples $(x_1, \ldots, x_n) \in \{0, \ldots, m\}^n$ which have no duplicates in $\{0, \ldots, m\}^n \setminus \{(x_1, \ldots, x_n)\}$. □

The proof of Theorem 5 effectively shows that the function f is computable in the limit. Limit-computable functions, also known as trial-and-error computable functions, have been thoroughly studied, see [9, pp. 233–235] for the main results. The function f_{ω_1} is also computable in the limit [12] and the following *MuPAD* code

```
input("input the value of n",n):
X:=[0]:
while TRUE do
Y:=combinat::cartesianProduct(X $i=1..n):
W:=combinat::cartesianProduct(X $i=1..n):
for s from 1 to nops(Y) do
for t from 1 to nops(Y) do
m:=0:
for i from 1 to n do
if Y[s][i]=1 and Y[t][i]<>1 then m:=1 end_if:
for j from i to n do
for k from 1 to n do
if Y[s][i]+Y[s][j]=Y[s][k] and Y[t][i]+Y[t][j]<>Y[t][k]
then m:=1 end_if:
if Y[s][i]*Y[s][j]=Y[s][k] and Y[t][i]*Y[t][j]<>Y[t][k]
then m:=1 end_if:
end_for:
end_for:
end_for:
if m=0 and max(Y[t][i] $i=1..n)<max(Y[s][i] $i=1..n)
```

```
then W:=listlib::setDifference(W,[Y[s]]) end_if:
end_for:
end_for:
print(max(max(W[z][u] $u=1..n) $z=1..nops(W))):
X:=append(X,nops(X)):
end_while:
```

performs an infinite computation of $f_{\omega_1}(n)$. The flowchart in Figure 1 describes an algorithm which computes $f_\kappa(n)$ in the limit for any $\kappa \in \{\omega_1\} \cup \{2, 3, 4, \ldots\}$.

MuPAD is a computer algebra system whose syntax is modelled on *Pascal*. The commercial version of *MuPAD* is no longer available as a stand-alone product, but only as the *Symbolic Math Toolbox* of *MATLAB*. Fortunately, all presented codes can be executed by *MuPAD Light*, which was offered for free for research and education until autumn 2005.

Theorem 6 ([12]). *Let $\kappa \in \{2, 3, 4, \ldots, \omega\}$. Let us consider the following three statements:*

(*a*) *There exists an algorithm $\mathscr{A}$ whose execution always terminates and which takes as input a Diophantine equation D and returns the answer* `YES` *or* `NO` *which indicates whether or not the equation D has a solution in non-negative integers, if the solution set Sol(D) satisfies card(Sol(D)) $< \kappa$.*
(*b*) *The function f_κ is majorized by a computable function.*
(*c*) *If a set $\mathscr{M} \subseteq \mathbb{N}^n$ has a κ-fold Diophantine representation, then $\mathscr{M}$ is computable.*

We claim that (*a*) *is equivalent to* (*b*) *and* (*a*) *implies* (*c*).

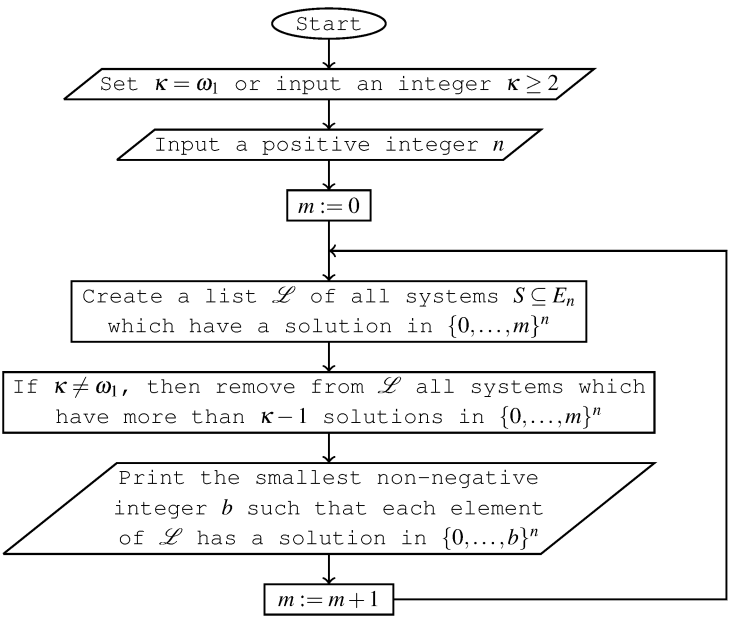


Fig. 1 An infinite computation of $f_\kappa(n)$, where $\kappa \in \{\omega_1\} \cup \{2, 3, 4, \ldots\}$

Proof. The implication $(a) \Rightarrow (c)$ is obvious. We prove the implication $(a) \Rightarrow (b)$. There is an algorithm Dioph which takes as input a positive integer m and a non-empty system $S \subseteq E_m$, and returns a Diophantine equation Dioph(m, S) which has the same solutions in non-negative integers $x_1, \ldots, x_m$. Item (a) implies that for each Diophantine equation D, if the algorithm $\mathscr{A}$ returns YES for D, then D has a solution in non-negative integers. Hence, if the algorithm $\mathscr{A}$ returns YES for Dioph(m, S), then we can compute the smallest non-negative integer $i(m, S)$ such that Dioph(m, S) has a solution in non-negative integers not greater than $i(m, S)$. If the algorithm $\mathscr{A}$ returns NO for Dioph(m, S), then we set $i(m, S) = 0$. The function

$$\mathbb{N} \setminus \{0\} \ni m \to \max\left\{i(m, S)\colon \emptyset \neq S \subseteq E_m\right\} \in \mathbb{N}$$

is computable and majorizes the function $f_{\boldsymbol{K}}$. We prove the implication $(b) \Rightarrow (a)$. Let a function h majorizes $f_{\boldsymbol{K}}$. By Lemma 2 for $\boldsymbol{K} = \mathbb{N}$, a Diophantine equation D is equivalent to a system $S \subseteq E_n$. The algorithm $\mathscr{A}$ checks whether or not S has a solution in non-negative integers $x_1, \ldots, x_n$ not greater than $h(n)$. □

The implication $(a) \Rightarrow (c)$ remains true with a weak formulation of item (a), where the execution of $\mathscr{A}$ may not terminate or $\mathscr{A}$ may return nothing or something irrelevant, if D has at least κ solutions in non-negative integers. The weakened item (a) implies that the flowchart in Figure 2 describes an algorithm whose execution terminates, if the set

$$Sol(D) := \{(x_1, \ldots, x_n) \in \mathbb{N}^n\colon\ D(x_1, \ldots, x_n) = 0\}$$

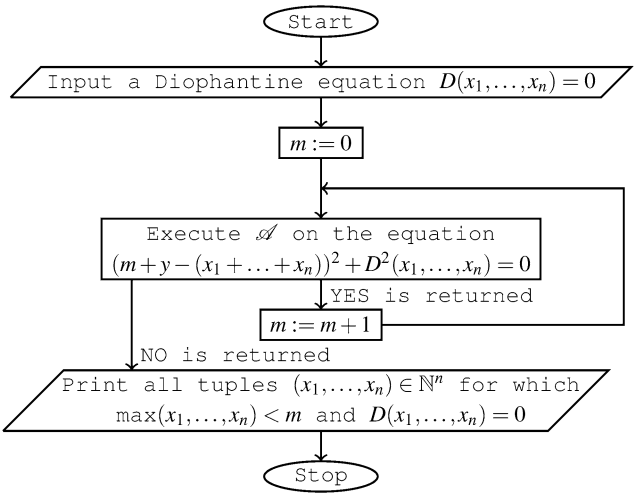


Fig. 2 An algorithm that conditionally finds all solutions to a Diophantine equation which has less than κ solutions in non-negative integers

has less than κ elements. If this condition holds, then the weakened item (a) guarantees that the execution of the flowchart in Figure 2 prints all elements of $Sol(D)$. However, the weakened item (a) is equivalent to the original one. Indeed, if the algorithm $\mathscr{A}$ satisfies the weakened item (a), then the flowchart in Figure 3 illustrates a new algorithm $\mathscr{A}$ that satisfies the original item (a).

Y. Matiyasevich in [6] studies Diophantine equations and Diophantine representations over $\mathbb{N} \setminus \{0\}$.

Theorem 7 ([6, p. 87]). *Suppose that there exists an effectively enumerable set having no finite-fold Diophantine representation. We claim that if a one-parameter Diophantine equation*

$$J(u, x_1, \ldots, x_m) = 0 \tag{3}$$

for each value of the parameter u has only finitely many solutions in $x_1, \ldots, x_m$, then there exists a number n such that in every solution of (3)

$$x_1 < u^n, \ldots, x_m < u^n$$

Theorem 7 is false for $u = 1$ when $J(u, x_1) = u + x_1 - 3$. Theorem 7 is missing in [7], the Springer edition of [6]. The author has no opinion on the validity of Theorem 7 for integers $u > 1$, but is not convinced by the proof in [6]. Theorem 7 restricted to integers $u > 1$ and reformulated for solutions in non-negative integers implies the following Corollary:

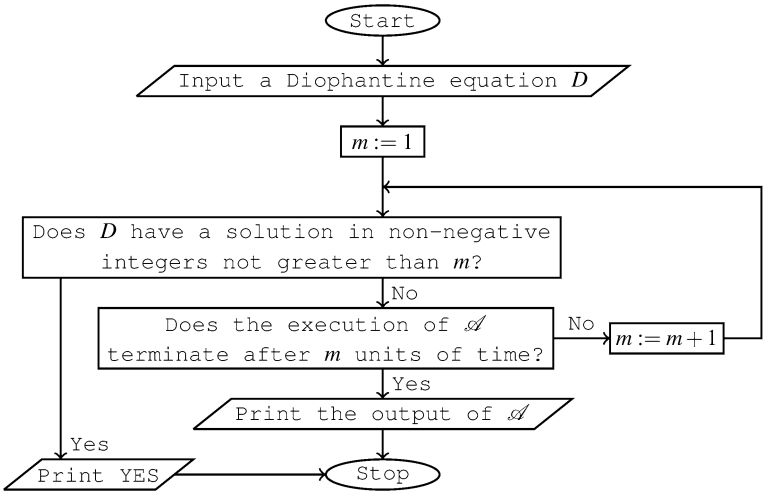


Fig. 3 The weakened item (a) implies the original one

Corollary. *If there exists a recursively enumerable set having no finite-fold Diophantine representation, then any set $\mathscr{M} \subseteq \mathbb{N}$ with a finite-fold Diophantine representation is computable.*

Let us pose the following two questions:

Question 1. Is there an algorithm $\mathscr{B}$ which takes as input a Diophantine equation D, returns an integer, and this integer is greater than the heights of non-negative integer solutions, if the solution set has less than κ elements? We allow a possibility that the execution of $\mathscr{B}$ does not terminate or $\mathscr{B}$ returns nothing or something irrelevant, if D has at least κ solutions in non-negative integers.

Question 2. Is there an algorithm $\mathscr{C}$ which takes as input a Diophantine equation D, returns an integer, and this integer is greater than the number of non-negative integer solutions, if the solution set is finite? We allow a possibility that the execution of $\mathscr{C}$ does not terminate or $\mathscr{C}$ returns nothing or something irrelevant, if D has infinitely many solutions in non-negative integers.

Obviously, a positive answer to Question 1 implies the weakened item (a). Conversely, the weakened item (a) implies that the flowchart in Figure 4 describes an appropriate algorithm $\mathscr{B}$.

Theorem 8 ([12]). *A positive answer to Question 1 for $\kappa = \omega$ is equivalent to a positive answer to Question 2.*

Proof. Trivially, a positive answer to Question 1 for $\kappa = \omega$ implies a positive answer to Question 2. Conversely, if a Diophantine equation $D(x_1, \ldots, x_n) = 0$ has only finitely many solutions in non-negative integers, then the number of non-negative integer solutions to the equation

$$D^2(x_1, \ldots, x_n) + (x_1 + \ldots + x_n - y - z)^2 = 0$$

is finite and greater than $\max(a_1, \ldots, a_n)$, where $(a_1, \ldots, a_n) \in \mathbb{N}^n$ is any solution to $D(x_1, \ldots, x_n) = 0$. □

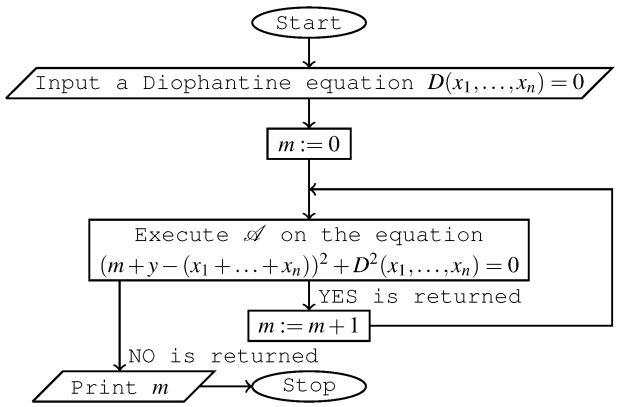


Fig. 4 The weakened item (a) implies a positive answer to Question 1

References

1. Appendix: A counterexample to the conjecture. http://www.cyf-kr.edu.pl/~rttyszka/IPL.pdf (2014). A part of the report by an anonymous referee of Inform. Process. Lett.
2. Davis, M., Matiyasevich, Y., Robinson, J.: Hilbert's tenth problem: diophantine equations: positive aspects of a negative solution. In: Mathematical Developments Arising from Hilbert problems, Proceedings of Symposium on Pure Mathematics, vol. 28, pp. 323–378. American Mathematical Society, Providence, RI (1976). Reprinted in: The collected works of Julia Robinson (ed. S. Feferman), American Mathematical Society, 1996, pp. 269–324
3. Lagarias, J.C.: On the computational complexity of determining the solvability or unsolvability of the equation $X^2 - DY^2 = -1$. Trans. Am. Math. Soc. **260**(2), 485–508 (1980)
4. Matiyasevich, Y.: Hilbert's tenth problem. MIT Press, Cambridge, MA (1993)
5. Matiyasevich, Y.: Hilbert's tenth problem: what was done and what is to be done. In: Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Contemporary Mathematics, vol. 270, pp. 1–47. American Mathematical Society, Providence, RI (2000)
6. Matiyasevich, Y.: Towards finite-fold Diophantine representations. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Math. Inst. Steklov. **377**, 78–90 (2010)
7. Matiyasevich, Y.: Towards finite-fold Diophantine representations. J. Math. Sci. (N. Y.) **171**(6), 745–752 (2010)
8. Ribenboim, P.: The Little Book of Bigger Primes. Springer, New York (2004)
9. Soare, R.I.: Interactive computing and relativized computability. In: Copeland, B.J., Posy, C.J., Shagrir, O. (eds.) Computability: Turing, Gödel, Church and Beyond, pp. 203–260. MIT Press, Cambridge, MA (2013)
10. Tyszka, A.: Conjecturally computable functions which unconditionally do not have any finite-fold Diophantine representation. Inf. Process. Lett. **113**(19–21), 719–722 (2013)
11. Tyszka, A.: Does there exist an algorithm which to each Diophantine equation assigns an integer which is greater than the modulus of integer solutions, if these solutions form a finite set? Fund. Inf. **125**(1), 95–99 (2013)
12. Tyszka, A.: Mupad codes which implement limit-computable functions that cannot be bounded by any computable function. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems, vol. 2, pp. 623–629. IEEE Computer Society Press (2014)

Image Encryption Scheme Based on Non-autonomous Chaotic Systems

Christos K. Volos, Ioannis M. Kyprianidis, Ioannis Stouboulos, and Viet-Thanh Pham

Abstract In this chapter, the great sensitivity of nonlinear systems, and especially of chaotic systems, on the initial conditions and on the variation of their parameters, was used to design a novel image encryption scheme. Until now, a great number of chaotic autonomous continuous systems or discrete dynamical systems, have been used in various image encryption processes, as a source of random numbers. However, in this work, a Chaotic Random Bit Generator (CRBG), which is based on a non-autonomous dynamical system, is used. For ridding from the system the influence of the external source and increasing the security of the proposed generator, the Poincaré section for sampling the signal has been used. As a dynamical system, the very well-known Duffing–van der Pol system has been chosen, presenting very good statistical results. The aforementioned CRBG is the "heart" of the proposed image encryption scheme. Finally, the security analysis of the proposed encryption scheme, based on histogram analysis, correlation of two adjacent pixels, differential analysis and information entropy, demonstrate the robustness of the proposed chaotic encryption scheme against all kinds of statistical, cryptanalytic, and brute-force attacks.

Keywords: Image encryption • Chaotic random bit generator • Non-autonomous dynamical system • Duffing-van der Pol system • FIPS-140-2 • Security analysis

1 Introduction

In the last decades, information sharing and especially images information sharing became more and more prevalent under the rapid development of Internet, networks and mobile communication technologies. However, in open networks, there is a

C.K. Volos (✉) • I.M. Kyprianidis • I. Stouboulos
Department of Physics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
e-mail: volos@physics.auth.gr; imkypr@auth.gr; stouboulos@physics.auth.gr

V.-T. Pham
School of Electronics and Telecommunications, Hanoi University of Science and Technology, 01 Dai Co Viet, Hanoi, Vietnam
e-mail: pvt3010@gmail.com

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_25

potential risk of making sensitive information, such as online personal photographs, industrial drawings and medical images, vulnerable to unauthorized interceptions. Also, images for military use such as drawings of military establishments, photographs which are produced by satellites or from military missions, must be also kept private from enemies attacks. So, the development of robust cryptographic schemes is essential to the provision of image's security.

Furthermore, as it is known, digital images have some very characteristic features such as:

- Bulk data capacity,
- Strong pixel correlation,
- High redundancy, and
- Existence of patterns and backgrounds.

In more detail, image data are usually bulky and very large-sized. So, the encryption of such bulky data with the traditional ciphers incurs significant overhead, and it is too expensive for real-time multimedia applications. Also, in the case of digital images, adjacent pixels often have similar gray-scale values and strong correlations, or image blocks have similar patterns. Such an extremely high data redundancy of multimedia makes the conventional encryption schemes fail to obscure all visible information. Furthermore, in many real-life multimedia applications, it is very important that very light encryption should be made to preserve some perceptual information. This is impossible to be achieved with traditional encryption schemes alone, which most likely degrade the data to a perceptually unrecognizable content.

Therefore, because of these features, traditional ciphers like Data Encryption Scheme (DES) [1], International Data Encryption Algorithm (IDEA) [2], and Advanced Encryption Scheme (AES) [3] are not suitable for real time image encryption, as these ciphers require a large computational time and high computing power.

So, nowadays, since traditional encryption schemes are not fit for modern image requirement, many research teams have been devoted to investigate better solutions on image encryption processes, such as digital watermarking [4–7] and chaotic encryption [8–11].

Furthermore, in the last decades, nonlinear systems and especially chaotic systems have aroused tremendous interest because of their applications in several disciplines including meteorology, physics, engineering, economics, biology, and philosophy [12]. Chaos theory studies the behavior of dynamical systems that are highly sensitive on initial conditions, an effect which is popularly referred to as the "Butterfly Effect." This means that small differences in initial conditions, such as those due to rounding errors in numerical computation, yield widely diverging outcomes for such dynamical systems, rendering long-term prediction impossible in general. This happens even though these systems are deterministic, meaning that their future behavior is fully determined by their initial conditions, with no random elements involved. In other words, the deterministic nature of these systems does not make them predictable.

So, the main advantage of the encryption schemes, which are based on chaos, lies on the observation that a chaotic signal looks like noise for an unauthorized user who ignores its mechanism of generation. Secondly, the time evolution of the chaotic signal strongly depends on system's initial conditions and parameters. So, slight variations in these quantities yield quite different time evolutions. This means that system's initial conditions and parameters can be efficiently used as keys in an encryption system based on chaos. Also, the generation of a chaotic signal is often of low cost, which makes it suitable for the encryption of large bulky data. According to the classification of chaotic systems, the chaotic encryption schemes, which have been proposed, can be divided into two categories: analog chaotic cryptosystems utilizing continuous dynamical systems [13, 14] and digital chaotic cryptosystems utilizing discrete dynamical systems [15–17].

Also, it is known that cryptography and chaos have a structural relationship due to their many similar properties [18]. As a result of this close relationship several chaotic cryptosystems have been presented. One of the most interesting ways through which chaotic cryptosystems can be realized is via the implementation of a Chaotic Random Bit Generator (CRBG). Until now, the great majority of such generators is based on autonomous nonlinear dynamical systems, in order to use the independence of these systems from external sources. However, in the present work a CRBG, which is based on a non-autonomous dynamical system, is used. For ridding from the system the influence of the external source and increasing the security of the proposed generator, the Poincar*é* section for sampling the time series has been used.

So, in response to the aforementioned challenges, the objective of this chapter is the presentation of a gray-scale image encryption scheme realized with a non-autonomous chaotic system, the Duffing–van der Pol, which is used in the CRBG. The produced chaotic bitstream is a result of the X-OR function in the outputs of two threshold circuits that use two same variables (x) by the two Duffing–van der Pol's Poincar*é* maps. Next, this bit sequence is subjected to the de-skewing technique to extract unbiased bits with no correlation and so to increase their complexity, as it is confirmed by the statistical test suite, FIPS-140-2. The produced bits sequence is used to encrypt and decrypt digital images. Statistical analysis by using histogram analysis, correlation of two adjacent pixels, differential analysis, and information entropy confirmed the robustness of the encryption process against various known statistical attacks.

This chapter is organized as follows. In Sect. 2, the definition of chaotic systems in general and the description of the Duffing–van der Pol system, which has been used, are given. Section 3 introduces the CRBG that is the base of the proposed image encryption scheme and the results of the use of the statistical tests suite FIPS-140-2 in the proposed generator. Section 4 demonstrates step by step the proposed encryption process of a gray-scale image by using the CRBG. In Sect. 5, the necessary security analysis of the proposed chaotic image encryption schemes is presented. Finally, conclusion remarks are drawn in the last section.

2 The Duffing–Van Der Pol System

Chaos refers to some dynamical phenomena considered to be complex and unpredictable. Although it was precluded by Poincaré at the end of the nineteenth century [19], chaos theory begins to take form in the second half of the twentieth century after observations of the evolution of different physical systems [20, 21]. These systems revealed that despite the knowledge of their evolution rules and initial conditions, their future seemed to be arbitrary and unpredictable. That opened quite a revolution in modern physics, terminating with Laplace's ideas of causal determinism [22].

Until now, chaos has been observed in weather and climate [20], population growth in ecology [23], economy [24], to mention only a few examples. It also has been observed in the laboratory in a number of systems such as electrical circuits [25], lasers [26], chemical reactions [27], fluid dynamics [28], mechanical systems, and magneto-mechanical devices [29]. So, chaos theory provides the means to explain various phenomena in nature and make use of chaotic dynamical systems in many different scientific fields.

From a mathematical viewpoint a nonlinear dynamical system, in order to be considered as chaotic, must fulfill the following three conditions [30].

- It must be topologically mixing,
- Its chaotic orbits must be dense, and
- It must be very sensitive on initial conditions.

Firstly, the term topologically mixing means that the chaotic dynamical system, especially the chaotic designated area of the trajectory will eventually cover part of any particular region. The second feature of chaotic systems is that its chaotic orbits have to be dense. This means that the trajectory of a dynamical system is dense, if it comes arbitrarily close to any point in the domain. Finally, the most important feature of chaotic systems, as it is mentioned, is the sensitivity on initial conditions. This means that a small variation on a system's initial conditions will produce a totally different chaotic trajectory.

In this chapter, the second order nonlinear, non-autonomous Duffing–van der Pol system [31], which is described by the following set of differential equations (1), (2), is used.

$$\frac{dx}{dt} = y \tag{1}$$

$$\frac{dy}{dt} = \mu(1-x^2)y - x^3 + Bcos(\omega_N z) \tag{2}$$

The aforementioned system is called Duffing–van der Pol, because it contains in the second equation the term, $\mu(1-x^2)y$, which is a characteristic feature of the van der Pol oscillator and the cubic term x^3 of Duffing's equation.

The dynamic behavior of the Duffing–van der Pol system is investigated numerically by employing the fourth order Runge–Kutta algorithm. The system's

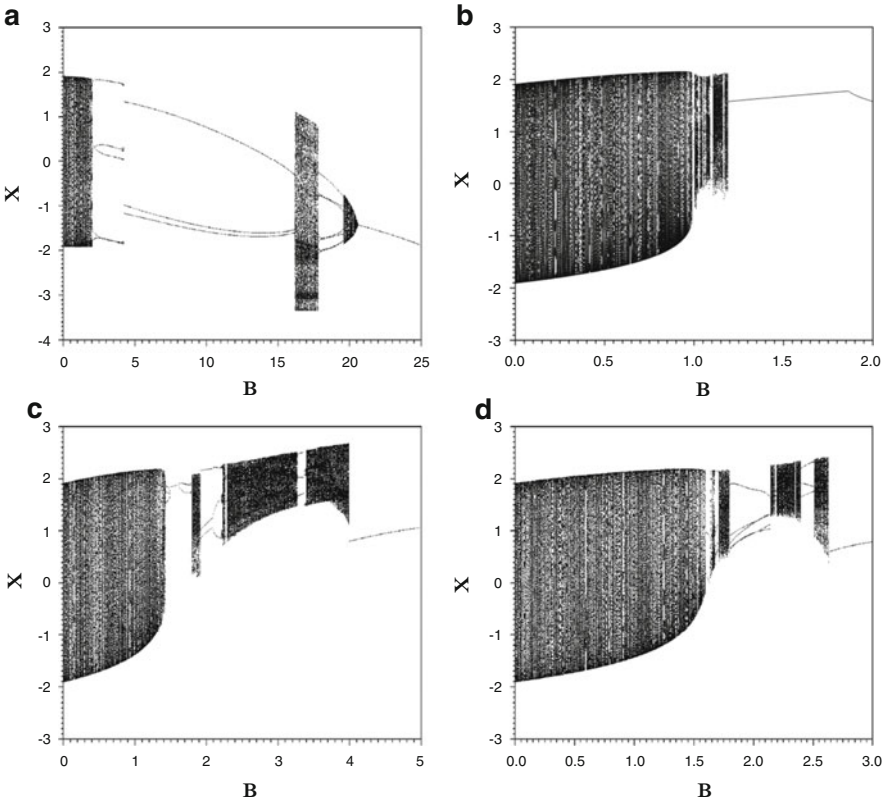


Fig. 1 The bifurcation diagram of x vs. B, for $\mu = 0.2$, while (**a**) $\omega_N = 4.0$, (**b**) $\omega_N = 0.9$, (**c**) $\omega_N = 0.7$ and (**d**) $\omega_N = 0.6$

rich dynamical behavior is revealed in Fig. 1, which shows the bifurcation diagrams of x versus the parameter B, for various values of ω_N, while $\mu = 0.2$. Periodic and chaotic regions alternate as the parameter B increases while interesting dynamical phenomena, such as routes to chaos and crisis phenomena (boundary, internal), are also displayed. This richness of dynamic behavior makes the proposed non-autonomous system a suitable candidate for use in this CRBG.

The base of the proposed CRBG is the well-known Poincaré map named by Henri Poincaré. It is the intersection of an orbit in the state space of a continuous dynamical system with a certain lower dimensional subspace, called the Poincaré cross-section, transversal to the flow of the system. So, in this case, the Poincaré map is produced by using the Poincaré cross-section which is defined by

$$\sum = \{(x, y, \theta = \omega_N \tau_N) \in R^2 \times S^1\} \tag{3}$$

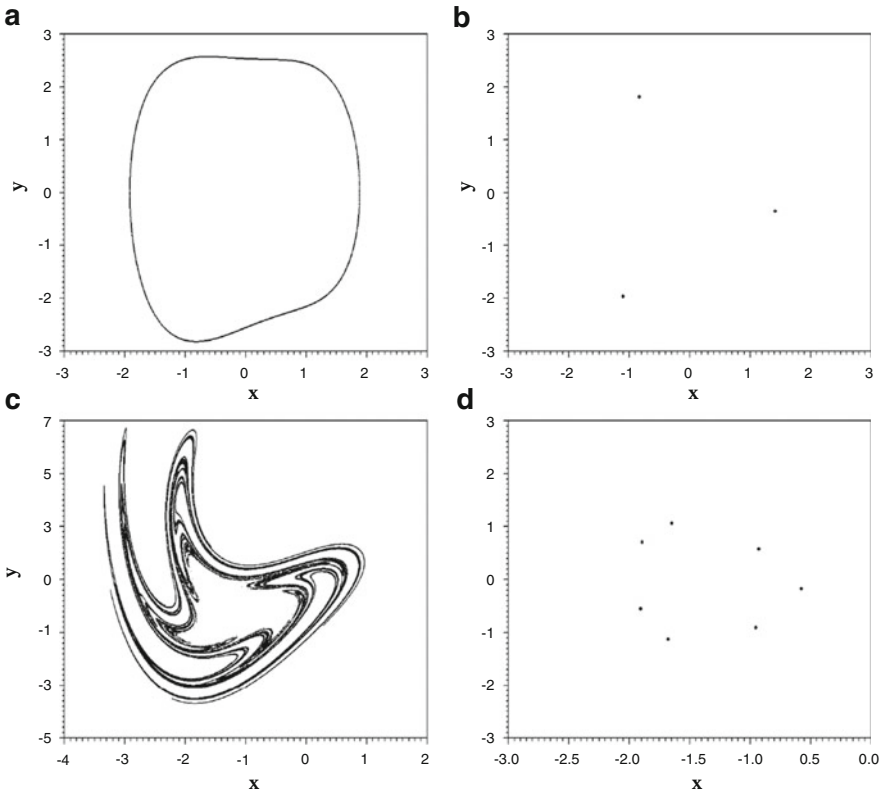


Fig. 2 Poincaré maps of y vs. x, for $\mu = 0.2$ and $\omega_N = 4$, while (**a**) $B = 1$, (**b**) $B = 8$, (**c**) $B = 17$ and (**d**) $B = 19$

where $\tau_N = NT + \tau_0$ is the sampling time, τ_0 the initial time determining the location of the Poincaré cross-section on which the coordinates (x, y) of the attractors are projected, and $T = 2\pi/\omega_N$ is the period of the voltage source.

In Fig. 2 a number of Poincaré maps for various values of the parameter B of the bifurcation diagram of Fig. 1a, in the case of $\mu = 0.2$ and $\omega_N = 4.0$, are displayed. From this figure the great utility of the Poincaré map can be seen, since for different system's dynamic behavior the Poincaré map has a totally different form. As it is known, the various forms of the Poincaré map depending on system's dynamic behavior are:

- Discrete number of points for Periodic behavior.
- Closed curve for Quasi-periodic behavior.
- Strange attractor for Chaotic behavior.

3 The Chaotic Random Bit Generator

As it was mentioned before, one of the methods to obtain aperiodic sequences is to use chaos which is defined as "random" phenomenon generated by simple deterministic systems. Until now there have been many works on random number generation based on chaos by using either continuous or discrete chaotic systems [32–52].

In the case of discrete-time chaotic systems, truly random generators which were realized by analog circuits generate aperiodic random sequences. However, it is difficult to generate random sequences with good statistical properties due to nonidealities of analog circuit elements and inevitable noise [35, 53]. Thus, there have also been several works on post-processing of the chaos-based random numbers [54, 55].

The proposed random bit generator is based on a non-autonomous continuous chaotic system. So, for this reason, the values of system's parameters were selected so that the system is in chaotic state. In Fig. 3 the phase portrait and the respective Poincaré map for a chosen set of system's parameters are shown.

In Fig. 4, the proposed CRBG is presented. This generator consists of five blocks. The first block (S1) includes two chaotic non-autonomous systems (Duffing–van der Pol) running side by side with different set of parameters (μ_1, B_1, ω_{N1}) and (μ_2, B_2, ω_{N2}), respectively. In the second block (S2) the Poincaré map for the two systems is used. The state variables x_{i1} from the two systems are partitioned into two subspaces each other by using the two threshold functions in the third block (S3), which are described as:

- $\sigma^{x_1} = 0$, if $x_{i1} < x_{T1}$ or $\sigma^{x_1} = 1$, if $x_{i1} \geq x_{T1}$
- $\sigma^{x_2} = 0$, if $x_{i2} < x_{T2}$ or $\sigma^{x_2} = 1$, if $x_{i2} \geq x_{T2}$

where x_{T1} and x_{T2} are the threshold values for the variables x_{i1} and x_{i2}, respectively.

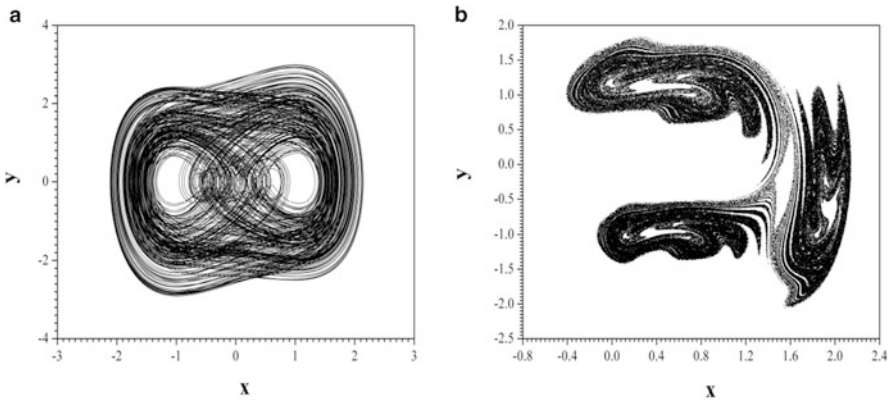


Fig. 3 (**a**) The phase portrait and the (**b**) the Poincaré map of y vs. x, for $\mu = 0.2$, $B = 1.175$ and $\omega_N = 0.92$

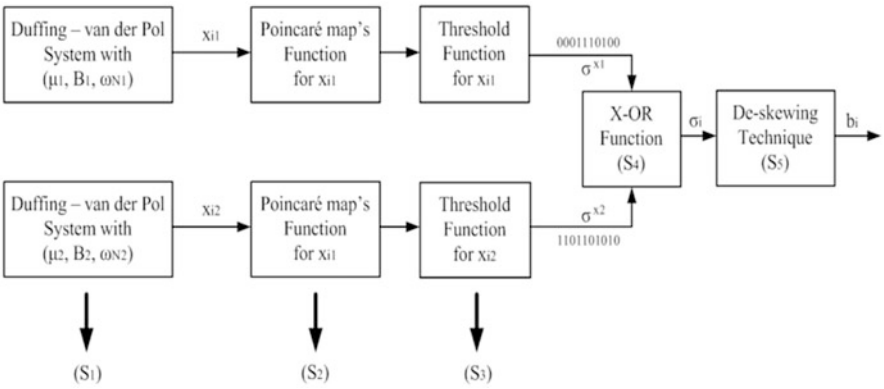


Fig. 4 The block diagram of the proposed CRBG

The fourth block (S4) produces the bit sequence σ_i by using the X-OR function in σ^{x_1} and σ^{x_2}, which are finally subjected into the well-known De-skewing technique for eliminating the correlation in the output of the sources of random bits. Von Neumann has been the first author to state this problem [56]. He proposed a digital post-processing that balances the distribution of bits. This technique consists of converting the bit pair "01" into the output "0", "10" into the output "1" and of discarding bit pairs "00" and "11".

This CRBG, due to its simple structure, can be implemented in hardware or by using a microcontroller. However, in this chapter, for testing reasons, the proposed CRBG is implemented in a software environment.

3.1 Statistical Tests

From the beginnings of computing, random number generation has been a subject of great interest. Especially, from a practical point of view, "randomness" occurs to the extent that something cannot be predicted. In the literature there are several informal definitions of "randomness," usually based on either a lack of discernible patterns in a sequence or the unpredictability of the sequence or various aspects of it by, generally, the most puissant possible adversary. Poincaré pointed out that the classic random outcome of a die throwing or a flipping coin comes from the sensitive dependence on the initial condition. A small perturbation causes a large difference in the final outcome, thereby making prediction difficult. This sensitive dependence, as it was pointed out, is a hallmark of Chaos.

Furthermore, in the generation process of random numbers nobody can be sure, if the produced numbers are really random. That is exactly why a background theory is needed for it, and a set of standardized statistical tests must certify the numbers as random. So, in order to gain the confidence that a newly developed random

bit generator is cryptographically secure, due to its high level of randomness, it should be subjected to a variety of statistical tests designed to detect the specific characteristics expected of truly random sequences. There are several options available in the literature for analyzing the "randomness" of the newly developed random bit generators. The four most popular options are:

- The FIPS-140-2 (Federal Information Processing Standards) suite of statistical tests of the National Institute of Standards and Technology (NIST) [57],
- The DIEHARD suite of statistical tests, which was created by the statistician George Marsaglia [58],
- The Crypt-XS suite of statistical tests, which was developed by researchers at the Information Security Research Centre at Queensland University of Technology in Australia [59], and
- The Donald Knuth's statistical tests set, which includes several empirical statistical tests [60].

In this chapter the "randomness" of the produced bit sequences, by the proposed CRBG, is analyzed by using the FIPS-140-2 suite of statistical tests. The results of the use of the four more important statistical tests (Monobit test, Poker test, Runs test, and Long run test), which are part of the FIPS-140-2, are presented in detail. According to FIPS-140-2, the examined CRBG will produce a bitstream, $b_i = b_0$, $b_1, b_2, \ldots, b_{n-1}$, of length n (at least 20,000 bits), which must satisfy the following standards [57].

- Monobit Test: The number n_1 of 1's in the bitstream must be $9,725 < n_1 < 10,275$.
- Poker Test: This test determines whether the sequences of length n ($n = 4$) show approximately the same number of times in the bitstream. The bounds of this statistic are then $2.16 < X3 < 46.17$.
- Runs Test: This test determines whether the number of 0's (Gap) and 1's (Block) of various lengths in the bitstream are as expected for a random sequence (Table 1).
- Long Run Test: This test is passed if there are no runs longer than 26 bits.

Table 1 Required intervals for length of runs test, according to FIPS-140-2 statistical tests

Length of run	Required interval
1	2,315–2,685
2	1,114–1,386
3	527–723
4	240–384
5	103–209
6	103–209

From the information theory, it is known that the noise has maximum entropy. For this reason, the system's parameters and initial conditions are chosen so as the measure-theoretic entropy [61] of the CRBG, which is given by the following equation, is maximum.

$$H_n = \lim_{n\to\infty}\left(-\sum_{B^n} P\left(B^n\right)\ln P\left(B^n\right)/n\right) \tag{4}$$

where $P(B^n)$ is the probability of occurrence of a binary subsequence B of length n.

So, in this chapter, two Duffing–van der Pol systems with identical parameters $B = 2.0, \omega_N = 0.6$ and by using slightly different values for μ ($\mu_1 = 0.20$, while $\mu_2 = 0.19$), initial conditions $(x_{01}, y_{01}) = (0.5, 1.0)$, $(x_{02}, y_{02}) = (0.49, 0.99)$ and threshold values $(x_{n1}, x_{n2}) = (1.238, 1.313)$ are running side by side as it was mentioned. For the chosen sets of parameters the two systems present the expected chaotic behavior as it confirmed by the phase portraits and the respective Poincaré maps of Fig. 5.

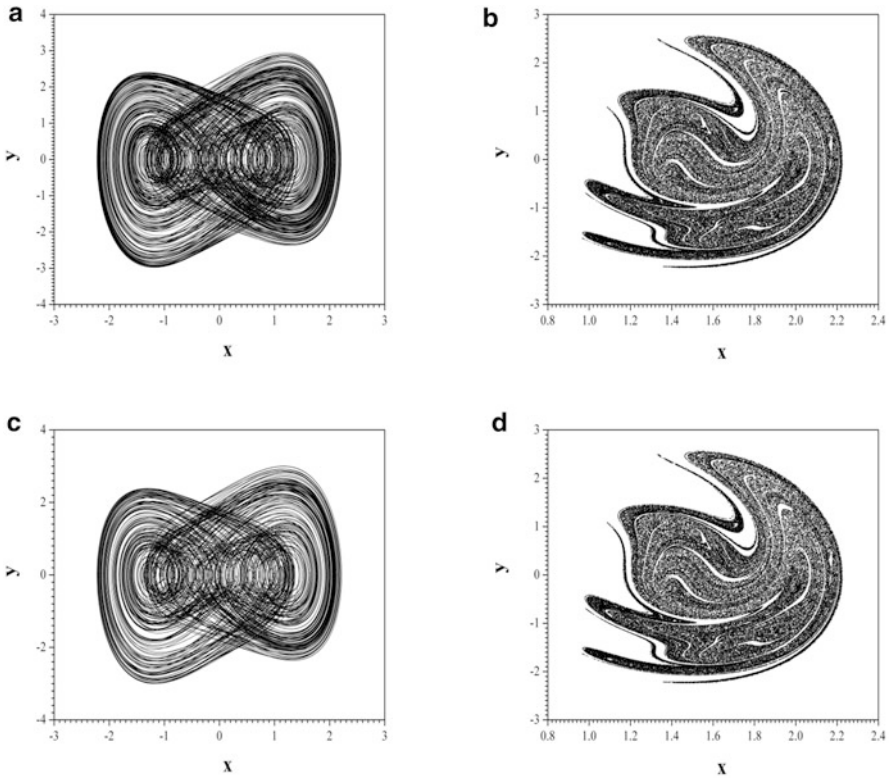


Fig. 5 (**a**) The phase portraits and the Poincaré maps of y vs. x, for $B = 2$, $\omega_I = 0.6$, while (**a**), (**b**) $\mu_1 = 0.20$, (**c**), (**d**) $\mu_2 = 0.19$

With the procedure described in the previous section, by using the aforementioned sets of systems' parameters, a bitstream of 200,000 bits is obtained which is divided in 10 bit sequences of length 20,000 bits. In Table 2 the measure-theoretic entropy for $n = 3$ and $n = 4$ and the detailed results of the 10 bit sequences which were subjected to the four tests of FIPS-140-2 test suite are presented. As a conclusion, all the bit sequences produced by the CRBG have numerically verified the specific characteristics expected of random bit sequences. Finally, Table 3 presents the analytical results for the four tests of FIPS-140-2, in the case of the first of the 10 bit sequences.

4 The Image Encryption Scheme

In this chapter a simple but effective encryption scheme for gray-scale images, which has been implemented in MATLAB, is presented. This encryption process is mainly based on X-OR function as in many other related works [62, 63]. The proposed encryption scheme includes the following steps.

- **Step 1**: The scheme finds the pixel size $M \times N$ of the image, where M and N represent the numbers of rows and columns of the image. The pixels are arranged by order from left to right and top to bottom. Then an image data set, in which each element is the decimal gray-scale value of the pixel (0-255), is produced. Finally each decimal value is converted into a binary equivalent number and in the end a one-dimensional matrix B is produced.
- **Step 2**: The matrix A which is a binary sequence produced by the chaotic TRBG, with the procedure that was described in Sect. 3, and the above-mentioned matrix B produces a third one-dimensional matrix C by using the X-OR function: $C = A \oplus B$.
- **Step 3**: The produced in the previous step matrix C is converted into the encrypted image by the inverse process of step 1.

For the image's decryption process the X-OR function must be applied again ($C \oplus B = A$). In Fig. 6 the plain gray-scale image (plane) of size 131×131 pixels, the encrypted and the decrypted images which are produced with the above encryption scheme are shown.

5 Statistical and Security Analysis

In this section the robustness of the proposed encryption scheme against many existing statistical, cryptanalytic, and brute-force attacks is demonstrated. Thus, the results of the security analysis on the proposed image encryption scheme, by using histogram analysis, correlation of two adjacent pixels, differential analysis, key space analysis and information entropy analysis, are presented.

Table 2 Results of the measure-theoretic entropy and FIPS-140-2 tests suite, for 10 bit sequences produced by the proposed CRBG

Tests	1	2	3	4	5	6	7	8	9	10
H_3	0.692430	0.692375	0.692005	0.692410	0.692343	0.692502	0.692252	0.692579	0.692281	0.692417
H_4	0.692111	0.691208	0.691521	0.691468	0.691421	0.690959	0.690333	0.691490	0.691419	0.690893
Monobit Test	ν	ν	ν	ν	ν	ν	ν	ν	ν	ν
Poker test	ν	ν	ν	ν	ν	ν	ν	ν	ν	ν
Runs test	ν	ν	ν	ν	ν	ν	ν	ν	ν	ν
Long run test	ν	ν	ν	ν	ν	ν	ν	ν	ν	ν
Results	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed

Table 3 Analytical results of FIPS-140-2 tests, for the first bit sequence of Table 2

Monobit test	Poker test	Runs test	Long run test
$n_1 = 10{,}040$ (50.22 %)	28.51452	$B_1 = 2{,}580$ $B_2 = 1{,}215$ $B_3 = 504$ $B_4 = 295$ $B_5 = 151$ $B_6 = 190$	No
Passed	Passed	Passed	Passed

5.1 Statistical Analysis

This section is devoted to analyze the statistical behavior of the gray-scale encrypted image produced by the proposed scheme.

5.1.1 Histogram Analysis

Histogram analysis is an important metric used in the evaluation of the robustness of an image encryption scheme. An image histogram shows the distribution of the pixel values within an image. Figure 7 presents the histogram of the plain (plane) and the encrypted image, respectively. From this figure one can observe that the pixel values of the plain image are not uniformly distributed over the interval [0, 255]. Whereas, the histogram of the encrypted image shows uniform distribution of the pixel values. Based on these results, we conclude that the encrypted image do not provide any useful information about their corresponding plain image. So, the proposed image encryption scheme is secure from any statistical attack.

5.1.2 Correlation Analysis

Another important metric, which is used in the evaluation of an image encryption scheme, is the correlation analysis. Each pixel of any image has a high correlation with its adjacent pixels either in horizontal, vertical, or diagonal directions. For testing the correlation in a plain and encrypted image, respectively, the correlation coefficient γ [64] of each pair of pixels was calculated, by using the following formulas.

$$E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$D(x) = \frac{1}{N}\sum_{i=1}^{N} [x_i - E(x)]^2 \tag{6}$$

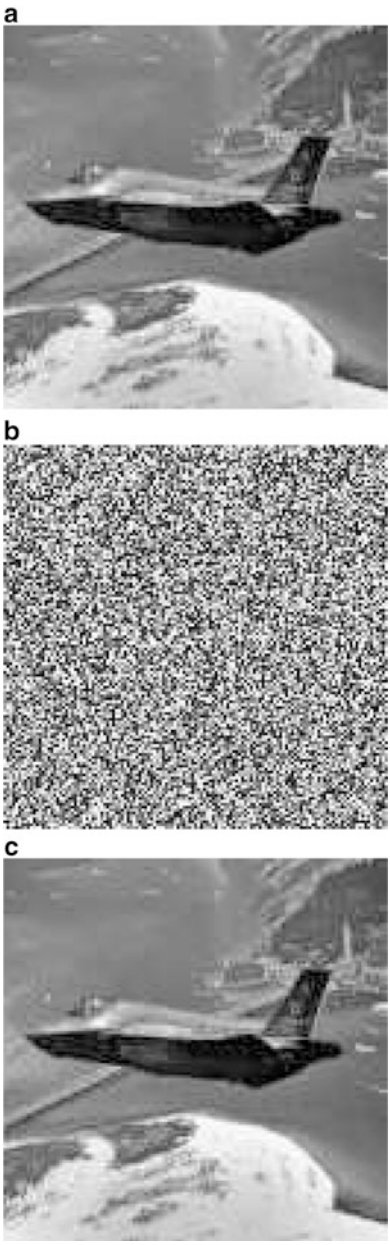


Fig. 6 (**a**) The plain image (plane), (**b**) the encrypted image, and (**c**) the decrypted image

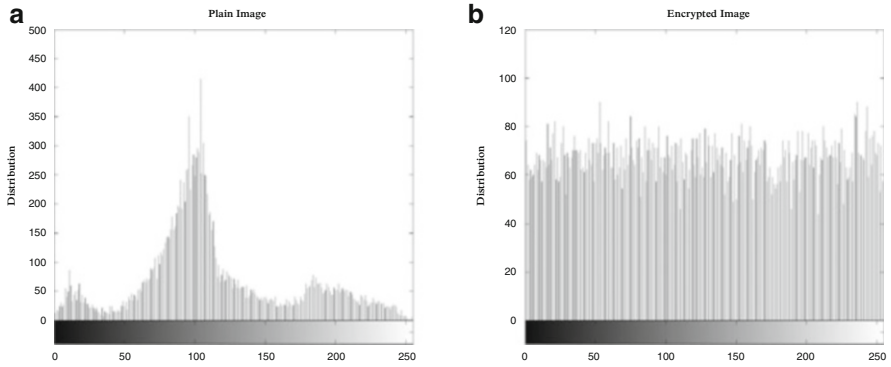


Fig. 7 (**a**) Histogram of the plain and (**b**) the encrypted image

$$\mathrm{cov}(x, y) = \frac{1}{N} \sum_{i=1}^{N} [x_i - E(x)] [y_i - E(y)] \tag{7}$$

$$\gamma(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{D(x)\sqrt{D(y)}}} \tag{8}$$

In the aforementioned equations, x and y are the gray values of two adjacent pixels in the image and N is the total number of adjacent pairs of pixels. In Fig. 8 the correlations of two horizontal, vertical and diagonal pixels in the plain and the encrypted image are shown respectively. Also, the results of the correlation coefficient of the encrypted image, which has been decreased significantly, in regard to the correlation coefficient of the plain image, are presented in Table 4. It is obvious that the correlation coefficient of the encrypted image in any direction is approximately equal to zero, so the correlated relationship is very low. Thus, the proposed encryption scheme is robust against this type of statistical attacks.

5.1.3 Entropy Analysis

Entropy is one of the most important randomness measures [65]. The entropy $H(s)$ of a source s is given by

$$H(s) = -\sum_{i=0}^{N-1} p(s_i) log_2 p(s_i) \tag{9}$$

where $p(s_i)$ is the probability of appearance of the symbol s_i.

The entropy of an image presents the distribution of the gray-scale values (0-255). As much uniform the distribution, is so much bigger the entropy is. Table 5

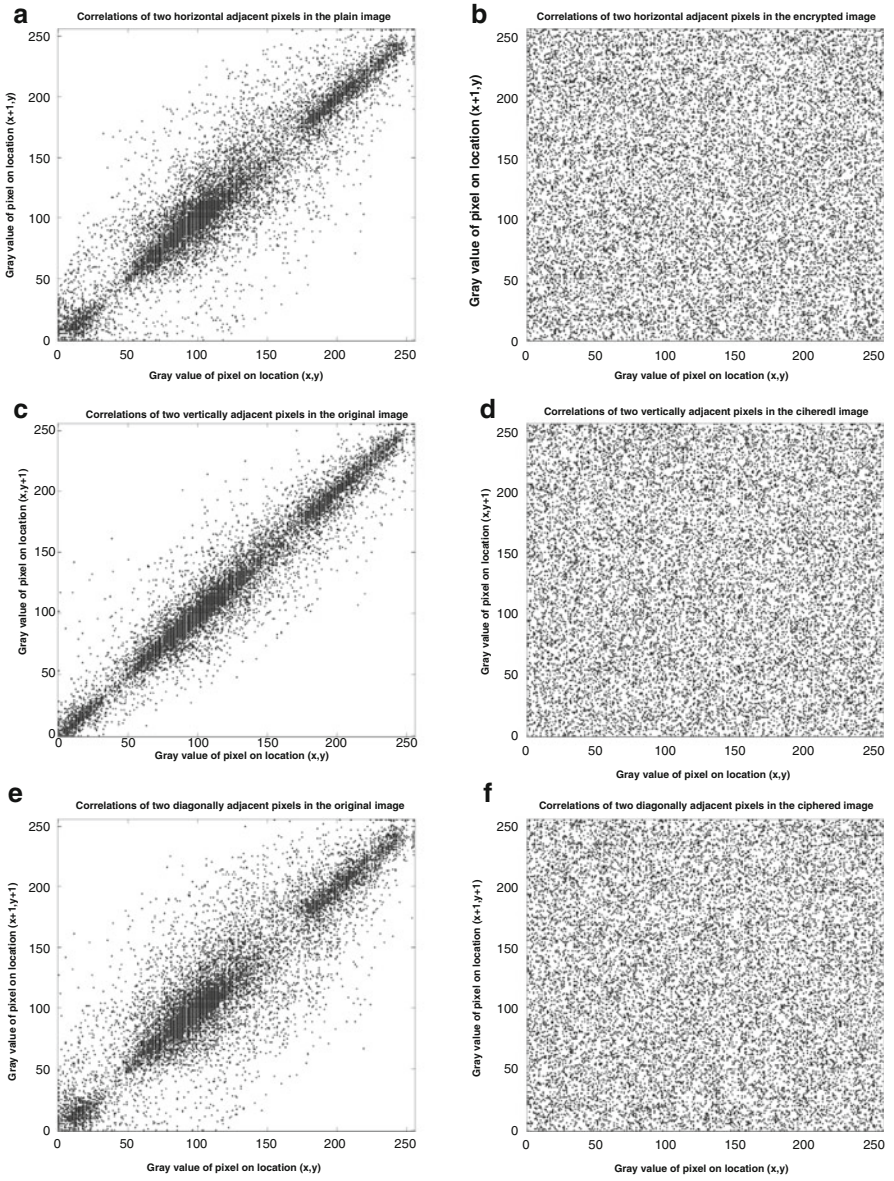


Fig. 8 Correlations of two (**a**), (**b**) horizontal, (**c**), (**d**) vertical and (**e**), (**f**) diagonal adjacent pixels in the plain and encrypted image

presents the entropy results for the plain and the encrypted image. Due to the fact that the entropy of the encrypted image is increased, we have come to the conclusion that the proposed encryption method is safe from an entropy attack.

Table 4 Correlation coefficients of two adjacent pixels in the plain and encrypted image

	Plain image	Encrypted image
Horizontal	0.9292	0.0006
Vertical	0.9715	0.0014
Diagonal	0.9071	0.0054

Table 5 Entropy results for the plain and the encrypted image

	Plain image	Encrypted image
Entropy	7.4003	7.9555

5.2 Security Analysis

In this section, the key-space and the sensitivity of the proposed scheme to a tiny change in the plain image is analyzed.

5.2.1 Key Space and Sensitivity

A direct method for cipher-image analysis is to launch the brute-force attack if there is enough time. So, the key space should be as large as possible. In this work, the key space of the proposed encryption scheme is consisted of the following parts:

- The systems' parameters: B_1, B_2, μ_1, μ_2, ω_{N1}, ω_{N2}.
- The systems' initial conditions: (x_{01}, y_{01}) and (x_{02}, y_{02}).
- The threshold values: (x_{n1}, x_{n2}).

So, the key space can be as large as 10^{168}, if the computational precision of the 64-bit double-precision number is set to 10^{-14}. This result shows that the key space of the proposed encryption scheme is very large and the scheme can resist against brute force attacks.

Also, an ideal image encryption scheme should have high sensitivity to every key even if only a tiny change has been made. In order to show this characteristic, only the value of the initial condition x_{01}, of the first system, has been changed (with 10^{-14} change) and the failure of recovering the plain image is presented in Fig. 9. Therefore, the proposed image encryption scheme is very sensitive on the initial conditions, which have been used.

5.2.2 Differential Analysis

The differential attack is one of the most famous attacks against an encrypted image. This method is based on a tiny change (modification of one pixel) in the encrypted image and the result is observed. Using this technique someone can find a relationship between the encrypted and plain image. So, if a minor change in the plain image can cause a significant change in the encrypted image, then the differential attack would become practically useless.

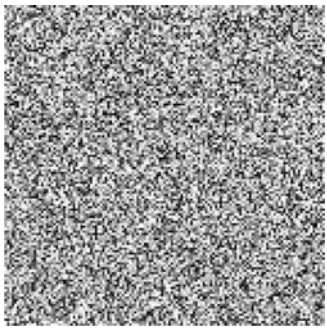


Fig. 9 The wrong recovered image

The robustness of the proposed encryption method against the differential attack is examined by changing one pixel in the plain image and two common numbers: the Number of Pixels Change Rate (NPCR) and the Unified Average Changing Intensity (UACI) [66], are calculated. Therefore, if $A(i,j)$ and $B(i,j)$ are the pixels in row—i and column—j of the encrypted images A and B, with only one pixel difference between the respective plain images, then the NPCR is calculated by the following formula:

$$\mathrm{NPCR}(A,B) = 100\% \left(\sum_{i,j} D(i,j) \right) / N \tag{10}$$

where N is the total number of pixels and $D(i,j)$ is produced by the following way: $D(i,j) = 1$ if $A(i,j) \neq B(i,j)$ or $D(i,j) = 0$ if $A(i,j) = B(i,j)$.

For two random selected images the NPCR is $\mathrm{NPCR} = (1 - 2^L) \times 100\,\%$, where L is the number of bits used for representing the pixels of an image. So, for a gray-scale image (8 bit/pixel), the NPCR is equal to 99.60938 %.

The second number (UACI) measures the average intensity of differences between the plain image and the encrypted image, calculated by the following formula:

$$\mathrm{UACI}(A,B) = \frac{1}{N} \left(\sum_{i,j} \frac{|A(i,j) - B(i,j)|}{2^L - 1} \right) 100\% \tag{11}$$

The expected value of UACI for two random selected images is:

$$\mathrm{UACI} = \left(\sum_{i-1}^{2^L - 1} i(i-1) \right) / 2^L (2^L - 1) \tag{12}$$

So, for a gray scale image the UACI is equal to 33.46354 %.

Table 6 The NPCR and UACI at two encryption rounds

Round	1 (%)	2 (%)
NPCR	0.0233	99.24
UACI	0.0043	33.02

Therefore, the values of these two numbers show that the encryption scheme is very weak to a differential attack. To improve this weakness of the proposed scheme the encryption process is evaluated in more than one round. The NPCR and UACI at different rounds of encryption process are calculated and listed in Table 6. In each round the bitstream is shifted only one bit. Table 6 shows that the performance is very satisfactory after only two rounds of encryption while the values of NPCR and UACI have the tendency to be equal to the calculated values of random selected images.

6 Conclusion

In this chapter a robust and efficient image encryption scheme, based on a CRBG, was studied. The main elements of this CRBG were two non-autonomous Duffing–van der Poll dynamical systems, which were running side by side, with different sets of parameters and initial conditions, for producing the random bit sequence.

Simulation results showed the excellent performance of the proposed scheme. Furthermore, these results demonstrated the robustness of the scheme against existing statistical-based attacks. Moreover, security analysis demonstrated the high sensitive dependence of the encryption scheme to a slightly modification in the secret key. In addition the key space is large enough to defeat brute force attacks. Therefore, the proposed encryption scheme is suitable for use in many applications including medical, military, or industrial image encryption.

References

1. Data Encryption Standard: NIST FIPS PUB 46-2. U.S. Department of Commerce (1993)
2. Lai, X., Massey, J.: A proposal for a new block encryption standard. In: Proceedings of Advances in Cryptology EUROCRYPT '90, pp. 389–404. Springer, New York (1991)
3. Advanced Encryption Standard: NIST FIPS PUB 197. U.S. Department of Commerce (2001)
4. Chrysochos, E., Fotopoulos, V., Xenos, M., Skodras, A.N.: Hybrid watermarking based on chaos and histogram modification, signal, image and video processing. Signal Image Video Process. (2012). doi:10.1007/s11760-012-0307-3
5. Fotopoulos, V., Stavrinou, M.L., Skodras, A.N.: Medical image authentication and self-correction through an adaptive reversible watermarking technique. In: Proceedings of 8th IEEE International Conference on Bioinformatics and Bioengineering, vol. 1-2, pp. 910–914 (2008)
6. Rawat, S., Raman, B.: A blind watermarking algorithm based on fractional Fourier transform and visual cryptography. Signal Process. **92**, 1480–1491 (2012)

7. Yeung, M.M., Pankanti, S.: Verification watermarks on fingerprint recognition and retrieval. J. Electron. Imaging **9**, 468–476 (2000)
8. Chen, T.-H., Wu, C.-S.: Efficient multi-secret image sharing based on boolean operations. Signal Process. **91**, 90–97 (2011)
9. Liao, X., Lai, S., Zhou, Q.: A novel image encryption algorithm based on self-adaptive wave transmission. Signal Process. **90**, 2714–2722 (2010)
10. Wang, X., Teng, L., Qin, X.: A Novel colour image encryption algorithm based on chaos. Signal Process. **92**, 1101–1108 (2012)
11. Zhang, L., Liao, X., Wang X.: An image encryption approach based on chaotic maps. Chaos Solitons Fractals **24**, 759–765 (2005)
12. Grebogi, C., Yorke, J.: The Impact of Chaos on Science and Society. United Nations University Press, Tokyo (1997)
13. Li, Z., Xu, D.: A secure communication scheme using projective chaos synchronization. Chaos Solitons Fractals **22**, 477–481 (2004)
14. Chen, J.Y., Wong, K.W., Cheng, L.M., Shuai, J.W.: A secure communication scheme based on the phase synchronization of chaotic systems. Chaos **13**, 508–514 (2003)
15. Baptista, M.S.: Cryptography with chaos. Phys. Lett. A **240**, 50–54 (1998)
16. Habutsu, T., Nishio, Y., Sasase, I., Mori, S.: A secret key cryptosystem by iterating a chaotic map. In: Proceedings of Advances in Cryptology-CRYPTO '91, pp. 127–140. Springer, New York (1991)
17. Yen, J.C., Guo, J.I.: A new key-based design for image encryption and decryption. In: Proceedings of IEEE Conference on Circuits and Systems, vol. 4, pp. 49–52 (2000)
18. Alvarez, G., Li, S.: Some basic cryptographic requirements for chaos based cryptosystems. Int. J. Bifurcation Chaos **16**, 2129–2151 (2006)
19. Poincare, J.H.: Sur le probleme des trois corps et les equations de la dynamique. Divergence des series de M. Lindstedt. Acta Math. **13**, 1–270 (1890)
20. Lorenz, E.N.: Deterministic non-periodic flow. J. Atmos. Sci. **20**, 130–141 (1963)
21. Mandelbrot, B.: The Fractal Geometry of Nature. W.H. Freeman Company, New York (1977)
22. Laplace, P.S.: Traité du Mécanique Céleste. Oeuvres Complétes de Laplace. Gauthier-Villars, Paris (1825)
23. May, R.M., McLean, A.R.: Theoretical Ecology: Principles and Applications. Blackwell, Oxford (2007)
24. Kyrtsou, C., Vorlow, C.: Complex dynamics in macroeconomics: a novel approach. In: Diebolt, C., Kyrtsou, C. (eds.) New Trends in Macroeconomics, pp. 223–245. Springer, Berlin (2005). ISBN-13: 978-3-540-21448-9
25. Van der Pol, B., Van der Mark, J.: Frequency demultiplication. Nature **120**, 363–364 (1927)
26. Casperson, L.W.: Gas laser instabilities and their interpretation. In: Proceedings of the NATO Advanced Study Institute, pp. 83–98. Springer, Berlin (1988)
27. Field, R.J., Gyӧrgyi, L.: Chaos in Chemistry and Biochemistry. World Scientific Publishing, Singapore (1993)
28. Baker, G.L.: Chaotic Dynamics: An Introduction. Cambridge University Press, Cambridge (1996)
29. Moon, F.C.: Chaotic vibrations: An Introduction for Applied Scientists and Engineers. Wiley, New York (1987)
30. Hasselblatt, B., Katok, A.: A First Course in Dynamics: With a Panorama of Recent Developments. Cambridge University Press, Cambridge (2003)
31. Ueda, Y., Akamatsu, N.: Chaotically transitional phenomena in the forced negative-resistance oscillator. IEEE Trans. Circuits Syst. **CAS-28**, 217–224 (1981)
32. Oishi, S., Inoue, H.: Pseudo-random number generators and chaos. Trans. Inst. Electr. Commun. Eng. Japan E **65**, 534–541 (1982)
33. Bernstein, G.M., Lieberman, M.A.: Secure random number generation using chaotic circuits. IEEE Trans. Circuits Syst. **37(9)**, 1157–1164 (1990)
34. Kohda, T., Tsuneda, A. Statistics of chaotic binary sequences. IEEE Trans. Inf. Theory **43(1)**, 104–112 (1997)

35. Tsuneda, A., Eguchi, K., Inoue. T.: Design of chaotic binary sequences with good statistical properties based on piecewise linear into maps. In: Proceedings of 7th International Conference on Microelectronics for Neural, Fuzzy and Bio-Inspired Systems, pp. 261–266 (1999)
36. Kolesov, V.V., Belyaev, R.V., Voronov, G.M.: Digital random-number generator based on the chaotic signal algorithm. J. Commun. Technol. Electron. **46**, 1258–1263 (2001)
37. Stojanovski, T., Kocarev, L.: Chaos-based random number generators - part I: analysis. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **48(3)**, 281–288 (2001)
38. Stojanovski, T., Pihl, J., Kocarev, L.: Chaos-based random number generators - part II: practical realizations. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **48(3)**, 382–385 (2001)
39. Bernardini, R., Cortelazzo, G.: Tools for designing chaotic systems for secure random number generation. IEEE Trans. Circuits Syst. **48(5)**, 552–564 (2001)
40. Gerosa, A., Bernardini, R., Pietri, S.: A fully integrated chaotic system for the generation of truly random numbers. IEEE Trans. Circuits Syst. I **49(7)**, 993–1000 (2001)
41. Li, S., Mou, X., Cai, Y.: Pseudo-random bit generator based on coupled chaotic systems and its application in stream-ciphers cryptography. In: Progress in Cryptology - INDOCRYPT 2001. Lecture Notes in Computer Science, vol. 2247, pp. 316–329 (2001)
42. Li, K., Soh, Y.C., Li, Z.G. Chaotic cryptosystem with high sensitivity to parameter mismatch. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **50**, 579–583 (2003)
43. Gentle, J.E.: Random Number Generation and Monte Carlo Method. Springer, New York (2003)
44. Kocarev, L.: Chaos-based cryptography: a brief overview. IEEE Circuits Syst. Mag. **1**, 6–21 (2001)
45. Fu, S.M., Chen, Z.Y., Zhou, Y.A.: Chaos based random number generators. Comput. Res. Dev. **41**, 749–754 (2004)
46. Huaping, L., Wang, S., Gang, H.: Pseudo-random number generator based on coupled map lattices. Int. J. Mod. Phys. B **18**, 2409–2414 (2004)
47. Yalcin, M.E., Suykens, J.A.K., Vandewalle, J.: True random bit generation from a double-scroll attractor. IEEE Trans. Circuits Syst. I **51(7)**, 1395–1404 (2004)
48. Wei, J., Liao, X., Wong, K., Xiang, T.: A new chaotic cryptosystem. Chaos Solitons Fractals **30**, 1143–1152 (2006)
49. Volos, C.K., Kyprianidis, I.M., Stouboulos, I.N.: Experimental demonstration of a chaotic cryptographic scheme. WSEAS Trans. Circuits Syst. **5**, 1654–1661 (2006)
50. Volos, C.K., Kyprianidis, I.M., Stouboulos, I.N.: Fingerprint images encryption process based on a chaotic true random bits generator. Int. J. Multimedia Intell. Secur. **1**, 320–335 (2010)
51. Volos, C.K., Kyprianidis, I.M., Stouboulos, I.N.: Image encryption process based on chaotic synchronization phenomena. Signal Process. **93**, 1328–1340 (2013)
52. Volos, C.K.: Chaotic random bit generator realized with a microcontroller. J. Comput. Model. **3(4)**, 115–136 (2013)
53. Kennedy, M.P., Rovatti, R., Setti, G.: Chaotic Electronics in Telecommunications. CRC Press, West Palm Beach, FL (2000)
54. Pareschi, F., Rovatti, R., Setti, G.: Simple and effective post-processing stage for random stream generated by a chaos-based RNG. In: Proceedings of 2006 International Symposium in Nonlinear Theory and its Applications, pp. 383–386 (2006)
55. Tang, K.W., Tang, W.: A low cost chaos-based random number generator realized in 8-bit precision environment. In: Proceedings of 2006 International Symposium in Nonlinear Theory and its Applications, pp. 395–398 (2006)
56. Von Neumann, J.: Various techniques used in connection with random digits. In: Forsythe, G.E. (eds.) Applied Mathematics Series, National Bureau of Standards, vol. 12, pp. 36–38 (1951)
57. NIST: Security Requirements for Cryptographic Modules. FIPS PUB 140-2. http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf (2001)
58. Marsaglia, G.: DIEHARD Statistical Tests. http://stst.fsu.edu/pub/diehard (1995)
59. Gustafson, H., Dawson, H.E., Nielsen, L., Caelli, W.: A computer package for measuring the strength of encryption algorithms. J. Comput. Secur. **13**, 687–697 (1994).

60. Knuth, D.: The Art of Computer Programming: Semiemperical Algorithms. Addison Wesley, Reading (1998)
61. Fraser, A.M.: Information and entropy in strange attractors. IEEE Trans. Inf. Theory **35**, 245–262 (1989)
62. Volos, C.K., Kyprianidis, I.M., Strouboulos, I.N.: Fingerprint images encryption process based on a chaotic true random bits generator. Int. J. Multimedia Intell. Secur. **1**, 320–335 (2010)
63. Volos, C.K., Kyprianidis, I.M., Strouboulos, I.N.: Image encryption scheme based on coupled chaotic systems. J. Appl. Math. Bioinforma. **3**, 123–149 (2013)
64. Chen, G.R., Mao, Y., Chui, C.: A symmetric image encryption scheme based on 3D chaotic cat map. Chaos Solitons Fractals **21**, 749–761 (2004)
65. Li, W.: On the relationship between complexity and entropy for Markov chains and regular languages. Complex Syst. **5**, 381–399 (1991)
66. Chen, G.R., Mao, Y., Chui, C.: A symmetric image encryption scheme based on chaotic maps with finite precision representation. Chaos Solitons Fractals **32**, 1518–1529 (2007)

Multiple Parameterize Yang-Hilbert-Type Integral Inequalities

Bicheng Yang

Abstract In this chapter, by using the way of weight functions and technique of Real Analysis, a multiple Yang-Hilbert-type integral inequality with a general non-homogeneous kernel and multi-parameters is given. The equivalent forms, the operator expressions with the norm, the reverses, a few cases with the particular parameters and some examples with the particular kernels are also considered.

Keywords: Multiple Yang-Hilbert-type integral inequality • Kernel • Weight function • Norm • Operator

1 Introduction

Suppose that $p > 1, \frac{1}{p} + \frac{1}{q} = 1, f(\geq 0) \in L^p(\mathbf{R}_+), g(\geq 0) \in L^q(\mathbf{R}_+), ||f||_p, ||g||_q > 0$. We have the following equivalent Hardy-Hilbert's integral inequalities (cf. [1]):

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y}dxdy < \frac{\pi}{\sin(\pi/p)}||f||_p||g||_q, \tag{1}$$

$$\left[\int_0^\infty \left(\int_0^\infty \frac{f(x)}{x+y}dx\right)^p dy\right]^{\frac{1}{p}} < \frac{\pi}{\sin(\pi/p)}||f||_p, \tag{2}$$

where the constant factor $\frac{\pi}{\sin(\pi/p)}$ is the best possible.

Define Hardy-Hilbert's integral operator $T : L^p(\mathbf{R}_+) \to L^p(\mathbf{R}_+)$ as follows: for $f \in L^p(\mathbf{R}_+)$, there exists a unified expression $h \in L^p(\mathbf{R}_+)$, such that $Tf = h$, satisfying for any $y \in \mathbf{R}_+ = (0, \infty)$,

B. Yang (✉)
Department of Mathematics, Guangdong University of Education, Guangzhou, Guangdong 510303, People's Republic of China
e-mail: bcyang@gdei.edu.cn; bcyang818@163.com

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_26

Then in view of (2), it follows

$$||Tf||_p < \frac{\pi}{\sin(\pi/p)}||f||_p$$

and $||T|| \leq \frac{\pi}{\sin(\pi/p)}$. Since the constant factor is the best possible, we have

$$||T|| = \frac{\pi}{\sin(\pi/p)}.$$

Inequality (1) and (2) with the operator expression are important in analysis and its applications (cf. [2, 3]). In 2002, [4] considered the property of Hardy-Hilbert's integral operator and gave an improvement of (1)(for $p = q = 2$). In 2004, by adding another pair of conjugate exponents $(r, s)(r > 1, \frac{1}{r} + \frac{1}{s} = 1)$ and an independent parameter $\lambda > 0$, Yang [5] gave a best extension of (1) as follows:

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x^\lambda + y^\lambda}dxdy < \frac{\pi}{\lambda \sin(\pi/r)}||f||_{p,\phi}||g||_{q,\psi}, \tag{3}$$

where $\phi(x) = x^{p(1-\frac{\lambda}{r})-1}, \psi(x) = x^{q(1-\frac{\lambda}{s})-1}$,

$$||f||_{p,\phi} = \left\{\int_0^\infty \phi(x)f^p(x)dx\right\}^{\frac{1}{p}} > 0,$$

$||g||_{q,\psi} > 0$. In 2007, [6] gave the following inequality with the best constant $B(\frac{\lambda}{2}, \frac{\lambda}{2})(\lambda > 0; B(u, v)$ is the beta function):

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(1+xy)^\lambda}dxdy$$
$$< B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right)\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty x^{1-\lambda}g^2(x)dx\right)^{\frac{1}{2}}. \tag{4}$$

Definition 1. If $n \in \mathbf{N}, \mathbf{R}_+^n := \{(x_1, \ldots, x_n)|x_i \in \mathbf{R}_+ \ (i = 1, \ldots, n)\}$, $\lambda \in \mathbf{R} = (-\infty, \infty)$, $k_\lambda(x_1, \ldots, x_n)$ is a measurable function in $\mathbf{R}_+^n$, such that for any $u > 0$ and $(x_1, \ldots, x_n) \in \mathbf{R}_+^n$,

$$k_\lambda(ux_1, \ldots, ux_n) = u^{-\lambda}k_\lambda(x_1, \ldots, x_n),$$

then we call $k_\lambda(x_1, \ldots, x_n)$ the homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^n$.

In 2009, [7] gave an extension of (4) in $\mathbf{R}^2$ with the kernel $\frac{1}{|1+xy|^\lambda}(0 < \lambda < 1)$. Yang [8] gave another extension of (4) to the general kernel $k_\lambda(1, xy)(\lambda > 0)$ with one pair of conjugate exponents (p, q), and obtained the following multiple Hilbert-type integral inequality: Suppose that $n \in \mathbf{N}\backslash\{1\} = \{2, 3, \ldots\}, p_i > 1, \sum_{i=1}^n$

$\frac{1}{p_i} = 1, \lambda > 0, k_\lambda(x_1, \ldots, x_n) \geq 0$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^n$, such that for any $(r_1, \ldots, r_n)(r_i > 1)$, satisfying $\sum_{i=1}^n \frac{1}{r_i} = 1$, and

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \ldots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1} du_1 \cdots du_{n-1} \in \mathbf{R}_+.$$

If $\phi_i(x) = x^{p_i(1-\frac{\lambda}{r_i})-1}, f_i(\geq 0) \in L_{\phi_i}^{p_i}(\mathbf{R}_+)$, $||f||_{p_i,\phi_i} > 0$ $(i = 1, \ldots, n)$, then we have the following inequality:

$$\int_{\mathbf{R}_+^n} k_\lambda(x_1, \ldots, x_n) \prod_{i=1}^n f_i(x_i) dx_1 \cdots dx_n < k_\lambda \prod_{i=1}^n ||f_i||_{p_i,\phi_i}, \tag{5}$$

where the constant factor k_λ is the best possible.

For $n = 2, k_\lambda(x, y) = \frac{1}{x^\lambda + y^\lambda}$ in (5), we obtain (3); for $\lambda = n - 1, r_i = \frac{(n-1)p_i}{p_i - 1}$ $(i = 1, \ldots, n)$, (5) reduces to the following multiple Hardy-Hilbert-type inequality (cf. [1]), which relates one group of conjugate exponents $(p_1, \ldots, p_n)$:

$$\int_{\mathbf{R}_+^n} k_{n-1}(x_1, \ldots, x_n) \prod_{i=1}^n f_i(x_i) dx_1 \cdots dx_n < k_1 \prod_{i=1}^n ||f_i||_{p_i}. \tag{6}$$

Benyi and Oh [9] also studied the corresponding multiple Hardy-Hilbert-type integral operator with the homogeneous kernel of degree $-n + 1$.

We call (5) together with the equivalent forms as multiple Yang-Hilbert-type integral inequalities with the homogeneous kernel, which relates two groups of conjugate exponents and a few independent parameters. And the corresponding operator are called multiple Yang-Hilbert-type operator.

Inequality (5) are some extensions of the results in [10–14]. In recent years, [15, 16] considered some Hilbert-type operators relating (1)–(3); some other kinds of Hilbert-type inequalities are provided by [17–22].

In this chapter, by using the way of weight functions and technique of Real Analysis, a multiple Yang-Hilbert-type integral inequality with a general non-homogeneous kernel and multi-parameters is given, which is an extension of (5). The equivalent forms, the operator expressions with the norm, the reverses, a few cases with the particular parameters, and some examples with the particular kernels are also considered.

2 Some Lemmas

Lemma 1. *If* $n \in \mathbf{N}\backslash\{1\}, \delta_i \in \{-1, 1\}, \lambda_i \in \mathbf{R}$ $(i = 1, \ldots, n), \sum_{i=1}^n \frac{1}{p_i} = 1$, *then we have*

$$A := \prod_{i=1}^{n} \left[x_i^{(\delta_i\lambda_i-1)(1-p_i)} \prod_{j=1(j\neq i)}^{n} x_j^{\delta_j\lambda_j-1} \right]^{\frac{1}{p_i}} = 1. \tag{7}$$

Proof. We find

$$\begin{aligned} A &= \prod_{i=1}^{n} \left[x_i^{(\delta_i\lambda_i-1)(1-p_i)+1-\delta_i\lambda_i} \prod_{j=1}^{n} x_j^{\delta_j\lambda_j-1} \right]^{\frac{1}{p_i}} \\ &= \prod_{i=1}^{n} \left[x_i^{(1-\delta_i\lambda_i)p_i} \prod_{j=1}^{n} x_j^{\delta_j\lambda_j-1} \right]^{\frac{1}{p_i}} \\ &= \prod_{i=1}^{n} x_i^{1-\delta_i\lambda_i} \left(\prod_{j=1}^{n} x_j^{\delta_j\lambda_j-1} \right)^{\sum_{i=1}^{n}\frac{1}{p_i}}, \end{aligned}$$

and then in view of $\sum_{i=1}^{n} \frac{1}{p_i} = 1$, (7) is valid.

The lemma is proved.

Lemma 2. *Suppose that* $n \in \mathbf{N}\backslash\{1\}, \lambda_i \in \mathbf{R}, \delta_i \in \{-1, 1\}$ $(i = 1, \ldots, n)$, $\lambda_n = \sum_{i=1}^{n-1} \lambda_i = \frac{\lambda}{2}$, $k_\lambda(x_1, \ldots, x_n) \geq 0$ *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}_+^n$. *If*

$$\begin{aligned} H(i) := &\int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \ldots, u_{i-1}, 1, u_{i+1}, \ldots, u_n) \\ &\times \prod_{j=1(j\neq i)}^{n} u_j^{\lambda_j-1} du_1 \cdots du_{i-1} du_{i+1} \cdots du_n (i = 1, \ldots, n), \end{aligned}$$

satisfying

$$k_\lambda := H(n) = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \ldots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j-1} du_1 \cdots du_{n-1} \in \mathbf{R},$$

then each $H(i) = H(n) = k_\lambda$ *and for any* $i = 1, \ldots, n$,

$$\begin{aligned} \omega_i(x_i) := &x_i^{\delta_i\lambda_i} \int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1^{\delta_1} x_n^{\delta_n}, \ldots, x_{n-1}^{\delta_{n-1}} x_n^{\delta_n}, 1) \\ &\times \prod_{j=1(j\neq i)}^{n} x_j^{\delta_j\lambda_j-1} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n = k_\lambda. \end{aligned} \tag{8}$$

Proof. Setting $u_j = u_n v_j (j \neq i, n)$ in the integral $H(i)$, we find

$$H(i) = \int_{\mathbf{R}_+^{n-1}} k_\lambda(v_1, \ldots, v_{i-1}, u_n^{-1}, v_{i+1}, \ldots, v_{n-1}, 1) \prod_{j=1(j\neq i)}^{n-1} v_j^{\lambda_j - 1} \times u_n^{-1-\lambda_i} dv_1 \cdots dv_{i-1} dv_{i+1} \cdots dv_{n-1} du_n.$$

Setting $v_i = u_n^{-1}$ in the above integral, we obtain $H(i) = H(n)$.

Since $\lambda - \lambda_n = \lambda_n$, we find

$$\omega_i(x_i) = x_i^{\delta_i \lambda_i} \int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1^{\delta_1}, \ldots, x_{n-1}^{\delta_{n-1}}, x_n^{-\delta_n}) x_n^{-\delta_n \lambda_n - 1} \times \prod_{j=1(j\neq i)}^{n-1} x_j^{\delta_j \lambda_j - 1} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

Setting $y_n = x_n^{-1}$ in the above expression, we obtain

$$\begin{aligned}\omega_i(x_i) &= x_i^{\delta_i \lambda_i} \int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1^{\delta_1}, \ldots, x_{n-1}^{\delta_{n-1}}, y_n^{\delta_n}) y_n^{\delta_n \lambda_n + 1} \\ &\quad \times \prod_{j=1(j\neq i)}^{n-1} x_j^{\delta_j \lambda_j - 1} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_{n-1} (y_n^{-2}) dy_n \\ &= x_i^{\delta_i \lambda_i} \int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1^{\delta_1}, \ldots, x_{n-1}^{\delta_{n-1}}, y_n^{\delta_n}) y_n^{\delta_n \lambda_n - 1} \\ &\quad \times \prod_{j=1(j\neq i)}^{n-1} x_j^{\delta_j \lambda_j - 1} dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_{n-1} dy_n.\end{aligned}$$

Setting $u_j = x_j^{\delta_j} x_i^{-\delta_i} (j \neq i, n)$ and $u_n = y_n^{\delta_n} x_i^{-\delta_i}$ in the above integral, we find

$$\begin{aligned}\omega_i(x_i) &= x_i^{\delta_i \lambda_i} \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1 x_i^{\delta_i}, \ldots, u_{i-1} x_i^{\delta_i}, x_i^{\delta_i}, u_{i+1} x_i^{\delta_i}, \ldots, u_n x_i^{\delta_i}) \\ &\quad \times \prod_{j=1(j\neq i)}^{n} (u_j^{\delta_j^{-1}} x_i^{\delta_i \delta_j^{-1}})^{\delta_j \lambda_j - 1} |\delta_j^{-1}| x_i^{\delta_i \delta_j^{-1}} u_j^{\delta_j^{-1} - 1} du_1 \cdots du_{i-1} du_{i+1} \cdots du_n \\ &= H(i) = H(n) = k_\lambda.\end{aligned}$$

The lemma is proved.

Lemma 3. *As the assumption of Lemma 2, it follows that*

$$k(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{n-1}) := \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1} u_j^{\tilde{\lambda}_j-1}du_1\cdots du_{n-1}$$

is finite in a neighborhood of $(\lambda_1,\ldots,\lambda_{n-1})$ *if any only if* $k(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{n-1})$ *is continuous at* $(\lambda_1,\ldots,\lambda_{n-1})$.

Proof. The sufficiency property is obvious. We prove the necessary property of the condition by mathematical induction in the following.

For $n = 2$, there exists $I := \{\tilde{\lambda}_1 | \tilde{\lambda}_1 = \lambda_1 + \delta, |\delta| \le \delta_0, \delta_0 > 0\}$, such that for any $\tilde{\lambda}_1 \in I$, $k(\tilde{\lambda}_1) \in \mathbf{R}$. Since for $\tilde{\lambda}_1 = \lambda_1 + \delta \in I(\delta \ne 0)$,

$$\begin{aligned} k(\lambda_1+\delta) &= \int_0^\infty k_\lambda(u_1,1)u_1^{\lambda_1+\delta-1}du_1 \\ &= \int_0^1 k_\lambda(u_1,1)u_1^{\lambda_1+\delta-1}du_1 + \int_1^\infty k_\lambda(u_1,1)u_1^{\lambda_1+\delta-1}du_1, \\ &\quad k_\lambda(u_1,1)u_1^{\lambda_1+\delta-1} \le k_\lambda(u_1,1)u_1^{\lambda_1-\delta_0-1}du_1, u_1 \in (0,1]; \\ &\quad k_\lambda(u_1,1)u_1^{\lambda_1+\delta-1} \le k_\lambda(u_1,1)u_1^{\lambda_1+\delta_0-1}du_1, u_1 \in (1,\infty), \end{aligned}$$

and $k(\lambda_1 - \delta_0) + k(\lambda_1 + \delta_0) < \infty$, then by Lebesgue control convergence theorem (cf. [23]) , it follows

$$k(\lambda_1+\delta) = k(\lambda_1) + o(1)(\delta \to 0).$$

Assuming that for $n(\ge 2), k(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{n-1})$ is continuous at $(\lambda_1,\ldots,\lambda_{n-1})$, then for $n+1$, by the result of $n = 2$, since $k(\lambda_1+\delta_1,\ldots,\lambda_n+\delta_n)$ is finite in a neighborhood of $(\lambda_1,\ldots,\lambda_n)$, we find

$$\begin{aligned} &\lim_{\delta_n\to 0} k(\lambda_1+\delta_1,\ldots,\lambda_n+\delta_n) \\ &= \lim_{\delta_n\to 0}\int_0^\infty \left(\int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1,\ldots,u_n,1)\prod_{j=1}^{n-1} u_j^{\lambda_j+\delta_j-1}du_1\cdots du_{n-1}\right) u_n^{\lambda_n+\delta_n-1}du_n \\ &= \int_0^\infty \left(\int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1,\ldots,u_n,1)\prod_{j=1}^{n-1} u_j^{\lambda_j+\delta_j-1}du_1\cdots du_{n-1}\right) u_n^{\lambda_n-1}du_n \\ &= \int_{\mathbf{R}_+^{n-1}} \left(\int_0^\infty k_\lambda(u_1,\ldots,u_n,1)u_n^{\lambda_n-1}du_n\right)\prod_{j=1}^{n-1} u_j^{\lambda_j+\delta_j-1}du_1\cdots du_{n-1}, \end{aligned}$$

then by the assumption for n, it follows

$$\lim_{\delta_n \to 0} k(\lambda_1 + \delta_1, \ldots, \lambda_n + \delta_n) = k(\lambda_1, \ldots, \lambda_n) + o(1)$$
$$(\delta_i \to 0, i = 1, \ldots, n-1).$$

By mathematical induction, we prove that for $n \in \mathbf{N} \setminus \{1\}$, $k(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{n-1})$ is continuous at $(\lambda_1, \ldots, \lambda_{n-1})$.

The lemma is proved.

Lemma 4. *As the assumption of Lemma 2, define*

$$E_i := \{x \in \mathbf{R}_+; x^{\delta_i} \geq 1\}(i = 1, \ldots, n).$$

If there exists a $\eta > 0$, *such that*

$$\max_{1 \leq i \leq n-1} \{|\eta_i|\} < \eta, k(\lambda_1 + \eta_1, \ldots, \lambda_{n-1} + \eta_{n-1}) \in \mathbf{R},$$

$p_i \in \mathbf{R} \backslash \{0, 1\}(i = 1, \ldots, n)$, *and* $0 < \varepsilon < \eta \min_{1 \leq i \leq n}\{|p_i|\}$, *then we have*

$$I_\varepsilon := \varepsilon \int_{E_{n-1}} \cdots \int_{E_1} \left[\int_{\mathbf{R}_+ \backslash E_n} x_n^{\delta_n(\lambda_n + \frac{\varepsilon}{p_n}) - 1} k_\lambda (x_1^{\delta_1} x_n^{\delta_n}, \ldots, x_{n-1}^{\delta_{n-1}} x_n^{\delta_n}, 1) dx_n \right]$$
$$\times \prod_{j=1}^{n-1} x_j^{\delta_j(\lambda_j - \frac{\varepsilon}{p_j}) - 1} dx_1 \cdots dx_{n-1} = k_\lambda + o(1)(\varepsilon \to 0^+). \tag{9}$$

Proof. Setting $y_n = x_n^{-1}$ in the integral of (9), we find

$$I_\varepsilon = \varepsilon \int_{E_{n-1}} \cdots \int_{E_1} \left[\int_{E_n} y_n^{-\delta_n(\lambda_n + \frac{\varepsilon}{p_n}) - 1} k_\lambda (x_1^{\delta_1} y_n^{-\delta_n}, \ldots, x_{n-1}^{\delta_{n-1}} y_n^{-\delta_n}, 1) dy_n \right]$$
$$\times \prod_{j=1}^{n-1} x_j^{\delta_j(\lambda_j - \frac{\varepsilon}{p_j}) - 1} dx_1 \cdots dx_{n-1}.$$

Setting $u_j = x_j^{\delta_j} y_n^{-\delta_n} (j = 1, \ldots, n-1)$ in the above integral, since $\lambda - \lambda_n = \lambda_n$, we find

$$I_\varepsilon = \varepsilon \int_{E_n} y_n^{-1-\delta_n \varepsilon} \left[\int_{y_n^{-\delta_n}}^{\infty} \cdots \int_{y_n^{-\delta_n}}^{\infty} k_\lambda (u_1, \ldots, u_{n-1}, 1) \right.$$
$$\left. \times \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1} \right] dy_n$$

$$= \varepsilon \int_1^\infty x_n^{-1-\varepsilon} \left[\int_{x_n^{-1}}^\infty \cdots \int_{x_n^{-1}}^\infty k_\lambda(u_1, \ldots, u_{n-1}, 1) \right.$$

$$\left. \times \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1} \right] dx_n (x_n = y_n^{\delta_n}). \tag{10}$$

Setting some sets

$$D_j := \{(u_1, \ldots, u_{n-1}) | u_j \in (0, x_n^{-1}), u_k \in (0, \infty)(k \neq j)\}$$

and functions

$$A_j(x_n) := \int \cdots \int_{D_j} k_\lambda(u_1, \ldots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1}$$

$(j = 1, \ldots, n-1)$, then by (10), it follows

$$I_\varepsilon \geq \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1, \ldots, u_{n-1}, 1) \prod_{j=1}^{n-1} u_j^{\lambda_j - \frac{\varepsilon}{p_j} - 1} du_1 \cdots du_{n-1}$$

$$- \varepsilon \sum_{j=1}^{n-1} \int_1^\infty x_n^{-1} A_j(x_n) dx_n. \tag{11}$$

Without loss of generality, we estimate the case of $j = n$, namely,

$$\int_1^\infty x_n^{-1} A_{n-1}(x_n) dx_n = O(1).$$

In fact, setting $\alpha > 0$, such that $|\frac{\varepsilon}{p_{n-1}} + \alpha| < \eta$, since

$$-u_{n-1}^\alpha \ln u_{n-1} \to 0 (u_{n-1} \to 0^+),$$

there exists a constant $M > 0$, such that

$$-u_{n-1}^\alpha \ln u_{n-1} \leq M (u_{n-1} \in (0, 1]),$$

and then by Fubini theorem, it follows

$$0 \leq \int_1^\infty x_n^{-1} A_{n-1}(x_n) dx_n$$

$$= \int_1^\infty x_n^{-1} \left[\int_{\mathbf{R}_+^{n-2}} \int_0^{x_n^{-1}} k_\lambda(u_1, \ldots, u_{n-1}, 1) \right.$$

$$\times\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}du_{n-1}du_1\cdots du_{n-2}\Bigg]dx_n$$

$$=\int_0^1\int_{\mathbf{R}_+^{n-2}}k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}\left(\int_1^{u_{n-1}^{-1}}x_n^{-1}dx_n\right)du_1\cdots du_{n-1}$$

$$=\int_0^1\int_{\mathbf{R}_+^{n-2}}k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}(-\ln u_{n-1})du_1\cdots du_{n-1}$$

$$\le M\int_0^1\int_{\mathbf{R}_+^{n-2}}k_\lambda(u_1,\ldots,u_{n-1},1)$$

$$\times\prod_{j=1}^{n-2}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}u_{n-1}^{\lambda_{n-1}-(\frac{\varepsilon}{p_{n-1}}+\alpha)-1}du_1\cdots du_{n-1}$$

$$\le M\int_{\mathbf{R}_+^{n-1}}k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-2}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}u_{n-1}^{\lambda_{n-1}-(\frac{\varepsilon}{p_{n-1}}+\alpha)-1}du_1\cdots du_{n-1}$$

$$=M\cdot k\left(\lambda_1-\frac{\varepsilon}{p_1},\ldots,\lambda_{n-2}-\frac{\varepsilon}{p_{n-2}},\lambda_{n-1}-(\frac{\varepsilon}{p_{n-1}}+\alpha)\right)<\infty.$$

Hence by (10), we have

$$I_\varepsilon\ge\int_{\mathbf{R}_+^{n-1}}k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}du_1\cdots du_{n-1}-o_1(1).$$

Since by Lemma 3, we find

$$I_\varepsilon\le\varepsilon\int_1^\infty x_n^{-1-\varepsilon}\Bigg[\int_0^\infty\cdots\int_0^\infty k_\lambda(u_1,\ldots,u_{n-1},1)$$

$$\times\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}du_1\cdots du_{n-1}\Bigg]dx_n$$

$$=\int_0^\infty\cdots\int_0^\infty k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1}u_j^{\lambda_j-\frac{\varepsilon}{p_j}-1}du_1\cdots du_{n-1}$$

$$=k(\lambda_1-\frac{\varepsilon}{p_1},\ldots,\lambda_{n-1}-\frac{\varepsilon}{p_{n-1}})=k_\lambda+o_2(1),$$

then we have (9).

The lemma is proved.

Lemma 5. *Suppose that* $n \in \mathbf{N}\backslash\{1\}, \delta_i \in \{-1,1\}, p_i \in \mathbf{R}\backslash\{0,1\}$ $(i=1,\ldots,n)$, $\sum_{i=1}^{n}\frac{1}{p_i}=1$, $\frac{1}{q_n}=1-\frac{1}{p_n}$, $(\lambda_1,\ldots,\lambda_n)\in\mathbf{R}^n$, $\lambda_n=\sum_{i=1}^{n-1}\lambda_i=\frac{\lambda}{2}$, $k_\lambda(x_1,\ldots,x_n)$ (≥ 0) *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}_+^n$, *such that*

$$k_\lambda=\int_{\mathbf{R}_+^{n-1}}k_\lambda(u_1,\ldots,u_{n-1},1)\prod_{j=1}^{n-1}u_j^{\lambda_j-1}du_1\cdots du_{n-1}\in\mathbf{R}.$$

If $f_i\geq 0$ *are measurable functions in* $\mathbf{R}_+$ $(i=1,\ldots,n-1)$, *putting*

$$\tilde{k}(x_1,\ldots,x_n):=k_\lambda(x_1^{\delta_1}x_n^{\delta_n},\ldots,x_{n-1}^{\delta_{n-1}}x_n^{\delta_n},1),$$

then (i) for $p_i>1(i=1,\ldots,n)$, *we have*

$$\begin{aligned}J&:=\left\{\int_0^\infty x_n^{\delta_n\lambda_n q_n-1}\left[\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\prod_{i=1}^{n-1}f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n}dx_n\right\}^{\frac{1}{q_n}}\\&\leq k_\lambda\prod_{i=1}^{n-1}\left\{\int_0^\infty x^{p_i(1-\delta_i\lambda_i)-1}f^{p_i}(x)dx\right\}^{\frac{1}{p_i}};\end{aligned}\tag{12}$$

(ii) for $0<p_1<1, p_i<0(i=2,\ldots,n)$, *we have the reverse of (12).*

Proof.

(i) For $p_i>1(i=1,\ldots,n)$, by Hölder's inequality (cf. [24]) and (7), it follows

$$\begin{aligned}&\left[\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\prod_{i=1}^{n-1}f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n}\\&=\left\{\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\prod_{i=1}^{n-1}\left[x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n}x_j^{\delta_j\lambda_j-1}\right]^{\frac{1}{p_i}}f_i(x_i)\right.\\&\quad\left.\times\left[x_n^{(\delta_n\lambda_n-1)(1-p_n)}\prod_{j=1}^{n-1}x_j^{\delta_j\lambda_j-1}\right]^{\frac{1}{p_n}}dx_1\cdots dx_{n-1}\right\}^{q_n}\\&\leq\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\prod_{i=1}^{n-1}\left[x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n}x_j^{\delta_j\lambda_j-1}\right]^{\frac{q_n}{p_i}}\end{aligned}$$

$$\times f_i^{q_n}(x_i)dx_1\cdots dx_{n-1}$$

$$\times\left\{\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)x_n^{(\delta_n\lambda_n-1)(1-p_n)}\prod_{j=1}^{n-1}x_j^{\delta_j\lambda_j-1}dx_1\cdots dx_{n-1}\right\}^{q_n-1}$$

$$=(k_\lambda)^{q_n-1}x_n^{1-\delta_n q_n\lambda_n}\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)$$

$$\times\prod_{i=1}^{n-1}\left[x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n}x_j^{\delta_j\lambda_j-1}\right]^{\frac{q_n}{p_i}}f_i^{q_n}(x_i)dx_1\cdots dx_{n-1}. \tag{13}$$

Then it follows

$$J\leq(k_\lambda)^{\frac{1}{p_n}}\left\{\int_0^\infty\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\right.$$

$$\left.\times\prod_{i=1}^{n-1}\left[x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n}x_j^{\delta_j\lambda_j-1}\right]^{\frac{q_n}{p_i}}f_i^{q_n}(x_i)dx_1\cdots dx_{n-1}dx_n\right\}^{\frac{1}{q_n}}$$

$$=(k_\lambda)^{\frac{1}{p_n}}\left\{\int_{\mathbf{R}_+^{n-1}}\left(\int_0^\infty\tilde{k}(x_1,\ldots,x_n)x_n^{\delta_n\lambda_n-1}dx_n\right)\right.$$

$$\left.\times\prod_{i=1}^{n-1}\left[x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n-1}x_j^{\delta_j\lambda_j-1}\right]^{\frac{q_n}{p_i}}f_i^{q_n}(x_i)dx_1\cdots dx_{n-1}\right\}^{\frac{1}{q_n}}. \tag{14}$$

For $n\geq 3$, in view of $\sum_{i-1}^{n-1}\frac{q_n}{p_i}=1$, by Hölder's inequality again, it follows

$$J\leq(k_\lambda)^{\frac{1}{p_n}}\left\{\prod_{i=1}^{n-1}\left[\int_{\mathbf{R}_+^{n-1}}\left(\int_0^\infty\tilde{k}(x_1,\ldots,x_n)x_n^{\delta_n\lambda_n-1}dx_n\right)\right.\right.$$

$$\left.\left.\times x_i^{(\delta_i\lambda_i-1)(1-p_i)}\prod_{j=1(j\neq i)}^{n-1}x_j^{\delta_j\lambda_j-1}f_i^{p_i}(x_i)dx_1\cdots dx_{n-1}\right]^{\frac{q_n}{p_i}}\right\}^{\frac{1}{q_n}}$$

$$\leq(k_\lambda)^{\frac{1}{p_n}}\prod_{i=1}^{n-1}\left\{\int_0^\infty\left[\int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\right.\right.$$

$$\times x_i^{\delta_i\lambda_i}\prod_{j=1(j\neq i)}^{n} x_j^{\delta_j\lambda_j-1}dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n\Bigg]x_i^{p_i(1-\delta_i\lambda_i)-1}f_i^{p_i}(x_i)dx_i\Bigg\}^{\frac{1}{p_i}}$$

$$= (k_\lambda)^{\frac{1}{p_n}}\prod_{i=1}^{n-1}\left\{\int_0^\infty \omega_i(x_i)x_i^{p_i(1-\delta_i\lambda_i)-1}f_i^{p_i}(x_i)dx_i\right\}^{\frac{1}{p_i}}.$$

Then by (7), we have (12) (Note: for $n = 2$, we do not use Hölder's inequality again in the above).

(ii) For $0 < p_1 < 1, p_i < 0 (i = 2, \ldots, n)$, by the reverse Hölder's inequality and the same way, we obtain the reverses of (12).
The lemma is proved.

3 Main Results and Applications

As the assumption of Lemma 5, setting

$$\phi_i(x) := x^{p_i(1-\delta_i\lambda_i)-1}(x \in \mathbf{R}_+ = (0,\infty); i = 1,\ldots,n),$$

then we find $[\phi_n(x)]^{q_n-1} = x^{\delta_n q_n\lambda_n-1}$. If $p_i > 1$ $(i = 1,\ldots,n)$, define the following real function spaces:

$$L_{\phi_i}^{p_i}(\mathbf{R}_+) := \left\{f;\ ||f||_{p_i,\phi_i} = \{\int_0^\infty \phi_i(x)|f(x)|^{p_i}dx\}^{\frac{1}{p_i}} < \infty\right\}\ (i = 1,\ldots,n),$$

$$\prod_{i=1}^{n-1}L_{\phi_i}^{p_i}(\mathbf{R}_+) := \left\{(f_1,\ldots,f_{n-1}); f_i \in L_{\phi_i}^{p_i}(\mathbf{R}_+), i = 1,\ldots,n-1\right\},$$

and a multiple Yang-Hilbert-type integral operator

$$T:\prod_{i=1}^{n-1}L_{\phi_i}^{p_i}(\mathbf{R}_+) \to L_{\phi_n^{q_n-1}}^{q_n}(\mathbf{R}_+)$$

as follows: For $f = (f_1,\ldots,f_{n-1}) \in \prod_{i=1}^{n-1}L_{\phi_i}^{p_i}(\mathbf{R}_+)$, there exists a unified expression Tf, satisfying for $x_n \in (0,\infty)$,

$$(Tf)(x_n) := \int_{\mathbf{R}_+^{n-1}}\tilde{k}(x_1,\ldots,x_n)\prod_{i=1}^{n-1}f_i(x_i)dx_1\cdots dx_{n-1}. \tag{15}$$

Then by (12), it follows $Tf \in L^{q_n}_{\phi_n^{q_n-1}}(\mathbf{R}_+)$. T is bounded satisfying

$$||Tf||_{q_n,\phi_n^{q_n-1}} \le k_\lambda \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i}$$

and then $||T|| \le k_\lambda$, where

$$||T|| := \sup_{f(\neq\theta)\in\prod_{i=1}^{n-1} L^{p_i}_{\phi_i}(\mathbf{R}_+)} \frac{||Tf||_{q_n,\phi_n^{q_n-1}}}{\prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i}}. \tag{16}$$

Define the formal inner product of $T(f_1,\ldots,f_{n-1})$ and f_n as

$$(T(f_1,\ldots,f_{n-1}),f_n) := \int_{\mathbf{R}_+^n} \tilde{k}(x_1,\ldots,x_n) \prod_{i=1}^{n} f_i(x_i)dx_1\cdots dx_n. \tag{17}$$

Theorem 1. *As the assumption of Lemma 5, suppose that for any* $(\lambda_1,\ldots,\lambda_n) \in \mathbf{R}^n$, $\lambda_n = \sum_{i=1}^{n-1} \lambda_i = \frac{\lambda}{2}$, *and*

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} k_\lambda(u_1,\ldots,u_{n-1},1) \prod_{j=1}^{n-1} u_j^{\lambda_j-1} du_1\cdots du_{n-1} \in \mathbf{R}_+. \tag{18}$$

If $f_i(\ge 0) \in L^{p_i}_{\phi_i}(\mathbf{R}_+)$, $||f||_{p_i,\phi_i} > 0$ $(i = 1,\ldots,n)$, *then*

(i) for $p_i > 1 (i = 1,\ldots,n)$, *we have* $||T|| = k_\lambda$ *and the following equivalent inequalities:*

$$||T(f_1,\ldots,f_{n-1})||_{q_n,\phi_n^{q_n-1}} < k_\lambda \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i}, \tag{19}$$

$$(T(f_1,\ldots,f_{n-1}),f_n) < k_\lambda \prod_{i=1}^{n} ||f_i||_{p_i,\phi_i}, \tag{20}$$

where the constant factor k_λ *is the best possible;*

(ii) for $0 < p_1 < 1, p_i < 0 (i = 2,\ldots,n)$, *using the formal symbols in the case of (i), we have the equivalent reverses of (19) and (20) with the same best constant factor.*

Proof.

(i) For all $p_i > 1$, if (12) takes the form of equality, then for $n \ge 3$ in (14), there exist C_i and $C_k (i \ne k)$, such that they are not all zero and

$$C_i x_i^{(\delta_i\lambda_i-1)(1-p_i)} \prod_{j=1(j\neq i)}^{n-1} x_j^{\delta_j\lambda_j-1} f_j^{p_j}(x_j)$$

$$= C_k x_k^{(\delta_k\lambda_k-1)(1-p_k)} \prod_{j=1(j\neq k)}^{n-1} x_j^{\delta_j\lambda_j-1} f_j^{p_j}(x_j) \text{ a.e. in } \mathbf{R}_+^n,$$

namely,

$$C_i x_i^{p_i(1-\delta_i\lambda_i)} f_i^{p_i}(x_i) = C_k x_k^{p_k(1-\delta_k\lambda_k)} f_k^{p_k}(x_k) = C \text{ a.e. in } \mathbf{R}_+^n.$$

Assuming that $C_i > 0$, then

$$x_i^{p_i(1-\delta_i\lambda_i)-1} f_i^{p_i}(x_i) = C/(C_i x_i),$$

which contradicts that $0 < ||f||_{p_i,\phi_i} < \infty$ (Note: for $n = 2$, we consider (13) for $f_k^{p_i}(x_k) = 1$ in the above). Hence we have (19).
By Hölder's inequality, it follows

$$(Tf, f_n) = \int_0^\infty \left(x_n^{\delta_n\lambda_n-\frac{1}{q_n}} \int_{\mathbf{R}_+^{n-1}} \tilde{k}(x_1,\ldots,x_n) \prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1} \right)$$

$$\times \left(x_n^{\frac{1}{q_n}-\delta_n\lambda_n} f_n(x_n) \right) dx_n \le ||T(f_1,\ldots,f_{n-1})||_{q_n,\phi_n^{q_n-1}} ||f_n||_{p_n,\phi_n}, \tag{21}$$

and then by (19), we have (20). Assuming that (20) is valid, setting

$$f_n(x_n) := x_n^{\delta_n q_n\lambda_n-1} \left[\int_{\mathbf{R}_+^{n-1}} \tilde{k}(x_1,\ldots,x_n) \prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1} \right]^{q_n-1},$$

then it follows that

$$J = \left\{ \int_0^\infty x_n^{p_n(1-\delta_n\lambda_n)-1} f_n^{p_n}(x_n)dx_n \right\}^{\frac{1}{q_n}}.$$

By (12), it follows $J < \infty$. If $J = 0$, then (19) is trivially valid. Assuming that $0 < J < \infty$, by (20), it follows

$$\int_0^\infty x_n^{p_n(1-\delta_n\lambda_n)-1} f_n^{p_n}(x_n)dx_n = J^{q_n} = (Tf, f_n) < k_\lambda \prod_{i=1}^n ||f_i||_{p_i,\phi_i},$$

$$\left\{\int_0^\infty x_n^{p_n(1-\delta_n\lambda_n)-1} f_n^{p_n}(x_n)dx_n\right\}^{\frac{1}{q_n}} = J < k_\lambda \prod_{i=1}^{n-1} ||f_i||_{p_i,\phi_i},$$

and then (19) is valid, which is equivalent to (20).
We put

$$E_i := \{x \in \mathbf{R}_+; x^{\delta_i} \in [1,\infty)\}(i = 1,\ldots,n).$$

For $\varepsilon > 0$ small enough, setting $\tilde{f}_i(x_i)$ as follows:

$$\tilde{f}_i(x_i) = 0, x_i \in \mathbf{R}_+\backslash E_i;$$
$$\tilde{f}_i(x_i) = x_i^{\delta_i(\lambda_i-\frac{\varepsilon}{p_i})-1}, x \in E_i(i = 1,\ldots,n-1),$$
$$\tilde{f}_n(x_n) = x_n^{\delta_n(\lambda_n+\frac{\varepsilon}{p_n})-1}, x \in \mathbf{R}_+\backslash E_n; \tilde{f}_n(x_n) = 0, x_n \in E_n,$$

if there exists a positive constant $k \le k_\lambda$, such that (20) is still valid when replacing k_λ by k, then in particular, by Lemma 4, we have

$$k_\lambda + o(1) = I_\varepsilon = \varepsilon(T(\tilde{f}_1,\ldots,\tilde{f}_{n-1}),\tilde{f}_n) < \varepsilon k \prod_{i=1}^{n} ||\tilde{f}_i||_{p_i,\phi_i} = k, \tag{22}$$

and $k_\lambda \le k(\varepsilon \to 0^+)$. Hence $k = k_\lambda$ is the best value of (20). We confirm that the constant factor k_λ in (19) is the best possible, otherwise we would reach a contradiction by (21) that the constant factor in (20) is not the best possible. Therefore, we have $||T|| = k_\lambda$.

(ii) For $0 < p_1 < 1, p_i < 0(i = 2,\ldots,n)$, by using the reverse Hölder's inequality and the same way, we have the equivalent reverses of (19) and (20) with the same best constant factor.
The theorem is proved.

Remark 1.

(i) For $\delta_i = 1(i = 1,\ldots,n)$ in (19) and (20), we have the following equivalent inequalities with the non-homogeneous kernel and best possible constant factor k_λ (cf. [25]):

$$\int_{\mathbf{R}_+^n} k_\lambda(x_1x_n,\ldots,x_{n-1}x_n,1)\prod_{i=1}^{n} f_i(x_i)dx_1\cdots dx_n$$
$$< k_\lambda \prod_{i=1}^{n}\left\{\int_0^\infty x_i^{p_i(1-\lambda_i)-1} f^{p_i}(x_i)dx_i\right\}^{\frac{1}{p_i}}, \tag{23}$$

$$\left\{\int_0^{\infty} x_n^{\lambda_n q_n-1}\left[\int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1x_n,\ldots,x_{n-1}x_n,1)\prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n} dx_n\right\}^{\frac{1}{q_n}}$$
$$< k_\lambda\prod_{i=1}^{n-1}\left\{\int_0^{\infty} x^{p_i(1-\lambda_i)-1}f^{p_i}(x)dx\right\}^{\frac{1}{p_i}}; \tag{24}$$

(ii) For $\delta_i = 1(i = 1,\ldots,n-1), \delta_n = -1$ in (19) and (20), replacing $x_n^{\lambda} f_n(x_n)$ by $f_n(x_n)$, we have the following equivalent inequalities with the homogeneous kernel and a best possible constant factor k_λ :

$$\int_{\mathbf{R}_+^n} k_\lambda(x_1,\ldots,x_n)\prod_{i=1}^{n} f_i(x_i)dx_1\cdots dx_n$$
$$< k_\lambda\prod_{i=1}^{n}\left\{\int_0^{\infty} x_i^{p_i(1-\lambda_i)-1}f^{p_i}(x_i)dx_i\right\}^{\frac{1}{p_i}}, \tag{25}$$

$$\left\{\int_0^{\infty} x_n^{\lambda_n q_n-1}\left[\int_{\mathbf{R}_+^{n-1}} k_\lambda(x_1,\ldots,x_n)\prod_{i=1}^{n-1} f_i(x_i)dx_1\cdots dx_{n-1}\right]^{q_n} dx_n\right\}^{\frac{1}{q_n}}$$
$$< k_\lambda\prod_{i=1}^{n-1}\left\{\int_0^{\infty} x^{p_i(1-\lambda_i)-1}f^{p_i}(x)dx\right\}^{\frac{1}{p_i}}. \tag{26}$$

For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}(i = 1,\ldots,n)$, inequality (25) reduces to (5) ($r_n = 2$).

4 Some Examples

Example 1. For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}(i = 1,\ldots,n), r_n = 2, \sum_{i=1}^{n}\frac{1}{r_i} = 1$,

$$k_\lambda(x_1,\ldots,x_n) = \frac{1}{(\sum_{i=1}^{n} x_i)^{\lambda}},$$

by mathematical induction, we can show that

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1}}{(\sum_{i=1}^{n-1} u_i + 1)^{\lambda}} du_1\cdots du_{n-1} = \frac{1}{\Gamma(\lambda)}\prod_{i=1}^{n}\Gamma\left(\frac{\lambda}{r_i}\right). \tag{27}$$

In fact, for $n = 2$, we obtain

$$k_\lambda = \int_{\mathbf{R}_+} \frac{1}{(u_1+1)^\lambda} u_1^{\frac{\lambda}{r_1}-1} du_1 = \frac{1}{\Gamma(\lambda)} \Gamma\left(\frac{\lambda}{r_1}\right) \Gamma\left(\frac{\lambda}{r_2}\right).$$

Assuming that for $n(\geq 2)$, (27) is valid, then for $n+1$, it follows

$$\begin{aligned}
k_\lambda &= \int_{\mathbf{R}_+^n} \frac{1}{(\sum_{i=1}^n u_i + 1)^\lambda} \prod_{j=1}^n u_j^{\frac{\lambda}{r_j}-1} du_1 \cdots du_n \\
&= \int_{\mathbf{R}_+^{n-1}} \prod_{j=2}^n u_j^{\frac{\lambda}{r_j}-1} \left\{ \int_{\mathbf{R}_+} \frac{u_1^{\frac{\lambda}{r_1}-1} du_1}{[u_1 + (\sum_{i=2}^n u_i + 1)]^\lambda} \right\} du_2 \cdots du_n \\
&= \int_{\mathbf{R}_+^{n-1}} \frac{1}{(\sum_{i=2}^n u_i + 1)^\lambda} \prod_{j=2}^n u_j^{\frac{\lambda}{r_j}-1} \left[\int_{\mathbf{R}_+} \frac{v_1^{\frac{\lambda}{r_1}-1} dv_1}{(v_1+1)^\lambda} \right] du_2 \cdots du_n \\
&= \frac{\Gamma(\frac{\lambda}{r_1})\Gamma(\lambda - \frac{\lambda}{r_1})}{\Gamma(\lambda)} \int_{R_+^{n-1}} \frac{1}{(\sum_{i=2}^n u_i + 1)^{\lambda(1-\frac{1}{r_1})}} \prod_{j=2}^n u_j^{\frac{\lambda}{r_j}-1} du_2 \cdots du_n \\
&= \frac{\Gamma(\frac{\lambda}{r_1})\Gamma(\lambda - \frac{\lambda}{r_1})}{\Gamma(\lambda)} \frac{1}{\Gamma(\lambda - \frac{\lambda}{r_1})} \prod_{i=2}^{n+1} \Gamma\left(\frac{\lambda}{r_i}\right) \\
&= \frac{1}{\Gamma(\lambda)} \prod_{i=1}^{n+1} \Gamma\left(\frac{\lambda}{r_i}\right).
\end{aligned}$$

Then by mathematical induction, (27) is valid for $n \in \mathbf{N}\backslash\{1\}$.

Example 2. For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i} (i = 1, \ldots, n), r_n = 2, \sum_{i=1}^n \frac{1}{r_i} = 1$,

$$k_\lambda(x_1, \ldots, x_n) = \frac{1}{\sum_{i=1}^n x_i^\lambda},$$

we can show that

$$k_\lambda = \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1}}{\sum_{i=1}^{n-1} u_i^\lambda + 1} du_1 \cdots du_{n-1} = \frac{1}{\lambda^{n-1}} \prod_{i=1}^n \Gamma\left(\frac{1}{r_i}\right). \tag{28}$$

In fact, setting $v_i = u_i^{\lambda}(i = 1, \ldots, n-1)$ in the above integral, we find $u_i = v_i^{\frac{1}{\lambda}}, du_i = \frac{1}{\lambda} v_i^{\frac{1}{\lambda}-1} dv_i$ and

$$k_\lambda = \frac{1}{\lambda^{n-1}} \int_{\mathbf{R}_+^{n-1}} \frac{\prod_{j=1}^{n-1} v_j^{\frac{1}{r_j}-1}}{\sum_{i=1}^{n-1} v_i + 1} dv_1 \cdots dv_{n-1}.$$

In view of (27), for $\lambda = 1$, we have (28).

Example 3. For $\lambda > 0, \lambda_i = \frac{\lambda}{r_i}(i = 1, \ldots, n), r_n = 2, \sum_{i=1}^{n} \frac{1}{r_i} = 1$,

$$k_\lambda(x_1, \ldots, x_n) = \frac{1}{(\max_{1 \le i \le n}\{x_i\})^{\lambda}},$$

by mathematical induction, we can show that

$$\begin{aligned} k_\lambda &= \int_{R_+^{n-1}} \frac{1}{(\max_{1 \le i \le n-1}\{u_i\} + 1)^{\lambda}} \prod_{j=1}^{n-1} u_j^{\frac{\lambda}{r_j}-1} du_1 \cdots du_{n-1} \\ &= \frac{1}{\lambda^{n-1}} \prod_{i=1}^{n} r_i. \end{aligned} \tag{29}$$

In fact, for $n = 2$, we obtain

$$\begin{aligned} k_\lambda &= \int_{R_+} \frac{u_1^{\frac{\lambda}{r_1}-1}}{(\max\{u_1, 1\})^{\lambda}} du_1 \\ &= \int_0^1 u_1^{\frac{\lambda}{r_1}-1} du_1 + \int_1^{\infty} u_1^{\frac{-\lambda}{r_2}-1} du_1 = \frac{r_1 r_2}{\lambda}. \end{aligned}$$

Assuming that for $n(\ge 2)$, (29) is valid, then for $n+1$, it follows

$$\begin{aligned} k_\lambda &= \int_{R_+^{n-1}} \prod_{j=2}^{n} u_j^{\frac{\lambda}{r_j}-1} \left[\int_0^{\infty} \frac{1}{(\max_{1 \le i \le n}\{u_i, 1\})^{\lambda}} u_1^{\frac{\lambda}{r_1}-1} du_1 \right] du_2 \cdots du_n \\ &= \int_{R_+^{n-1}} \prod_{j=2}^{n} u_j^{\frac{\lambda}{r_j}-1} \left[\int_0^{\max\{u_2, \ldots, u_n, 1\}} \frac{1}{(\max_{2 \le i \le n}\{u_i, 1\})^{\lambda}} u_1^{\frac{\lambda}{r_1}-1} du_1 \right. \\ &\quad \left. + \int_{\max\{u_2, \ldots, u_n, 1\}}^{\infty} \frac{1}{u_1^{\lambda}} u_1^{\frac{\lambda}{r_1}-1} du_1 \right] du_2 \cdots du_n \\ &= \frac{r_1^2}{\lambda(r_1 - 1)} \int_{R_+^{n-1}} \frac{1}{(\max_{2 \le i \le n}\{u_i, 1\})^{\lambda\left(1-\frac{1}{r_1}\right)}} \prod_{j=2}^{n} u_j^{\frac{\lambda}{r_j}-1} du_2 \cdots du_n \end{aligned}$$

$$
\begin{aligned}
&= \frac{r_1^2}{\lambda(r_1-1)}\left(\frac{r_1}{r_1-1}\right)^{n-1} \\
&\quad \times \int_{R_+^{n-1}} \frac{1}{(\max_{2\le i\le n}\{v_i,1\})^{\lambda}} \prod_{j=2}^{n} v_j^{\frac{\lambda}{r_j}\frac{r_1}{r_1-1}-1} dv_2\cdots dv_n \\
&= \frac{r_1^2}{\lambda(r_1-1)}\left(\frac{r_1}{r_1-1}\right)^{n-1} \frac{1}{\lambda^{n-1}} \prod_{i=2}^{n+1} \frac{r_1-1}{r_1} r_i \\
&= \frac{1}{\lambda^n} \prod_{i=1}^{n+1} r_i.
\end{aligned}
$$

Then by mathematical induction, (29) is valid for $n \in \mathbf{N}\backslash\{1\}$.

Example 4. For $\lambda > 0, \lambda_i = \frac{-\lambda}{r_i}(i = 1,\ldots,n)$, $r_n = 2, \sum_{i=1}^{n} \frac{1}{r_i} = 1$,

$$
k_\lambda(x_1,\ldots,x_n) = \left(\min_{1\le i\le n}\{x_i\}\right)^{\lambda},
$$

by mathematical induction, we can show

$$
\begin{aligned}
k_{-\lambda} &= \int_{R_+^{n-1}} (\min\{u_1,\ldots,u_{n-1},1\})^{\lambda} \prod_{j=1}^{n-1} u_j^{\frac{-\lambda}{r_j}-1} du_1\cdots du_{n-1} \\
&= \frac{\prod_{i=1}^{n} r_i}{\lambda^{n-1}}. \qquad (30)
\end{aligned}
$$

In fact, for $n = 2$, we obtain

$$
k_{-\lambda} = \int_0^1 u_1^{\frac{\lambda}{r_2}-1} du_1 + \int_1^{\infty} u_1^{\frac{-\lambda}{r_1}-1} du_1 = \frac{1}{\lambda} r_1 r_2.
$$

Assuming that for $n(\ge 2)$, (30) is valid, then for $n+1$, it follows

$$
\begin{aligned}
k_{-\lambda} &= \int_{R_+^{n-1}} \prod_{j=2}^{n} u_j^{\frac{-\lambda}{r_j}-1} \left[\int_0^{\infty} (\min\{u_1,\ldots,u_n,1\})^{\lambda} u_1^{\frac{-\lambda}{r_1}-1} du_1\right] du_2\cdots du_n \\
&= \int_{R_+^{n-1}} \prod_{j=2}^{n} u_j^{\frac{-\lambda}{r_j}-1} \left[\int_0^{\min\{u_2,\ldots,u_n,1\}} u_1^{\lambda} u_1^{\frac{-\lambda}{r_1}-1} du_1\right. \\
&\quad \left. + \int_{\min\{u_2,\ldots,u_n,1\}}^{\infty} (\min\{u_2,\ldots,u_n,1\})^{\lambda} u_1^{\frac{-\lambda}{r_1}-1} du_1\right] du_2\cdots du_n
\end{aligned}
$$

$$= \frac{r_1^2}{\lambda(r_1-1)} \int_{R_+^{n-1}} (\min\{u_2,\ldots,u_n,1\})^{\lambda(1-\frac{1}{r_1})} \prod_{j=2}^{n} u_j^{\frac{-\lambda(1-\frac{1}{r_1})}{(1-\frac{1}{r_1})r_j}-1} du_2\cdots du_n$$

$$= \frac{r_1^2}{\lambda(r_1-1)} \frac{1}{[\lambda(1-\frac{1}{r_1})]^{n-1}} \prod_{i=2}^{n+1} \left(1-\frac{1}{r_1}\right) r_i$$

$$= \frac{1}{\lambda^n} \prod_{i=1}^{n+1} r_i.$$

Then by mathematical induction, (30) is valid for $n \in \mathbf{N}\backslash\{1\}$.

Remark 2.

(i) In particular, for $n = 2$ in (23), we have

$$\int_0^\infty \int_0^\infty k_\lambda(xy,1) f(x) g(y) dxdy$$
$$< k_\lambda \left\{ \int_0^\infty x^{p(1-\frac{\lambda}{2})-1} f^p(x) dx \right\}^{\frac{1}{p}} \left\{ \int_0^\infty x^{q(1-\frac{\lambda}{2})-1} g^q(x) dx \right\}^{\frac{1}{q}}, \quad (31)$$

where $k_\lambda = \int_0^\infty k_\lambda(u,1) u^{\frac{\lambda}{2}-1} du > 0 (\lambda \in \mathbf{R})$ is the best possible. Inequality (31) is an extension of (4) and (8.1.7) in [8].

(ii) In Examples 1–3, by Theorem 1, since for any $(\lambda_1,\ldots,\lambda_n) \in \mathbf{R}^n (\lambda_n = \sum_{i=1}^n \lambda_i = \frac{\lambda}{2})$, we obtain $0 < k_\lambda < \infty$, then we have $||T|| = k_\lambda$ and the equivalent inequalities (19) and (20) with the particular kernels and some equivalent reverses. In Example 4, still using Theorem 1, we find $0 < ||T|| = k_{-\lambda} < \infty$ and the equivalent inequalities (19) and (20) with the particular kernel and some equivalent reverses.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61370186), and 2013 Knowledge Construction Special Foundation Item of Guangdong Institution of Higher Learning College and University (No. 2013KJCX0140).

References

1. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities. Cambridge University Press, Cambridge (1934)
2. Mitrinović, D.S., Pečarić, J.E., Fink, A.M.: Inequalities Involving Functions and Their Integrals and Derivatives. Kluwer Academic Publishers, Boston (1991)
3. Yang, B.C.: A survey of the study of Hilbert-type inequalities with parameters. Adv. Math. **28**(3), 257–268 (2009)
4. Zhang, K.W.: A bilinear inequality. J. Math. Anal. Appl. **271**, 288–296 (2002)

5. Yang, B.C.: On a extension of Hilbert's integral inequality with some parameters. Aust. J. Math. Anal. Appl. **1**(1), Article ID 11, 1–8 (2004)
6. Yang, B.C.: A new Hilbert's type integral inequality. Soochow J. Math. **33**(4), 849–859 (2007)
7. Yang, B.C.: A Hilbert-type integral inequality with a non-homogeneous kernel. J. Xiamen Univ. (Nat. Sci.) **48**(2), 165–169 (2009)
8. Yang, B.C.: The Norm of Operator and Hilbert-Type Inequalities. Science Press, Beijing (2009)
9. Benyi, A., Oh, C.: Best constant for certain multilinear integral operator. J. Inequal. Appl. **2006**, 1–12 (2006). Article ID 28582
10. Hong, H.: All-side generalization about Hardy-Hilbert integral inequalities. Acta Math. Sin. **44**(4), 619–625 (2001)
11. He, L.P., Yu, J., Gao, M.Z.: An extension of Hilbert's integral inequality. J. Shaoguan Univ. (Nat. Sci.) **23**(3), 25–30 (2002)
12. Yang, B.C.: On a multiple Hardy-Hilbert's integral inequality. Chin. Ann. Math. **24A**(6), 25–30 (2003)
13. Yang, B.C., Rassias, T.M.: On the way of weight coefficient and research for Hilbert-type inequalities. Math. Inequal. Appl. **6**(4), 625–658 (2003)
14. Yang, B.C., Brnetić, I., Krnić, M., Pečarić, J.: Generalization of Hilbert and Hardy-Hilbert integral inequalities. Math. Inequal. Appl. **8**(2), 259–272 (2005)
15. Yang, B.C.: On the norm of an integral operator and applications. J. Math. Anal. Appl. **321**, 182–192 (2006)
16. Yang, B.C.: On the norm of a self-adjoint operator and a new bilinear integral inequality. Acta Math. Sin. Engl. Ser. **23**(7), 1311–1316 (2007)
17. Milovanovic, G.V., Rassias, M.T.: Some properties of a hypergeometric function which appear in an approximation problem. J. Glob. Optim. **57**, 1173–1192 (2013)
18. Rassias, M.T., Yang, B.C.: On half-discrete Hilbert's inequality. Appl. Math. Comput. **220**, 75–93 (2013)
19. Rassias, M.T., Yang, B.C.: A multidimensional half-discrete Hilbert-type inequality and the Riemann zeta function. Appl. Math. Comput. **225**, 263–277 (2013)
20. Rassias, M.T., Yang, B.C.: On a multidimensional half-discrete Hilbert-type inequality related to the hyperbolic cotangent function. Appl. Math. Comput. **242**, 800–813 (2013)
21. Rassias, M.T., Yang, B.C.: A multidimensional Hilbert-type integral inequality related to the Riemann zeta function. In: Daras, N.J. (ed.) Applications of Mathematics and Informatics in Science and Engineering, pp. 417–433. Springer, New York (2014)
22. Milovanovic, G.V., Rassias, M.T. (eds.): Analytic Number Theory, Approximation Theory and Special Functions. Springer, New York (2014)
23. Kuang, J.C.: Introduction to Real Analysis. Human Education Press, Changsha (1996)
24. Kuang, J.C.: Applied Inequalities. Shandong Science Technic Press, Jinan (2004)
25. Huang, Q.L., Yang, B.C.: A multiple Hilbert-type inequality with a non-homogeneous kernel. J. Inequal. Appl. **2013**, 73 (2013). doi:10.1186/1029-242X-2013-73

Parameterized Yang–Hilbert-Type Integral Inequalities and Their Operator Expressions

Bicheng Yang and Michael Th. Rassias

Abstract Applying methods of Real Analysis and Functional Analysis, we build two weight functions with parameters and provide two kinds of parameterized Yang–Hilbert-type integral inequalities with the best constant factors. Equivalent forms, the reverses, and the operator expressions are also given. In particular, the Hardy-type inequalities and Hardy-type operators are studied. Additionally, a number of examples with two kinds of particular kernels are considered.

Keywords: Hardy-type integral operator • Yang–Hilbert-type integral inequality • Hölder's inequality • Measurable function • Weight function • Equivalent form • Operator expression

1 Introduction

If $f(x),\ g(y) \geq 0$, satisfy

$$0 < \int_0^\infty f^2(x)dx < \infty$$

B. Yang (✉)
Department of Mathematics, Guangdong University of Education, Guangzhou, Guangdong 510303, China
e-mail: bcyang@gdei.edu.cn; bcyang818@163.com

M.Th. Rassias
Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000, USA

Department of Mathematics, ETH-Zürich, 8092, Zürich, Switzerland
e-mail: michailrassias@math.princeton.edu; michael.rassias@math.ethz.ch

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_27

and

$$0 < \int_0^\infty g^2(y)dy < \infty,$$

then we have the following well-known Hilbert's integral inequality (cf. [1])

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y}dxdy < \pi \left(\int_0^\infty f^2(x)dx \int_0^\infty g^2(y)dy \right)^{\frac{1}{2}}, \tag{1}$$

where the constant factor π is the best possible. The operator expression of (1) was studied in [2] and [3].

In 1925, by introducing one pair of conjugate exponents (p, q), that is $\frac{1}{p} + \frac{1}{q} = 1$, Hardy [4] provided an extension of (1) as follows:

For $p > 1$, $f(x), g(y) \geq 0$, satisfying

$$0 < \int_0^\infty f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty g^q(y)dy < \infty,$$

we have

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y}dxdy < \frac{\pi}{\sin(\pi/p)} \left(\int_0^\infty f^p(x)dx \right)^{\frac{1}{p}} \left(\int_0^\infty g^q(y)dy \right)^{\frac{1}{q}}, \tag{2}$$

where the constant factor $\frac{\pi}{\sin(\pi/p)}$ is still the best possible. Inequality (2) is known as Hardy-Hilbert's integral inequality, and is important in analysis and its applications (cf. [5, 6]).

Definition 1. If $\lambda \in \mathbf{R} = (-\infty, \infty)$, $k_\lambda(x, y)$ is a non-negative measurable function in $\mathbf{R}_+^2 = \mathbf{R}_+ \times \mathbf{R}_+$, satisfying

$$k_\lambda(tx, ty) = t^{-\lambda} k_\lambda(x, y),$$

for any $t, x, y > 0$, then we call $k_\lambda(x, y)$ the homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^2$.

In 1934, replacing $\frac{1}{x+y}$ in (2) by a general homogeneous kernel of degree-1, as $k_1(x, y)$, Hardy et al. presented an extension of (2) with the best possible constant factor

$$k_p = \int_0^\infty k_1(t, 1) t^{\frac{-1}{p}} dt \in \mathbf{R}_+ = (0, \infty)$$

obtaining (cf. [5, Th. 319]):

$$\int_0^\infty \int_0^\infty k_1(x,y)f(x)g(y)dxdy < k_p \left(\int_0^\infty f^p(x)dx\right)^{\frac{1}{p}} \left(\int_0^\infty g^q(y)dy\right)^{\frac{1}{q}}. \tag{3}$$

The following inequality with the non-homogeneous kernel $h(xy)$ similar to (3) was studied (cf. [5, Th. 350]):

For $h(t) > 0$, satisfying $\phi(s) := \int_0^\infty h(t)t^{s-1}dt \in \mathbf{R}_+, f(x), g(y) \geq 0$,

$$0 < \int_0^\infty x^{p-2}f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty g^q(y)dy < \infty,$$

we have

$$\int_0^\infty \int_0^\infty h(xy)f(x)g(y)dxdy < \phi(\frac{1}{p}) \left(\int_0^\infty x^{p-2}f^p(x)dx\right)^{\frac{1}{p}} \left(\int_0^\infty g^q(y)dy\right)^{\frac{1}{q}}. \tag{4}$$

Remark 1. Hardy could not prove that the constant factor in (4) is the best possible and did not consider the operator expressions of (3) and (4) (cf. [5, Chapter 9]). We shall call (3) and (4) Hardy-Hilbert-type integral inequalities, which only contain one pair of conjugate exponents (p, q).

In 1998, by introducing an independent parameter $\lambda > 0$, Yang [7, 8] gave an extension of (1) as follows:

For $f(x),\ g(y) \geq 0$, such that

$$0 < \int_0^\infty x^{1-\lambda}f^2(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{1-\lambda}g^2(y)dy < \infty,$$

we have

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(x+y)^\lambda}dxdy < B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right) \left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}, \tag{5}$$

where the constant factor $B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right)$ is the best possible, and

$$B(u, v) := \int_0^\infty \frac{t^{u-1}}{(t+1)^{u+v}} dt (u, v > 0) \tag{6}$$

is the beta function (cf. [9]).

In 2004, by introducing two pairs of conjugate exponents (p, q) and (r, s), that is $\frac{1}{p} + \frac{1}{q} = \frac{1}{r} + \frac{1}{s} = 1$, and an independent parameter $\lambda > 0$, Yang [10] gave the following extension of (3):

For $p, r > 1$, $f(x), g(y) \geq 0$, such that

$$0 < \int_0^\infty x^{p(1-\frac{\lambda}{r})-1} f^p(x) dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\frac{\lambda}{s})-1} g^q(y) dy < \infty,$$

it holds

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x^\lambda + y^\lambda} dxdy < \frac{\pi}{\lambda \sin(\pi/r)} \left[\int_0^\infty x^{p(1-\frac{\lambda}{r})-1} f^p(x) dx\right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\frac{\lambda}{s})-1} g^q(y) dy\right]^{\frac{1}{q}}, \tag{7}$$

where the constant factor

$$\frac{\pi}{\lambda \sin(\pi/r)}$$

is the best possible.

For $\lambda = 1, r = q, s = p$, inequality (7) reduces to (2); for $\lambda = 1, r = p, s = q$, inequality (7) reduces to the dual form of (2) as follows:

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x+y} dxdy < \frac{\pi}{\sin(\pi/p)} \left(\int_0^\infty x^{p-2} f^p(x) dx\right)^{\frac{1}{p}} \left(\int_0^\infty y^{q-2} g^q(y) dy\right)^{\frac{1}{q}}. \tag{8}$$

In 2009, replacing $1/(x^\lambda + y^\lambda)$ in (7), by a general homogeneous kernel of degree $-\lambda$, as $k_\lambda(x, y)$, Yang [11, 12] proved an extension of (7) in the following form:

For $\lambda > 0$, $f(x), g(y) \geq 0$, such that

$$0 < \int_0^\infty x^{p(1-\frac{\lambda}{r})-1} f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\frac{\lambda}{s})-1} g^q(y)dy < \infty,$$

it follows

$$\int_0^\infty \int_0^\infty k_\lambda(x,y) f(x)g(y)dxdy$$
$$< k_\lambda(r) \left[\int_0^\infty x^{p(1-\frac{\lambda}{r})-1} f^p(x)dx\right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\frac{\lambda}{s})-1} g^q(y)dy\right]^{\frac{1}{q}}, \quad (9)$$

where the constant factor

$$k_\lambda(r) := \int_0^\infty k_\lambda(t,1) t^{\frac{\lambda}{r}-1} dt \in \mathbf{R}_+$$

is the best possible.

In [13], Yang presented also the following new inequality with a non-homogeneous kernel similar to (4):

For $f(x), g(y) \geq 0$, satisfying

$$0 < \int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

we have

$$\int_0^\infty \int_0^\infty h(xy) f(x)g(y)dxdy$$
$$< \phi(\sigma) \left(\int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx\right)^{\frac{1}{p}} \left(\int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy\right)^{\frac{1}{q}}, \quad (10)$$

where the constant factor

$$\phi(\sigma) = \int_0^\infty h(t) t^{\sigma-1} dt \in \mathbf{R}_+$$

is the best possible.

Remark 2. For $\lambda = 1, r = q, s = p$, it follows that inequality (9) reduces to (3). Hence, (9) is an extension of (3) with two pairs of conjugate exponents and an independent parameter. In 2014, Yang [14] proved that inequalities (9) and (10) are equivalent for

$$h(u) = k_\lambda(u, 1),$$

and considered the operator expressions of (9) and (10). We call (9) together with (10) as Yang–Hilbert-type integral inequalities in the first quadrant. Also we can call some similar inequalities as Yang–Hilbert-type integral inequalities in the whole plane.

In 2007, Yang [15] introduced a Hilbert-type integral inequality in the whole plane as follows:

For $f(x), g(y) \geq 0$, such that

$$0 < \int_{-\infty}^{\infty} e^{-\lambda x} f^2(x)dx < \infty$$

and

$$0 < \int_0^{\infty} \int_{-\infty}^{\infty} e^{-\lambda y} g^2(y)dy < \infty,$$

it holds

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{f(x)g(y)}{(1 + e^{x+y})^\lambda} dxdy$$
$$< B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right) \left(\int_{-\infty}^{\infty} e^{-\lambda x} f^2(x)dx \int_{-\infty}^{\infty} e^{-\lambda y} g^2(y)dy\right)^{\frac{1}{2}}, \tag{11}$$

where the constant factor $B(\frac{\lambda}{2}, \frac{\lambda}{2})$ $(\lambda > 0)$ is the best possible.

For the case when $0 < \lambda < 1, p > 1, \frac{1}{p} + \frac{1}{q} = 1$, Yang [16] proved in 2008, the following Hilbert-type integral inequality in the whole plane:

For $p > 1$, $f(x), g(y) \geq 0$, satisfying

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\frac{\lambda}{2})-1} f^p(x)dx < \infty$$

and

$$0 < \int_{-\infty}^{\infty} |y|^{q(1-\frac{\lambda}{2})-1} g^q(y)dy < \infty,$$

we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{|1+xy|^{\lambda}}f(x)g(y)dxdy$$
$$<k_{\lambda}\left[\int_{-\infty}^{\infty}|x|^{p(1-\frac{\lambda}{2})-1}f^{p}(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\frac{\lambda}{2})-1}g^{q}(y)dy\right]^{\frac{1}{q}}, \quad (12)$$

where the constant

$$k_{\lambda}=B\left(\frac{\lambda}{2},\frac{\lambda}{2}\right)+2B\left(1-\lambda,\frac{\lambda}{2}\right)$$

is still the best possible.

Additionally, Yang et al. [17–26] provided also some other Hilbert-type integral inequalities in the whole plane. Rassias et al. [27–32] presented as well some different new Hilbert-type inequalities.

In this paper, applying methods of Real Analysis and Functional Analysis, we build two weight functions with parameters, and provide two kinds of parameterized Yang–Hilbert-type integral inequalities with the best constant factors. Equivalent forms, the reverses, and the operator expressions are also given. In particular, the Hardy-type inequalities and Hardy-type operators are studied. Furthermore, a number of examples with two kinds of particular kernels are considered.

2 Yang–Hilbert-Type Integral Inequalities in the First Quadrant

In this section, we present a weight function and study some Yang–Hilbert-type integral inequalities in the first quadrant with parameters and the best constant factors. Equivalent forms, the reverses, the Hardy-type inequalities, the operator expressions, and some particular examples are also discussed.

2.1 Definition of Weight Function and a Lemma

Definition 2. If $\sigma \in \mathbf{R}, h(t)$ is a non-negative measurable function in $\mathbf{R}_+$, define the following weight function:

$$\omega(\sigma,y):=y^{\sigma}\int_{0}^{\infty}h(xy)x^{\sigma-1}dx(y\in\mathbf{R}_+). \quad (13)$$

Setting $t = xy$ in (13), we obtain

$$\omega(\sigma, y) = k(\sigma) := \int_0^\infty h(t)t^{\sigma-1}dt. \tag{14}$$

Lemma 1. *If $p > 0$ $(p \neq 1)$, $\frac{1}{p} + \frac{1}{q} = 1, \sigma \in \mathbf{R}$, both $h(t)$ and $f(t)$ are non-negative measurable functions in $\mathbf{R}_+$, and $k(\sigma)$ is defined by (14), then, (i) for $p > 1$, we have the following inequality:*

$$J := \int_0^\infty y^{p\sigma-1}\left(\int_0^\infty h(xy)f(x)dx\right)^p dy \leq k^p(\sigma)\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx; \tag{15}$$

(ii) for $0 < p < 1$, we have the reverse of (15).

Proof. (i) By the weighted Hölder's inequality (cf. [33]) and (13), it follows that

$$\begin{aligned}\int_0^\infty h(xy)f(x)dx &= \int_0^\infty h(xy)\left[\frac{x^{(1-\sigma)/q}}{y^{(1-\sigma)/p}}f(x)\right]\left[\frac{y^{(1-\sigma)/p}}{x^{(1-\sigma)/q}}\right]dx\\ &\leq \left[\int_0^\infty h(xy)\frac{x^{(1-\sigma)p/q}}{y^{1-\sigma}}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty h(xy)\frac{y^{(1-\sigma)q/p}}{x^{1-\sigma}}dx\right]^{\frac{1}{q}}\\ &= (\omega(\sigma,y))^{\frac{1}{q}}y^{\frac{1}{p}-\sigma}\left[\int_0^\infty h(xy)\frac{x^{(1-\sigma)(p-1)}}{y^{1-\sigma}}f^p(x)dx\right]^{\frac{1}{p}}.\end{aligned} \tag{16}$$

Then by (14) and Fubini's theorem (cf. [34]), we have

$$\begin{aligned}J &\leq k^{p-1}(\sigma)\int_0^\infty\int_0^\infty h(xy)\frac{x^{(1-\sigma)(p-1)}}{y^{1-\sigma}}f^p(x)dxdy\\ &= k^{p-1}(\sigma)\int_0^\infty\left[\int_0^\infty h(xy)\frac{x^{(1-\sigma)(p-1)}}{y^{1-\sigma}}dy\right]f^p(x)dx\\ &= k^{p-1}(\sigma)\int_0^\infty \omega(\sigma,x)x^{p(1-\sigma)-1}f^p(x)dx.\end{aligned} \tag{17}$$

By (14), we obtain (15).

(ii) For $0 < p < 1$, by the reverse of the weighted Hölder's inequality (cf. [33]), combined with (13) and (14), we obtain the reverse of (16) and (17). Then we get the reverse of (15) by using (14). This completes the proof of the lemma. □

2.2 Two Equivalent Inequalities as well as the Reverses with the Best Possible Constant Factors

Theorem 1. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1, \sigma \in \mathbf{R}, h(t) \geq 0$, *and*

$$k(\sigma) = \int_0^\infty h(t)t^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$, *such that*

$$0 < \int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} I &:= \int_0^\infty \int_0^\infty h(xy)f(x)g(y)dxdy \\ &< k(\sigma)\left[\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \end{aligned} \tag{18}$$

$$J = \int_0^\infty y^{p\sigma-1}\left(\int_0^\infty h(xy)f(x)dx\right)^p dy < k^p(\sigma)\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx, \tag{19}$$

where the constant factors $k(\sigma)$ *and* $k^p(\sigma)$ *are the best possible.*

Proof. We first proved that (16) preserves the form of a strict inequality for any $y \in \mathbf{R}_+$. Otherwise, there exists a $y > 0$, such that (16) becomes an equality. Then, there exist two constants A and B, such that they are not all zero, and (cf. [33])

$$A\frac{x^{(1-\sigma)p/q}}{y^{1-\sigma}}f^p(x) = B\frac{y^{(1-\sigma)q/p}}{x^{1-\sigma}} \quad \text{a. e. in } \mathbf{R}_+.$$

If $A = 0$, then $B = 0$, which is impossible. Suppose that $A \neq 0$. Then it follows that

$$x^{p(1-\sigma)-1}f^p(x) = y^{(1-\sigma)q}\frac{B}{Ax} \quad \text{a. e. in } \mathbf{R}_+,$$

which contradicts the fact that

$$0 < \int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx < \infty,$$

in virtue of

$$\int_0^\infty \frac{1}{x}dx = \infty.$$

Hence, both (16) and (17) preserve the forms of strict inequalities, and thus we have (19).

By Hölder's inequality (cf. [33]), we obtain

$$\begin{aligned} I &= \int_0^\infty \left(y^{\sigma-\frac{1}{p}} \int_0^\infty h(xy)f(x)dx\right)(y^{\frac{1}{p}-\sigma}g(y))dy \\ &\le J^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}. \end{aligned} \tag{20}$$

Then by (19), we get (18). On the other hand, assuming that (18) is valid, we set

$$g(y) := y^{p\sigma-1}\left(\int_0^\infty h(xy)f(x)dx\right)^{p-1}, y \in \mathbf{R}_+.$$

Then we obtain

$$J = \int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy.$$

By (15), in view of

$$0 < \int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx < \infty,$$

it follows that $J < \infty$. If $J = 0$, then (19) is trivially valid; if $J > 0$, then by (18), we have

$$\begin{aligned} 0 &< \int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy = J = I \\ &< k(\sigma)\left[\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \end{aligned}$$

$$J^{\frac{1}{p}} = \left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{p}} < k(\sigma)\left[\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}},$$

and then (19) follows, which is equivalent to (18).

For any $n \in \mathbf{N}$ (where $\mathbf{N}$ is the set of positive integers), we define the functions $f_n(x)$ and $g_n(y)$ as follows:

$$f_n(x) := \begin{cases} 0, x \in (0,1) \\ x^{\sigma-\frac{1}{np}-1}, x \in [1,\infty) \end{cases}, \; g_n(y) := \begin{cases} y^{\sigma+\frac{1}{nq}-1}, y \in (0,1] \\ 0, y \in (1,\infty) \end{cases}.$$

Then we find

$$\begin{aligned} L_n &:= \left[\int_0^\infty x^{p(1-\sigma)-1} f_n^p(x)dx\right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g_n^q(y)dy\right]^{\frac{1}{q}} \\ &= \left(\int_1^\infty x^{-\frac{1}{n}-1}dx\right)^{\frac{1}{p}} \left(\int_0^1 y^{\frac{1}{n}-1}dy\right)^{\frac{1}{q}} = n. \end{aligned}$$

By Fubini's theorem, it follows that

$$\begin{aligned} I_n &:= \int_0^\infty \int_0^\infty h(xy) f_n(x) g_n(y)dxdy \\ &= \int_1^\infty x^{\sigma-\frac{1}{np}-1} \left(\int_0^1 h(xy) y^{\sigma+\frac{1}{nq}-1}dy\right) dx \\ &= \int_1^\infty x^{-\frac{1}{n}-1} \left(\int_0^x h(t) t^{\sigma+\frac{1}{nq}-1}dt\right) dx \\ &= \int_1^\infty x^{-\frac{1}{n}-1} \left(\int_0^1 h(t) t^{\sigma+\frac{1}{nq}-1}dt + \int_1^x h(t) t^{\sigma+\frac{1}{nq}-1}dt\right) dx \\ &= n\int_0^1 h(t) t^{\sigma+\frac{1}{nq}-1}dt + \int_1^\infty x^{-\frac{1}{n}-1} \left(\int_1^x h(t) t^{\sigma+\frac{1}{nq}-1}dt\right) dx \\ &= n\int_0^1 h(t) t^{\sigma+\frac{1}{nq}-1}dt + \int_1^\infty \left(\int_t^\infty x^{-\frac{1}{n}-1}dx\right) h(t) t^{\sigma+\frac{1}{nq}-1}dt \\ &= n\left(\int_0^1 h(t) t^{\sigma+\frac{1}{nq}-1}dt + \int_1^\infty h(t) t^{\sigma-\frac{1}{np}-1}dt\right). \end{aligned}$$

If there exists a positive number $k \leq k(\sigma)$, such that (18) is still valid when replacing $k(\sigma)$ by k, then in particular, it follows that

$$\frac{1}{n} I_n < k\frac{1}{n} L_n,$$

and

$$\int_0^1 h(t) t^{\sigma+\frac{1}{nq}-1}dt + \int_1^\infty h(t) t^{\sigma-\frac{1}{np}-1}dt < k.$$

Since both

$$\{h(t)t^{\sigma+\frac{1}{nq}-1}\}_{n=1}^{\infty}\ (t\in(0,1])$$

and

$$\{h(t)t^{\sigma-\frac{1}{np}-1}\}_{n=1}^{\infty}\ (t\in(1,\infty))$$

are non-negative and increasing, then by Levi's theorem (cf. [34]), we get

$$\begin{aligned}k(\sigma)&=\int_0^1 h(t)t^{\sigma-1}dt+\int_1^\infty h(t)t^{\sigma-1}dt\\&=\lim_{n\to\infty}\left(\int_0^1 h(t)t^{\sigma+\frac{1}{nq}-1}dt+\int_1^\infty h(t)t^{\sigma-\frac{1}{np}-1}dt\right)\le k.\end{aligned}$$

Thus $k=k(\sigma)$ is the best possible constant factor of (18).

The constant factor in (19) is still the best possible. Otherwise, by (20) we would reach the contradiction that the constant factor in (18) is not the best possible.

This completes the proof of the theorem. □

Theorem 2. *Replacing $p>1$ by $0<p<1$ in Theorem 1, we obtain the equivalent reverses of (18) and (19). If there exists a constant $\delta_0>0$, such that for any $\tilde{\sigma}\in(\sigma-\delta_0,\sigma]$,*

$$k(\tilde{\sigma})=\int_0^\infty h(t)t^{\tilde{\sigma}-1}dt\in\mathbf{R}_+,$$

then the constant factors in the reverses of (18) and (19) are the best possible.

Proof. By Lemma 1 and the reverse of Hölder's inequality, we get the reverses of (18), (19), and (20). Similarly, we can set $g(y)$ as in Theorem 1, and prove that the reverses of (18) and (19) are equivalent.

For $n>\frac{2}{\delta_0|q|}(n\in\mathbf{N})$, we set $f_n(x)$ and $g_n(y)$ as in Theorem 1. If there exists a positive number $k\ge k(\sigma)$, such that the reverse of (18) is valid when replacing $k(\sigma)$ by k, then it follows that

$$\frac{1}{n}I_n>k\frac{1}{n}L_n,$$

and

$$\int_0^1 h(t)t^{\sigma+\frac{1}{nq}-1}dt+\int_1^\infty h(t)t^{\sigma-\frac{1}{np}-1}dt>k. \tag{21}$$

Since $\{h(t)t^{\sigma-\frac{1}{np}-1}\}_{n=1}^{\infty}(t \in (1,\infty))$ is still non-negative and increasing, by Levi's theorem it follows that

$$\lim_{n\to\infty}\int_1^{\infty} h(t)t^{\sigma-\frac{1}{np}-1}dt = \int_1^{\infty} h(t)t^{\sigma-1}dt.$$

Since

$$0 \le h(t)t^{\sigma+\frac{1}{nq}-1} \le h(t)t^{(\sigma-\frac{\delta_0}{2})-1} \left(t \in (0,1], n > \frac{2}{\delta_0|q|}\right),$$

and

$$0 \le \int_0^1 h(t)t^{(\sigma-\frac{\delta_0}{2})-1}dt \le k\left(\sigma-\frac{\delta_0}{2}\right) < \infty,$$

then by Lebesgue's dominated convergence theorem (cf. [34]), it follows that

$$\lim_{n\to\infty}\int_0^1 h(t)t^{\sigma+\frac{1}{nq}-1}dt = \int_0^1 h(t)t^{\sigma-1}dt.$$

In view of the above results and (21), we have

$$k(\sigma) = \lim_{n\to\infty}\left(\int_0^1 h(t)t^{\sigma+\frac{1}{nq}-1}dt + \int_1^{\infty} h(t)t^{\sigma-\frac{1}{np}-1}dt\right) \ge k.$$

Then $k = k(\sigma)$ is the best possible constant factor for the reverse of (18).

Following the same method, we can prove that the constant factor in the reverse of (19) is the best possible, by the use of the reverse of (20).
This completes the proof of the theorem. □

2.3 *Yang–Hilbert-Type Integral Inequalities in the First Quadrant with Multi-Variables*

Theorem 3. *Suppose that* $p > 1,\ \frac{1}{p}+\frac{1}{q} = 1,\ h(t) \ge 0, \sigma \in \mathbf{R}$,

$$k(\sigma) = \int_0^{\infty} h(t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$\delta_i \in \{-1,1\},\ 0 \le a_i < b_i \le \infty,\ v_i'(s) > 0\ (s \in (a_i,b_i))$ *and*

$$v_i(a_i^+) = \lim_{s\to a_i+} v_i(s) = 0,$$

$$v_i(b_i^-) = \lim_{s\to b_i-} v_i(s) = \infty\ (i = 1,2).$$

If $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\delta_1\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\delta_2\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2}\int_{a_1}^{b_1} h(v_1^{\delta_1}(x)v_2^{\delta_2}(y))f(x)g(y)dxdy \\
&< k(\sigma)\left[\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\delta_1\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx\right]^{\frac{1}{p}} \\
&\times\left[\int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\delta_2\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy\right]^{\frac{1}{q}},
\end{aligned} \tag{22}$$

$$\begin{aligned}
&\int_{a_2}^{b_2} \frac{v_2'(y)}{(v_2(y))^{1-p\delta_2\sigma}}\left(\int_{a_1}^{b_1} h(v_1^{\delta_1}(x)v_2^{\delta_2}(y))f(x)dx\right)^p dy \\
&< k^p(\sigma)\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\delta_1\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx,
\end{aligned} \tag{23}$$

where the constant factors $k(\sigma)$ *and* $k^p(\sigma)$ *are the best possible.*

Proof. Setting

$$x = v_1^{\delta_1}(s),\ y = v_2^{\delta_2}(t)$$

in (18), since $\delta_i \in \{-1, 1\}$, we get

$$dx = \delta_1 v_1^{\delta_1-1}(s)v_1'(s)ds,\ \ dy = \delta_2 v_2^{\delta_2-1}(t)v_2'(t)dt,$$

and

$$\begin{aligned}
I &= |\delta_1\delta_2|\int_{a_2}^{b_2}\int_{a_1}^{b_1} h(v_1^{\delta_1}(s)v_2^{\delta_2}(t))f(v_1^{\delta_1}(s))g(v_2^{\delta_2}(t))v_1^{\delta_1-1}(s)v_1'(s)v_2^{\delta_2-1}(t)v_2'(t)dsdt \\
&= \int_{a_2}^{b_2}\int_{a_1}^{b_1} h(v_1^{\delta_1}(s)v_2^{\delta_2}(t))(f(v_1^{\delta_1}(s))v_1^{\delta_1-1}(s)v_1'(s))(g(v_2^{\delta_2}(t))v_2^{\delta_2-1}(t)v_2'(t))dsdt,
\end{aligned}$$

$$I_1 := \int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx = \int_{a_1}^{b_1} (v_1^{\delta_1}(s))^{p(1-\sigma)-1} f^p(v_1^{\delta_1}(s)) v_1^{\delta_1-1}(s) v_1'(s)ds,$$

$$I_2 := \int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy = \int_{a_2}^{b_2} (v_2^{\delta_2}(t))^{q(1-\sigma)-1} g^q(v_2^{\delta_2}(t)) v_2^{\delta_2-1}(t) v_2'(t)dt.$$

Setting

$$F(s) = f(v_1^{\delta_1}(s)) v_1^{\delta_1-1}(s) v_1'(s), \quad G(t) = g(v_2^{\delta_2}(t)) v_2^{\delta_2-1}(t) v_2'(t),$$

we obtain

$$f^p(v_1^{\delta_1}(s)) = v_1^{p(1-\delta_1)}(s)(v_1'(s))^{-p} F^p(s),$$

$$g^q(v_2^{\delta_2}(t)) = v_2^{q(1-\delta_2)}(t)(v_2'(t))^{-q} G^q(t),$$

and then it follows that

$$I = \int_{a_2}^{b_2} \int_{a_1}^{b_1} h(v_1^{\delta_1}(s) v_2^{\delta_2}(t)) F(s) G(t) ds dt,$$

$$I_1 = \int_{a_1}^{b_1} \frac{(v_1(s))^{p(1-\delta_1\sigma)-1}}{(v_1'(s))^{p-1}} F^p(s)ds, \quad I_2 = \int_{a_2}^{b_2} \frac{(v_2(t))^{q(1-\delta_2\sigma)-1}}{(v_2'(t))^{q-1}} G^q(t)dt.$$

Substituting the above results in (18), resetting

$$s = x, \ t = y, \ F(s) = f(x), \ G(t) = g(y),$$

we obtain (22). Similarly, we have (23).

On the other hand, if we set

$$v_1^{\delta_1}(x) = x, \ v_2^{\delta_2}(y) = y, \ a_i = 0, \ b_i = \infty \ (i = 1, 2)$$

in (22), we obtain (18). Hence, the inequalities (22) and (18) are equivalent. It is evident that the inequalities (23) and (19) are equivalent. Hence, the inequalities (22) and (23) are equivalent. Since the constant factors in (18) and (19) are the best possible, it follows that the constant factors in (22) and (23) are also the best possible by using the equivalency.

This completes the proof of the theorem. □

Theorem 4. *Replacing $p > 1$ by $0 < p < 1$ in Theorem 3, we obtain the equivalent reverses of (22) and (23). If there exists a constant $\delta_0 > 0$, such that for any value of $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,*

$$k(\tilde{\sigma}) = \int_0^\infty h(t) t^{\tilde{\sigma}-1} dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (22) and (23) are the best possible.

In particular, for $\delta_1 = \delta_2 = 1$ in Theorems 3 and 4, we get the following integral inequalities with the non-homogeneous kernel:

Corollary 1. *Suppose that* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $h(t) \geq 0, \sigma \in \mathbf{R}$,

$$k(\sigma) = \int_0^\infty h(t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$0 \leq a_i < b_i \leq \infty$, $v_i'(s) > 0 (s \in (a_i, b_i))$, $v_i(a_i^+) = 0$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2}\int_{a_1}^{b_1} h(v_1(x)v_2(y))f(x)g(y)dxdy \\ &< k(\sigma)\left[\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx\right]^{\frac{1}{p}} \\ &\times \left[\int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy\right]^{\frac{1}{q}}, \end{aligned} \tag{24}$$

$$\begin{aligned} &\int_{a_2}^{b_2} \frac{v_2'(y)}{(v_2(y))^{1-p\sigma}} \left(\int_{a_1}^{b_1} h(v_1(x)v_2(y))f(x)dx\right)^p dy \\ &< k^p(\sigma)\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \end{aligned} \tag{25}$$

where the constant factors $k(\sigma)$ *and* $k^p(\sigma)$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (24) and (25).

If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k(\tilde{\sigma}) = \int_0^\infty h(t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (24) and (25) are the best possible.

In particular, for $\delta_1 = -1, \delta_2 = 1$ in Theorems 3 and 4, setting

$$h(t) = k_\lambda(1, t)$$

(cf. Definition 1), we find

$$h\left(\frac{v_2(y)}{v_1(x)}\right) = k_\lambda\left(1, \frac{v_2(y)}{v_1(x)}\right) = v_1^\lambda(x) k_\lambda(v_1(x), v_2(y)).$$

Replacing $f(x)$ by $v_1^{-\lambda}(x) f(x)$, it follows that $[v_1(x)]^{p(1+\sigma)-1} f^p(x)$ is replaced by

$$[v_1(x)]^{p(1+\sigma)-1} [v_1^{-\lambda}(x) f(x)]^p = [v_1(x)]^{p(1-\mu)-1} f^p(x),$$

and we have the following integral inequalities with the homogeneous kernel:

Corollary 2. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1, \mu, \sigma \in \mathbf{R}, \mu + \sigma = \lambda, k_\lambda(x, y)$ *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}_+^2$,

$$k_\lambda(\sigma) = \int_0^\infty k_\lambda(1, t) t^{\sigma-1} dt \in \mathbf{R}_+,$$

$0 \le a_i < b_i \le \infty, v_i'(s) > 0$ $(s \in (a_i, b_i)), v_i(a_i^+) = 0, v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x), g(y) \ge 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y) dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2} \int_{a_1}^{b_1} k_\lambda(v_1(x), v_2(y)) f(x) g(y) dx dy \\ &< k_\lambda(\sigma) \left[\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx\right]^{\frac{1}{p}} \\ &\quad \times \left[\int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y) dy\right]^{\frac{1}{q}}, \end{aligned} \tag{26}$$

$$\int_{a_2}^{b_2} \frac{v_2'(y)}{(v_2(y))^{1-p\sigma}} \left(\int_{a_1}^{b_1} k_\lambda(v_1(x), v_2(y)) f(x) dx \right)^p dy$$
$$< k_\lambda^p(\sigma) \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx, \tag{27}$$

where the constant factors $k_\lambda(\sigma)$ and $k_\lambda^p(\sigma)$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (26) and (27). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k_\lambda(\tilde{\sigma}) = \int_0^\infty k_\lambda(1, t) t^{\tilde{\sigma}-1} dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (26) and (27) are the best possible.

Setting

$$a_i = 0,\ b_i = \infty\ (i = 1, 2),\ v_1(x) = x,\ v_2(y) = y$$

in Corollary 2, we have

Corollary 3. *Suppose that $p > 1, \frac{1}{p} + \frac{1}{q} = 1, \mu, \sigma \in \mathbf{R}, \mu + \sigma = \lambda,\ k_\lambda(x, y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^2$, and*

$$k_\lambda(\sigma) = \int_0^\infty k_\lambda(1, t) t^{\sigma-1} dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_0^\infty x^{p(1-\mu)-1} f^p(x) dx < \infty$$

and

$$0 < \int_0^\infty g^{q(1-\sigma)-1}(y) dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_0^\infty \int_0^\infty k_\lambda(x, y) f(x) g(y) dx dy$$
$$< k_\lambda(\sigma) \left[\int_0^\infty x^{p(1-\mu)-1} f^p(x) dx \right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g^q(y) dy \right]^{\frac{1}{q}}, \tag{28}$$

$$\int_0^\infty y^{p\sigma-1}\left(\int_0^\infty k_\lambda(x,y)f(x)dx\right)^p dy < k_\lambda^p(\sigma)\int_0^\infty x^{p(1-\mu)-1}f^p(x)dx, \tag{29}$$

where the constant factors $k_\lambda(\sigma)$ *and* $k_\lambda^p(\sigma)$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (28) and (29). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k_\lambda(\tilde{\sigma}) = \int_0^\infty k_\lambda(1,t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (28) and (29) are the best possible.

Remark 3. (a) It is evident that (18) and (28) are equivalent for $h(t) = k_\lambda(1,t)$. The same holds for (19) and (29).

(b) In the following, we list the functions $v_i(s)$ $(i = 1,2)$ which satisfy the conditions of Theorems 3 and 4:

(i) $v_i(s) = s^a$, $s \in (0,\infty)$ $(a \in \mathbf{R}_+)$, with $v_i'(s) = as^{a-1} > 0$;
(ii) $v_i(s) = \tan^a s$, $s \in (0, \frac{\pi}{2})$ $(a \in \mathbf{R}_+)$, with $v_i'(s) = a\tan^{a-1} s \sec^2 s > 0$;
(iii) $v_i(s) = \ln^a s$, $s \in (1,\infty)$ $(a \in \mathbf{R}_+)$, with $v_i'(s) = \frac{a}{s}\ln^{a-1} s > 0$;
(iv) $v_i(s) = e^{as} - 1$, $s \in (0,\infty)$ $(a \in \mathbf{R}_+)$, with $v_i'(s) = ae^{as} > 0$.

2.4 Hardy-Type Integral Inequalities with Multi-Variables

In the following two sections, if the constant factors in the inequalities (operator inequalities) are related to $k^{(1)}(\sigma)$ (or $k_\lambda^{(1)}(\sigma)$), then we call them Hardy-type inequalities (operators) of the *first* kind; if the constant factors in the inequalities (operator inequalities) are related to $k^{(2)}(\sigma)$ (or $k_\lambda^{(2)}(\sigma)$), then we call them Hardy-type inequalities (operators) of the *second* kind.

If $h(t) = 0$ $(t > 1)$, then we have $h(xy) = 0$ $(x > \frac{1}{y} > 0)$, and

$$k(\sigma) = \int_0^\infty h(t)t^{\sigma-1}dt = \int_0^1 h(t)t^{\sigma-1}dt.$$

Setting

$$k^{(1)}(\sigma) := \int_0^1 h(t)t^{\sigma-1}dt, \tag{30}$$

by Theorems 1 and 2, we obtain the following first kind Hardy-type integral inequalities with non-homogeneous kernel:

Corollary 4. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1,\ h(t) \geq 0, \sigma \in \mathbf{R}$,

$$k^{(1)}(\sigma) = \int_0^1 h(t)t^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_0^\infty \left(\int_0^{\frac{1}{y}} h(xy)f(x)dx\right) g(y)dy = \int_0^\infty \left(\int_0^{\frac{1}{x}} h(xy)g(y)dy\right) f(x)dx$$

$$< k^{(1)}(\sigma)\left[\int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx\right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy\right]^{\frac{1}{q}}, \tag{31}$$

$$\int_0^\infty y^{p\sigma-1}\left(\int_0^{\frac{1}{y}} h(xy)f(x)dx\right)^p dy < (k^{(1)}(\sigma))^p \int_0^\infty x^{p(1-\sigma)-1} f^p(x)dx; \tag{32}$$

where the constant factors $k^{(1)}(\sigma)$ *and* $(k^{(1)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (31) and (32). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k^{(1)}(\tilde{\sigma}) = \int_0^1 h(t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (31) and (32) are the best possible.

If $h(t) = 0$ $(t > 1)$ in Corollary 1, then

$$h(v_1(x)v_2(y)) = 0 \quad \left(v_1(x) > \frac{1}{v_2(y)} > 0\right),$$

and therefore we obtain the following general results:

Corollary 5. *Suppose that* $p > 1,\ \frac{1}{p} + \frac{1}{q} = 1,\ h(t) \geq 0, \sigma \in \mathbf{R}$,

$$k^{(1)}(\sigma) = \int_0^1 h(t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$0 \leq a_i < b_i \leq \infty,\ v_i'(s) > 0\ (s \in (a_i, b_i)), v_i(a_i^+) = 0, v_i(b_i^-) = \infty\ (i = 1, 2)$. *If* $f(x),\ g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2} \left(\int_{a_1}^{v_1^{-1}(\frac{1}{v_2(y)})} h(v_1(x)v_2(y))f(x)dx \right) g(y)dy \\
&= \int_{a1}^{b_1} \left(\int_{a_2}^{v_2^{-1}(\frac{1}{v_1(x)})} h(v_1(x)v_2(y))g(y)dy \right) f(x)dx \\
&< k^{(1)}(\sigma) \left[\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx \right]^{\frac{1}{p}} \\
&\quad \times \left[\int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy \right]^{\frac{1}{q}}, \qquad (33)
\end{aligned}$$

$$\begin{aligned}
&\int_{a_2}^{b_2} \frac{v_2'(y)}{(v_2(y))^{1-p\sigma}} \left(\int_{a_1}^{v_1^{-1}(\frac{1}{v_2(y)})} h(v_1(x)v_2(y))f(x)dx \right)^p dy \\
&< (k^{(1)}(\sigma))^p \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \qquad (34)
\end{aligned}$$

where the constant factors $k^{(1)}(\sigma)$ *and* $(k^{(1)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (33) and (34). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k^{(1)}(\tilde{\sigma}) = \int_0^1 h(t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (33) and (34) are the best possible.

If $h(t) = 0$ $(0 < t < 1)$, then $h(xy) = 0$ $(0 < x < \frac{1}{y})$, and

$$k(\sigma) = \int_0^\infty h(t)t^{\sigma-1}dt = \int_1^\infty h(t)t^{\sigma-1}dt.$$

Setting

$$k^{(2)}(\sigma) := \int_1^\infty h(t)t^{\sigma-1}dt, \tag{35}$$

by Theorems 1 and 2, we have the following second kind Hardy-type integral inequalities with the non-homogeneous kernel:

Corollary 6. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1, h(t) \geq 0, \sigma \in \mathbf{R}$,

$$k^{(2)}(\sigma) = \int_1^\infty h(t)t^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_0^\infty \left(\int_{\frac{1}{y}}^\infty h(xy)f(x)dx\right)g(y)dy = \int_0^\infty \left(\int_{\frac{1}{x}}^\infty h(xy)g(y)dy\right)f(x)dx$$
$$< k^{(2)}(\sigma)\left[\int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \tag{36}$$

$$\int_0^\infty y^{p\sigma-1}\left(\int_{\frac{1}{y}}^\infty h(xy)f(x)dx\right)^p dy < (k^{(2)}(\sigma))^p \int_0^\infty x^{p(1-\sigma)-1}f^p(x)dx, \tag{37}$$

where the constant factors $k^{(2)}(\sigma)$ *and* $(k^{(2)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we derive the equivalent reverses of (36) and (37).

If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k^{(2)}(\tilde{\sigma}) = \int_1^\infty h(t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (36) and (37) are the best possible.

If $h(t) = 0$ $(0 < t < 1)$ in Corollary 1, then

$$h(v_1(x)v_2(y)) = 0 \left(0 < v_1(x) < \frac{1}{v_2(y)}\right).$$

We have the following general results:

Corollary 7. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1, h(t) \geq 0, \sigma \in \mathbf{R}$,

$$k^{(2)}(\sigma) = \int_1^\infty h(t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$0 \leq a_i < b_i \leq \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = 0$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2} \left(\int_{v_1^{-1}(\frac{1}{v_2(y)})}^{b_1} h(v_1(x)v_2(y))f(x)dx \right) g(y)dy \\
&= \int_{a1}^{b_1} \left(\int_{v_2^{-1}(\frac{1}{v_1(x)})}^{b_2} h(v_1(x)v_2(y))g(y)dy \right) f(x)dx \\
&< k^{(2)}(\sigma) \left[\int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx \right]^{\frac{1}{p}} \\
&\quad \times \left[\int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy \right]^{\frac{1}{q}}, \qquad (38)
\end{aligned}$$

$$\int_{a_2}^{b_2} \frac{v_2'(y)}{(v_2(y))^{1-p\sigma}} \left(\int_{v_1^{-1}(\frac{1}{v_2(y)})}^{b_1} h(v_1(x)v_2(y)) f(x)dx \right)^p dy$$
$$< (k^{(2)}(\sigma))^p \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \tag{39}$$

where the constant factors $k^{(2)}(\sigma)$ and $(k^{(2)}(\sigma))^p$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we have the equivalent reverses of (38) and (39). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k^{(2)}(\tilde{\sigma}) = \int_1^\infty h(t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (38) and (39) are the best possible.

Similarly, if $k_\lambda(1,t) = 0$ $(t > 1)$, then

$$k_\lambda(x,y) = x^{-\lambda} k_\lambda\left(1, \frac{y}{x}\right) = 0 \ (y > x > 0),$$

by Corollary 3, we have the following First kind of Hardy-type integral inequalities with the homogeneous kernel:

Corollary 8. *Suppose that $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $k_\lambda(x,y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^2$,*

$$k_\lambda^{(1)}(\sigma) = \int_0^1 k_\lambda(1,t)t^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_0^\infty x^{p(1-\mu)-1} f^p(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_0^\infty \left(\int_y^\infty k_\lambda(x,y) f(x)dx \right) g(y)dy = \int_0^\infty \left(\int_x^\infty k_\lambda(x,y) g(y)dy \right) f(x)dx$$
$$< k_\lambda^{(1)}(\sigma) \left[\int_0^\infty x^{p(1-\mu)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \tag{40}$$

$$\int_0^\infty y^{p\sigma-1}\left(\int_y^\infty k_\lambda(x,y)f(x)dx\right)^p dy < (k_\lambda^{(1)}(\sigma))^p \int_0^\infty x^{p(1-\mu)-1}f^p(x)dx, \tag{41}$$

where the constant factors $k_\lambda^{(1)}(\sigma)$ *and* $(k_\lambda^{(1)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above cases, we obtain the equivalent reverses of (28) and (29). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k_\lambda^{(1)}(\tilde{\sigma}) = \int_0^1 k_\lambda(1,t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (40) and (41) are the best possible.

If $k_\lambda(1,t) = 0$ $(t > 1)$ in Corollary 2, then

$$k_\lambda(v_1(x), v_2(y)) = 0 \ (0 < v_1(x) < v_2(y)),$$

and we have the following general results:

Corollary 9. *Suppose that* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $k_\lambda(x,y)$ *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}_+^2$,

$$k_\lambda^{(1)}(\sigma) = \int_0^1 k_\lambda(1,t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$0 \le a_i < b_i \le \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = 0$, $v_i(b_i^-) = \infty$ $(i = 1,2)$. *If* $f(x), g(y) \ge 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{a_2}^{b_2}\left(\int_{v_1^{-1}(v_2(y))}^{b_1} k_\lambda(v_1(x), v_2(y))f(x)dx\right)g(y)dy$$
$$= \int_{a_1}^{b_1}\left(\int_{v_2^{-1}(v_1(x))}^{b_2} k_\lambda(v_1(x), v_2(y))g(y)dy\right)f(x)dx$$

$$< k_\lambda^{(1)}(\sigma)\left[\int_{a_1}^{b_1}\frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx\right]^{\frac{1}{p}}$$
$$\times\left[\int_{a_2}^{b_2}\frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}}g^q(y)dy\right]^{\frac{1}{q}}, \tag{42}$$

$$\int_{a_2}^{b_2}\frac{v_2'(y)}{(v_2(y))^{1-p\sigma}}\left(\int_{v_1^{-1}(v_2(y))}^{b_1}k_\lambda(v_1(x),v_2(y))f(x)dx\right)^p dy$$
$$< (k_\lambda^{(1)}(\sigma))^p\int_{a_1}^{b_1}\frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx, \tag{43}$$

where the constant factors $k_\lambda^{(1)}(\sigma)$ *and* $(k_\lambda^{(1)}(\sigma))^p$ *are the best possible.*

Replacing $p>1$ by $0<p<1$ in the above cases, we have the equivalent reverses of (42) and (43). If there exists a constant $\delta_0>0$, such that for any $\tilde{\sigma}\in(\sigma-\delta_0,\sigma]$,

$$k_\lambda^{(1)}(\tilde{\sigma})=\int_0^1 k_\lambda(1,t)t^{\tilde{\sigma}-1}dt\in\mathbf{R}_+,$$

then the constant factors in the reverses of (42) and (43) are the best possible.

Similarly, if $k_\lambda(1,t)=0$ $(0<t<1)$ in Corollary 3, then

$$k_\lambda(x,y)=0\ (x>y>0),$$

and we have the following second kind Hardy-type integral inequalities with the homogeneous kernel:

Corollary 10. *Suppose that* $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, $\mu,\sigma\in\mathbf{R}$, $\mu+\sigma=\lambda$, $k_\lambda(x,y)$ *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}_+^2$,

$$k_\lambda^{(2)}(\sigma)=\int_1^\infty k_\lambda(1,t)t^{\sigma-1}dt\in\mathbf{R}_+.$$

If $f(x),g(y)\geq 0$,

$$0<\int_0^\infty x^{p(1-\mu)-1}f^p(x)dx<\infty$$

and

$$0<\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy<\infty,$$

then we have the following equivalent inequalities:

$$\int_0^\infty \left(\int_0^y k_\lambda(x,y)f(x)dx\right) g(y)dy = \int_0^\infty \left(\int_0^x k_\lambda(x,y)g(y)dy\right) f(x)dx$$
$$< k_\lambda^{(2)}(\sigma)\left[\int_0^\infty x^{p(1-\mu)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_0^\infty y^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \tag{44}$$

$$\int_0^\infty y^{p\sigma-1}\left(\int_0^y k_\lambda(x,y)f(x)dx\right)^p dy < (k_\lambda^{(2)}(\sigma))^p \int_0^\infty x^{p(1-\mu)-1}f^p(x)dx; \tag{45}$$

where the constant factors $k_\lambda^{(2)}(\sigma)$ and $(k_\lambda^{(2)}(\sigma))^p$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above cases, we have the equivalent reverses of (44) and (45). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$k_\lambda^{(2)}(\tilde{\sigma}) = \int_1^\infty k_\lambda(1,t)t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (44) and (45) are the best possible.

If $k_\lambda(1,t) = 0$ $(0 < t < 1)$ in Corollary 2, then

$$k_\lambda(v_1(x), v_2(y)) = 0 \ (v_1(x) > v_2(y) > 0),$$

we have the following general results:

Corollary 11. *Suppose that $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $k_\lambda(x,y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}_+^2$,*

$$k_\lambda^{(2)}(\sigma) = \int_1^\infty k_\lambda(1,t)t^{\sigma-1}dt \in \mathbf{R}_+,$$

$0 \le a_i < b_i \le \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = 0$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. If $f(x)$, $g(y) \ge 0$, such that

$$0 < \int_{a_1}^{b_1} \frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2}\left(\int_{a_1}^{v_1^{-1}(v_2(y))} k_\lambda(v_1(x),v_2(y))f(x)dx\right)g(y)dy \\
&= \int_{a_1}^{b_1}\left(\int_{a_2}^{v_2^{-1}(v_1(x))} k_\lambda(v_1(x),v_2(y))g(y)dy\right)f(x)dx \\
&< k_\lambda^{(2)}(\sigma)\left[\int_{a_1}^{b_1}\frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx\right]^{\frac{1}{p}} \\
&\quad\times\left[\int_{a_2}^{b_2}\frac{(v_2(y))^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}}g^q(y)dy\right]^{\frac{1}{q}}, \qquad (46)
\end{aligned}$$

$$\begin{aligned}
&\int_{a_2}^{b_2}\frac{v_2'(y)}{(v_2(y))^{1-p\sigma}}\left(\int_{a_1}^{v_1^{-1}(v_2(y))} k_\lambda(v_1(x),v_2(y))f(x)dx\right)^p dy \\
&< (k_\lambda^{(2)}(\sigma))^p\int_{a_1}^{b_1}\frac{(v_1(x))^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx, \qquad (47)
\end{aligned}$$

where the constant factors $k_\lambda^{(2)}(\sigma)$ and $(k_\lambda^{(2)}(\sigma))^p$ are the best possible.

Replacing $p>1$ by $0<p<1$ in the above inequalities, we have the equivalent reverses of (46) and (47). If there exists a constant $\delta_0>0$, such that for any $\tilde{\sigma}\in(\sigma-\delta_0,\sigma]$,

$$k_\lambda^{(2)}(\tilde{\sigma})=\int_1^\infty k_\lambda(1,t)t^{\tilde{\sigma}-1}dt\in\mathbf{R}_+,$$

then the constant factors in the reverses of (46) and (47) are the best possible.

2.5 Yang–Hilbert-Type Operators and Hardy-Type Operators

Suppose that $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, $\mu,\sigma\in\mathbf{R}$, $\mu+\sigma=\lambda$. We set the following functions:

$$\varphi(x):=x^{p(1-\sigma)-1},\ \psi(y):=y^{q(1-\sigma)-1},\ \phi(x):=x^{p(1-\mu)-1}(x,y\in\mathbf{R}_+),$$

from which we obtain that $\psi^{1-p}(y)=y^{p\sigma-1}$.

Define the following real normed linear space:

$$L_{p,\varphi}(\mathbf{R}_+) := \left\{ f : ||f||_{p,\varphi} := \left\{ \int_0^\infty \varphi(x)|f(x)|^p dx \right\}^{\frac{1}{p}} < \infty \right\}.$$

Therefore,

$$L_{p,\psi^{1-p}}(\mathbf{R}_+) = \left\{ h : ||h||_{p,\psi^{1-p}} := \left\{ \int_0^\infty \psi^{1-p}(y)|h(y)|^p dy \right\}^{\frac{1}{p}} < \infty \right\},$$

$$L_{p,\phi}(\mathbf{R}_+) = \left\{ g : ||g||_{p,\phi} := \left\{ \int_0^\infty \phi(x)|g(x)|^p dx \right\}^{\frac{1}{p}} < \infty \right\}.$$

(a) In view of Theorem 1, for $f \in L_{p,\varphi}(\mathbf{R}_+)$, setting

$$H_1(y) := \int_0^\infty h(xy)|f(x)|dx \ \ (y \in \mathbf{R}_+),$$

by (19), we have

$$||H_1||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y) H_1^p(y) dy \right)^{\frac{1}{p}} < k(\sigma)||f||_{p,\varphi} < \infty. \tag{48}$$

Definition 3. Let us define the Yang–Hilbert-type integral operator with the non-homogeneous kernel $T_1 : L_{p,\varphi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$ as follows:

For any $f \in L_{p,\varphi}(\mathbf{R}_+)$, there exists a unique representation $T_1 f = H_1 \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_1 f(y) = H_1(y),$$

for any $y \in \mathbf{R}_+$.

In view of (48), it follows that

$$||T_1 f||_{p,\psi^{1-p}} = ||H_1||_{p,\psi^{1-p}} \le k(\sigma)||f||_{p,\varphi}$$

and then the operator T_1 is bounded satisfying

$$||T_1|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R}_+)} \frac{||T_1 f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \le k(\sigma).$$

Since the constant factor $k(\sigma)$ in (48) is the best possible, we have

$$||T_1|| = k(\sigma) = \int_0^\infty h(t) t^{\sigma-1} dt. \tag{49}$$

If we define the formal inner product of $T_1 f$ and g as

$$(T_1 f, g) := \int_0^\infty \left(\int_0^\infty h(xy) f(x) dx \right) g(y) dy = \int_0^\infty \int_0^\infty h(xy) f(x) g(y) dx dy,$$

then we can rewrite (18) and (19) as follows:

$$(T_1 f, g) < ||T_1|| \cdot ||f||_{p,\varphi} ||g||_{q,\psi}, \quad ||T_1 f||_{p,\psi^{1-p}} < ||T_1|| \cdot ||f||_{p,\varphi}.$$

(b) In view of Corollary 4, for $f \in L_{p,\varphi}(\mathbf{R}_+)$, setting

$$H_1^{(1)}(y) := \int_0^{\frac{1}{y}} h(xy) |f(x)| dx \quad (y \in \mathbf{R}_+),$$

by (32), we obtain

$$||H_1^{(1)}||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y) (H_1^{(1)}(y))^p dy \right)^{\frac{1}{p}} < k^{(1)}(\sigma) ||f||_{p,\varphi} < \infty. \tag{50}$$

Definition 4. Define the Hardy-type integral operator of the first kind with the non-homogeneous kernel

$$T_1^{(1)} : L_{p,\varphi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$$

as follows:

For any $f \in L_{p,\varphi}(\mathbf{R}_+)$, there exists a unique representation $T_1^{(1)} f = H_1^{(1)} \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_1^{(1)} f(y) = H_1^{(1)}(y),$$

for any $y \in \mathbf{R}_+$.

In view of (50), it follows that

$$||T_1^{(1)} f||_{p,\psi^{1-p}} = ||H_1^{(1)}||_{p,\psi^{1-p}} \le k^{(1)}(\sigma) ||f||_{p,\varphi}$$

and thus the operator $T_1^{(1)}$ is bounded satisfying

$$||T_1^{(1)}|| = \sup_{f(\neq\theta) \in L_{p,\varphi}(\mathbf{R}_+)} \frac{||T_1^{(1)} f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \le k^{(1)}(\sigma).$$

Since the constant factor $k^{(1)}(\sigma)$ in (50) is the best possible, we have

$$||T_1^{(1)}|| = k^{(1)}(\sigma) = \int_0^1 h(t)t^{\sigma-1}dt. \tag{51}$$

Setting the formal inner product of $T_1^{(1)}f$ and g as

$$(T_1^{(1)}f, g) = \int_0^\infty \left(\int_0^{\frac{1}{y}} h(xy)f(x)dx\right) g(y)dy,$$

we can rewrite (31) and (32) as follows:

$$(T_1^{(1)}f, g) < ||T_1^{(1)}|| \cdot ||f||_{p,\varphi}||g||_{q,\psi}, \quad ||T_1^{(1)}f||_{p,\psi^{1-p}} < ||T_1^{(1)}|| \cdot ||f||_{p,\varphi}. \tag{52}$$

(c) In view of Corollary 6, for $f \in L_{p,\varphi}(\mathbf{R}_+)$, setting

$$H_1^{(2)}(y) := \int_{\frac{1}{y}}^\infty h(xy)|f(x)|dx \ \ (y \in \mathbf{R}_+),$$

by (37), we have

$$||H_1^{(2)}||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y)(H_1^{(2)}(y))^p dy\right)^{\frac{1}{p}} < k^{(2)}(\sigma)||f||_{p,\varphi} < \infty. \tag{53}$$

Definition 5. Define the second kind Hardy-type integral operator with the non-homogeneous kernel

$$T_1^{(2)} : L_{p,\varphi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$$

as follows:

For any $f \in L_{p,\varphi}(\mathbf{R}_+)$, there exists a unique representation $T_1^{(2)}f = H_1^{(2)} \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_1^{(2)}f(y) = H_1^{(2)}(y),$$

for any $y \in \mathbf{R}_+$.

In view of (37), it follows that

$$||T_1^{(2)}f||_{p,\psi^{1-p}} = ||H_1^{(2)}||_{p,\psi^{1-p}} \le k^{(2)}(\sigma)||f||_{p,\varphi}$$

and hence the operator $T_1^{(2)}$ is bounded satisfying

$$||T_1^{(2)}|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R}_+)} \frac{||T_1^{(2)}f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \leq k^{(2)}(\sigma).$$

Since the constant factor $k^{(2)}(\sigma)$ in (53) is the best possible, we have

$$||T_1^{(2)}|| = k^{(2)}(\sigma) = \int_1^\infty h(t)t^{\sigma-1}dt. \tag{54}$$

Setting the formal inner product of $T_1^{(2)}f$ and g as

$$(T_1^{(2)}f, g) = \int_0^\infty \left(\int_{\frac{1}{y}}^\infty h(xy)f(x)dx\right) g(y)dy,$$

we can rewrite (36) and (37) as follows:

$$(T_1^{(2)}f, g) < ||T_1^{(2)}|| \cdot ||f||_{p,\varphi}||g||_{q,\psi}, \quad ||T_1^{(2)}f||_{p,\psi^{1-p}} < ||T_1^{(2)}|| \cdot ||f||_{p,\varphi}. \tag{55}$$

(d) In view of Corollary 3, for $f \in L_{p,\phi}(\mathbf{R}_+)$, setting

$$H_2(y) := \int_0^\infty k_\lambda(x, y)|f(x)|dx \quad (y \in \mathbf{R}_+),$$

by (29), we have

$$||H_2||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y)H_2^p(y)dy\right)^{\frac{1}{p}} < k_\lambda(\sigma)||f||_{p,\phi} < \infty. \tag{56}$$

Definition 6. Define the Yang–Hilbert-type integral operator with the homogeneous kernel $T_2 : L_{p,\phi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$ as follows:

For any $f \in L_{p,\phi}(\mathbf{R}_+)$, there exists a unique representation $T_2f = H_2 \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_2f(y) = H_2(y),$$

for any $y \in \mathbf{R}_+$.

In view of (56), it follows that

$$||T_2f||_{p,\psi^{1-p}} = ||H_2||_{p,\psi^{1-p}} \leq k_\lambda(\sigma)||f||_{p,\phi}$$

and thus the operator T_2 is bounded satisfying

$$||T_2|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R}_+)} \frac{||T_2 f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \leq k_\lambda(\sigma).$$

Since the constant factor $k_\lambda(\sigma)$ in (56) is the best possible, we have

$$||T_2|| = k_\lambda(\sigma) = \int_0^\infty k_\lambda(1,t)t^{\sigma-1}dt. \tag{57}$$

Setting the formal inner product of $T_2 f$ and g as

$$\begin{aligned}(T_2 f, g) &= \int_0^\infty \left(\int_0^\infty k_\lambda(x,y)f(x)dx\right) g(y)dy \\ &= \int_0^\infty \int_0^\infty k_\lambda(x,y)f(x)g(y)dxdy,\end{aligned}$$

we can rewrite (28) and (29) as follows:

$$(T_2 f, g) < ||T_2|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \quad ||T_2 f||_{p,\psi^{1-p}} < ||T_2|| \cdot ||f||_{p,\phi}.$$

(e) Due to Corollary 8, for $f \in L_{p,\phi}(\mathbf{R}_+)$,

$$H_2^{(1)}(y) := \int_y^\infty k_\lambda(x,y)|f(x)|dx \ \ (y \in \mathbf{R}_+),$$

by (41), we have

$$||H_2^{(1)}||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y)(H_2^{(1)}(y))^p dy\right)^{\frac{1}{p}} < k_\lambda^{(1)}(\sigma)||f||_{p,\phi} < \infty. \tag{58}$$

Definition 7. Define the Hardy-type integral operator of the first kind, with the homogeneous kernel $T_2^{(1)} : L_{p,\phi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$ as follows:

For any $f \in L_{p,\phi}(\mathbf{R}_+)$, there exists a unique representation $T_2^{(1)} f = H_2^{(1)} \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_2^{(1)} f(y) = H_2^{(1)}(y),$$

for any $y \in \mathbf{R}_+$.

By (41), it follows that

$$||T_2^{(1)} f||_{p,\psi^{1-p}} = ||H_2^{(1)}||_{p,\psi^{1-p}} \leq k_\lambda^{(1)}(\sigma)||f||_{p,\phi}$$

and then the operator $T_2^{(1)}$ is bounded satisfying

$$||T_2^{(1)}|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R}_+)} \frac{||T_2^{(1)}f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \leq k_\lambda^{(1)}(\sigma).$$

Since the constant factor $k_\lambda^{(1)}(\sigma)$ in (58) is the best possible, we have

$$||T_2^{(1)}|| = k_\lambda^{(1)}(\sigma) = \int_0^1 k_\lambda(1,t)t^{\sigma-1}dt. \tag{59}$$

Setting the formal inner product of $T_2^{(1)}f$ and g as

$$(T_2^{(1)}f, g) = \int_0^\infty \left(\int_y^\infty k_\lambda(x,y)f(x)dx\right) g(y)dy,$$

we can rewrite (40) and (41) as follows:

$$(T_2^{(1)}f, g) < ||T_2^{(1)}|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \quad ||T_2^{(1)}f||_{p,\psi^{1-p}} < ||T_2^{(1)}|| \cdot ||f||_{p,\phi}.$$

(f) By Corollary 10, for $f \in L_{p,\phi}(\mathbf{R}_+)$,

$$H_2^{(2)}(y) := \int_0^y k_\lambda(x,y)|f(x)|dx \quad (y \in \mathbf{R}_+),$$

and by (45), we have

$$||H_2^{(2)}||_{p,\psi^{1-p}} := \left(\int_0^\infty \psi^{1-p}(y)(H_2^{(2)}(y))^p dy\right)^{\frac{1}{p}} < k_\lambda^{(2)}(\sigma)||f||_{p,\phi} < \infty. \tag{60}$$

Definition 8. Define the Hardy-type integral operator of the second kind with the homogeneous kernel $T_2^{(2)} : L_{p,\phi}(\mathbf{R}_+) \to L_{p,\psi^{1-p}}(\mathbf{R}_+)$ as follows:

For any $f \in L_{p,\phi}(\mathbf{R}_+)$, there exists a unique representation $T_2^{(2)}f = H_2^{(2)} \in L_{p,\psi^{1-p}}(\mathbf{R}_+)$, satisfying

$$T_2^{(2)}f(y) = H_2^{(2)}(y),$$

for any $y \in \mathbf{R}_+$.

In view of (45), it follows that

$$||T_2^{(2)}f||_{p,\psi^{1-p}} = ||H_2^{(2)}||_{p,\psi^{1-p}} \leq k_\lambda^{(2)}(\sigma)||f||_{p,\phi}$$

and then the operator $T_2^{(2)}$ is bounded satisfying

$$||T_2^{(2)}|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R}_+)} \frac{||T_2^{(2)}f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \leq k_\lambda^{(2)}(\sigma).$$

Since the constant factor $k_\lambda^{(2)}(\sigma)$ in (60) is the best possible, we have

$$||T_2^{(2)}|| = k_\lambda^{(2)}(\sigma) = \int_1^\infty k_\lambda(1,t)t^{\sigma-1}dt. \tag{61}$$

Setting the formal inner product of $T_2^{(2)}f$ and g as

$$(T_2^{(2)}f, g) = \int_0^\infty \left(\int_0^y k_\lambda(x,y)f(x)dx\right) g(y)dy,$$

we can rewrite (44) and (45) as follows:

$$(T_2^{(2)}f, g) < ||T_2^{(2)}|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, ||T_2^{(2)}f||_{p,\psi^{1-p}} < ||T_2^{(2)}|| \cdot ||f||_{p,\phi}.$$

2.6 Some Examples

Example 1. (a) Set

$$h(t) = k_\lambda(1,t) = \frac{1}{(1+t)^\lambda} \ (\mu,\sigma > 0, \mu+\sigma = \lambda).$$

Then we have the kernels

$$h(xy) = \frac{1}{(1+xy)^\lambda}, k_\lambda(x,y) = \frac{1}{(x+y)^\lambda}$$

and obtain the constant factors

$$k(\sigma) = k_\lambda(\sigma) = \int_0^\infty \frac{t^{\sigma-1}}{(1+t)^\lambda}dt = B(\mu,\sigma) \in \mathbf{R}_+.$$

By (49) and (57), we have $||T_1|| = ||T_2|| = B(\mu,\sigma)$.

(b) Set

$$h(t) = k_\lambda(1,t) = \frac{-\ln t}{1-t^\lambda} \ (\mu,\sigma > 0, \mu+\sigma = \lambda).$$

Then we have the kernels

$$h(xy) = \frac{\ln(xy)}{(xy)^{\lambda} - 1}, \quad k_{\lambda}(x, y) = \frac{\ln(x/y)}{x^{\lambda} - y^{\lambda}}$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_{\lambda}(\sigma) &= \int_0^{\infty} \frac{(\ln t) t^{\sigma-1}}{t^{\lambda} - 1} dt \\ &= \frac{1}{\lambda^2} \int_0^{\infty} \frac{(\ln u) u^{(\sigma/\lambda)-1}}{u - 1} du = \left[\frac{\pi}{\lambda \sin \pi(\sigma/\lambda)} \right]^2 \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \left[\frac{\pi}{\lambda \sin \pi(\sigma/\lambda)} \right]^2.$$

(c) Set

$$h(t) = k_{\lambda}(1, t) = \frac{|\ln t|^{\beta}}{(\max\{1, t\})^{\lambda}} \quad (\beta > -1, \mu, \sigma > 0, \mu + \sigma = \lambda).$$

Then we have the kernels

$$h(xy) = \frac{|\ln(xy)|^{\beta}}{(\max\{1, xy\})^{\lambda}}, \quad k_{\lambda}(x, y) = \frac{|\ln(x/y)|^{\beta}}{(\max\{x, y\})^{\lambda}}$$

and by using the formula (cf. [9])

$$\Gamma(\alpha) := \int_0^{\infty} e^{-t} t^{\alpha-1} dt \quad (\alpha > 0)$$

we obtain the following constant factors

$$\begin{aligned} k(\sigma) = k_{\lambda}(\sigma) &= \int_0^{\infty} \frac{|\ln t|^{\beta} t^{\sigma-1}}{(\max\{1, t\})^{\lambda}} dt \\ &= \int_0^1 (-\ln t)^{\beta} t^{\sigma-1} dt + \int_1^{\infty} \frac{(\ln t)^{\beta} t^{\sigma-1}}{t^{\lambda}} dt \\ &= \int_0^1 (-\ln t)^{\beta} (t^{\sigma-1} + t^{\mu-1}) dt = \left(\frac{1}{\sigma^{\beta+1}} \frac{1}{\mu^{\beta+1}} \right) \int_0^{\infty} v^{\beta} e^{-v} dv \\ &= \left(\frac{1}{\sigma^{\beta+1}} + \frac{1}{\mu^{\beta+1}} \right) \Gamma(\beta + 1) \in \mathbf{R}_+. \end{aligned}$$

By (49) and (57), we have

$$||T_1|| = ||T_2|| = \left(\frac{1}{\sigma^{\beta+1}} + \frac{1}{\mu^{\beta+1}}\right)\Gamma(\beta+1).$$

Due to (51) and (59), we have

$$||T_1^{(1)}|| = ||T_2^{(1)}|| = \frac{1}{\sigma^{\beta+1}}\Gamma(\beta+1),$$

and by (54) and (61), it follows that

$$||T_1^{(2)}|| = ||T_2^{(2)}|| = \frac{1}{\mu^{\beta+1}}\Gamma(\beta+1).$$

(d) Set

$$h(t) = k_\lambda(1,t) = \frac{|\ln t|^\beta}{(\min\{1,t\})^\lambda} \quad (\beta > -1, \mu, \sigma < 0, \mu + \sigma = \lambda).$$

Then we have the kernels

$$h(xy) = \frac{|\ln(xy)|^\beta}{(\min\{1,xy\})^\lambda}, \quad k_\lambda(x,y) = \frac{|\ln(x/y)|^\beta}{(\min\{x,y\})^\lambda}$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) &= k_\lambda(\sigma) = \int_0^\infty \frac{|\ln t|^\beta t^{\sigma-1}}{(\min\{1,t\})^\lambda} dt \\ &= \int_0^1 \frac{(-\ln t)^\beta t^{\sigma-1}}{t^\lambda} dt + \int_1^\infty (\ln t)^\beta t^{\sigma-1} dt \\ &= \int_0^1 (-\ln t)^\beta (t^{-\mu-1} + t^{-\sigma-1}) dt = \left[\frac{1}{(-\mu)^{\beta+1}} \frac{1}{(-\sigma)^{\beta+1}}\right] \int_0^\infty v^\beta e^{-v} dv \\ &= \left[\frac{1}{(-\mu)^{\beta+1}} + \frac{1}{(-\sigma)^{\beta+1}}\right] \Gamma(\beta+1) \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \left[\frac{1}{(-\mu)^{\beta+1}} + \frac{1}{(-\sigma)^{\beta+1}}\right] \Gamma(\beta+1).$$

By (51) and (59), we have

$$||T_1^{(1)}|| = ||T_2^{(1)}|| = \frac{1}{(-\mu)^{\beta+1}}\Gamma(\beta+1),$$

and by (54) and (61), it follows that

$$||T_1^{(2)}|| = ||T_2^{(2)}|| = \frac{1}{(-\sigma)^{\beta+1}}\Gamma(\beta+1).$$

(e) Set

$$h(t) = k_\lambda(1,t) = \frac{|\ln t|}{1+t^\lambda} \quad (\mu,\sigma>0, \mu+\sigma=\lambda).$$

Then we have the kernels

$$h(xy) = \frac{|\ln(xy)|}{1+(xy)^\lambda}, \quad k_\lambda(x,y) = \frac{|\ln(x/y)|}{x^\lambda+y^\lambda}$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_\lambda(\sigma) &= \int_0^\infty \frac{|\ln t| t^{\sigma-1}}{1+t^\lambda}dt \\ &= \int_0^1 \frac{(-\ln t)t^{\sigma-1}}{t^\lambda+1}dt + \int_1^\infty \frac{(\ln t)t^{\sigma-1}}{t^\lambda+1}dt = \int_0^1 \frac{(-\ln t)(t^{\sigma-1}+t^{\mu-1})}{t^\lambda+1}dt \\ &= \int_0^1 (-\ln t)\sum_{k=0}^\infty (-1)^k (t^{k\lambda+\sigma-1}+t^{k\lambda+\mu-1})dt. \end{aligned}$$

By the fact that

$$\begin{aligned} &\sum_{k=0}^\infty \int_0^1 |(-1)^k(-\ln t)(t^{k\lambda+\sigma-1}+t^{k\lambda+\mu-1})|dt \\ &\quad = \sum_{k=0}^\infty \int_0^1 (-\ln t)\left(\frac{1}{k\lambda+\sigma}dt^{k\lambda+\sigma} + \frac{1}{k\lambda+\mu}dt^{k\lambda+\mu}\right) \\ &\quad = \sum_{k=0}^\infty \left[\frac{1}{(k\lambda+\sigma)^2} + \frac{1}{(k\lambda+\mu)^2}\right] \in \mathbf{R}_+, \end{aligned}$$

in combination with Theorem 7 (cf. [35], Chapter 5), we obtain

$$k(\sigma) = k_\lambda(\sigma) = \int_0^1 (-\ln t) \sum_{k=0}^{\infty} (-1)^k (t^{k\lambda+\sigma-1} + t^{k\lambda+\mu-1}) dt$$
$$= \sum_{k=0}^{\infty} \int_0^1 (-1)^k (-\ln t)(t^{k\lambda+\sigma-1} + t^{k\lambda+\mu-1}) dt$$
$$= \sum_{k=0}^{\infty} (-1)^k \left[\frac{1}{(k\lambda+\sigma)^2} + \frac{1}{(k\lambda+\mu)^2} \right] \in \mathbf{R}_+.$$

By (49) and (57), we have

$$||T_1|| = ||T_2|| = \sum_{k=0}^{\infty} (-1)^k \left[\frac{1}{(k\lambda+\sigma)^2} + \frac{1}{(k\lambda+\mu)^2} \right].$$

By (51) and (59), we have

$$||T_1^{(1)}|| = ||T_2^{(1)}|| = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(k\lambda+\sigma)^2},$$

and by (54) and (61), it follows that

$$||T_1^{(2)}|| = ||T_2^{(2)}|| = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(k\lambda+\mu)^2}.$$

(f) Set

$$h(t) = k_\lambda(t, 1) = \frac{1}{|1-t|^\lambda} \quad (\mu, \sigma > 0, \mu + \sigma = \lambda < 1).$$

Then, we have kernels

$$h(xy) = \frac{1}{|1-xy|^\lambda}, \quad k_\lambda(x, y) = \frac{1}{|x-y|^\lambda}$$

and obtain the constant factors

$$k(\sigma) = k_\lambda(\sigma) = \int_0^\infty \frac{t^{\sigma-1}}{|1-t|^\lambda} dt$$
$$= \int_0^1 \frac{t^{\sigma-1} + t^{\mu-1}}{(1-t)^\lambda} dt = B(1-\lambda, \sigma) + B(1-\lambda, \mu) \in \mathbf{R}_+.$$

In view of (49) and (57), we obtain

$$||T_1|| = ||T_2|| = B(1-\lambda,\sigma) + B(1-\lambda,\mu).$$

By (51) and (59), we have

$$||T_1^{(1)}|| = ||T_2^{(1)}|| = B(1-\lambda,\sigma),$$

and by (54) and (61), it follows that

$$||T_1^{(2)}|| = ||T_2^{(2)}|| = B(1-\lambda,\mu).$$

For (a)–(f), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{|\sigma|}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

(g) Set

$$h(t) = k_\lambda(1,t) = \frac{(\min\{t,1\})^\eta}{(\max\{t,1\})^{\lambda+\eta}} \quad (\eta > -\min\{\mu,\sigma\}, \mu+\sigma = \lambda).$$

Then we have the kernels

$$h(xy) = \frac{(\min\{1,xy\})^\eta}{(\max\{1,xy\})^{\lambda+\eta}}, \quad k_\lambda(x,y) = \frac{(\min\{x,y\})^\eta}{(\max\{x,y\})^{\lambda+\eta}}$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_\lambda(\sigma) &= \int_0^\infty \frac{(\min\{1,t\})^\eta t^{\sigma-1}}{(\max\{1,t\})^{\lambda+\eta}} dt \\ &= \int_0^1 t^\eta t^{\sigma-1} dt + \int_1^\infty \frac{t^{\sigma-1}}{t^{\lambda+\eta}} dt = \frac{1}{\sigma+\eta} + \frac{1}{\mu+\eta} \\ &= \frac{\lambda+2\eta}{(\sigma+\eta)(\mu+\eta)} \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we get

$$||T_1|| = ||T_2|| = \frac{\lambda+2\eta}{(\sigma+\eta)(\mu+\eta)}.$$

By (51) and (59), we have

$$||T_1^{(1)}|| = ||T_2^{(1)}|| = \frac{1}{\sigma+\eta},$$

and by (54) and (61), it follows that

$$||T_1^{(2)}|| = ||T_2^{(2)}|| = \frac{1}{\mu + \eta}.$$

Then we can derive the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{\sigma+\eta}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

In particular, (i) for $\eta = 0$,

$$h(t) = k_\lambda(1, t) = \frac{1}{(\max\{1, t\})^\lambda} \ \ (\mu, \sigma > 0, \mu + \sigma = \lambda),$$

we have

$$h(xy) = \frac{1}{(\max\{1, xy\})^\lambda}, \ \ k_\lambda(x, y) = \frac{1}{(\max\{x, y\})^\lambda}$$

and

$$||T_1|| = ||T_2|| = \frac{\lambda}{\sigma\mu};$$

(ii) for $\eta = -\lambda$,

$$h(t) = k_\lambda(t, 1) = \frac{1}{(\min\{1, t\})^\lambda} \ \ (\mu, \sigma < 0, \mu + \sigma = \lambda),$$

we have

$$h(xy) = \frac{1}{(\min\{1, xy\})^\lambda}, \ \ k_\lambda(x, y) = \frac{1}{(\min\{x, y\})^\lambda}$$

and

$$||T_1|| = ||T_2|| = \frac{-\lambda}{\sigma\mu};$$

(iii) for $\lambda = 0$,

$$h(t) = k_0(1, t) = \left(\frac{\min\{1, t\}}{\max\{1, t\}}\right)^\eta \ (\eta > |\sigma|),$$

we have

$$h(xy) = \left(\frac{\min\{1, xy\}}{\max\{1, xy\}}\right)^\eta, k_\lambda(x, y) = \left(\frac{\min\{x, y\}}{\max\{x, y\}}\right)^\eta$$

and

$$||T_1|| = ||T_2|| = \frac{2\eta}{\eta^2 - \sigma^2}.$$

Example 2. (a) Set

$$h(t) = k_0(1,t) = \ln\left(1 + \frac{\rho}{t^\eta}\right) \quad (\rho > 0, 0 < \sigma < \eta).$$

Then we have the kernels

$$h(xy) = \ln\left[1 + \frac{\rho}{(xy)^\eta}\right], \quad k_0(x,y) = \ln\left[1 + \rho\left(\frac{x}{y}\right)^\eta\right]$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_0(\sigma) &= \int_0^\infty t^{\sigma-1} \ln\left(1 + \frac{\rho}{t^\eta}\right) dt \\ &= \frac{1}{\sigma}\int_0^\infty \ln\left(1 + \frac{\rho}{t^\eta}\right) dt^\sigma \\ &= \frac{1}{\sigma}\left[t^\sigma \ln\left(1 + \frac{\rho}{t^\eta}\right)|_0^\infty - \int_0^\infty t^\sigma d\ln\left(1 + \frac{\rho}{t^\eta}\right)\right] \\ &= \frac{\eta}{\sigma}\int_0^\infty \frac{t^{\sigma-1}}{(t^\eta/\rho) + 1} dt = \frac{\rho^{\sigma/\eta}}{\sigma}\int_0^\infty \frac{u^{(\sigma/\eta)-1}}{u+1} du \\ &= \frac{\rho^{\sigma/\eta}\pi}{\sigma \sin \pi(\sigma/\eta)} \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{\rho^{\sigma/\eta}\pi}{\sigma \sin \pi(\sigma/\eta)}.$$

(b) Set

$$h(t) = k_0(1,t) = \arctan\left(\frac{\rho}{t^\eta}\right) \quad (\rho > 0, 0 < \sigma < \eta).$$

Then we have the kernels

$$h(xy) = \arctan\left(\frac{\rho}{(xy)^\eta}\right), \quad k_0(x,y) = \arctan\rho\left(\frac{x}{y}\right)^\eta$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_0(\sigma) &= \int_0^\infty t^{\sigma-1} \arctan\left(\frac{\rho}{t^\eta}\right) dt \\ &= \frac{1}{\sigma}\int_0^\infty \arctan\left(\frac{\rho}{t^\eta}\right) dt^\sigma \\ &= \frac{1}{\sigma}\left[t^\sigma \arctan\left(\frac{\rho}{t^\eta}\right)|_0^\infty - \int_0^\infty t^\sigma d \arctan\left(\frac{\rho}{t^\eta}\right)\right] \\ &= \frac{\eta}{\sigma\rho}\int_0^\infty \frac{t^{\eta+\sigma-1}}{(t^{2\eta}/\rho^2)+1} dt = \frac{\rho^{\sigma/\eta}}{2\sigma}\int_0^\infty \frac{u^{[(\eta+\sigma)/(2\eta)]-1}}{u+1} du \\ &= \frac{\rho^{\sigma/\eta}\pi}{2\sigma \sin \pi[(\eta+\sigma)/(2\eta)]} = \frac{\rho^{\sigma/\eta}\pi}{2\sigma \cos \pi[\sigma/(2\eta)]} \in \mathbf{R}_+. \end{aligned}$$

By (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{\rho^{\sigma/\eta}\pi}{2\sigma \cos \pi[\sigma/(2\eta)]}.$$

(c) Set

$$h(t) = k_0(1,t) = e^{-\rho t^\eta} \quad (\rho, \sigma, \eta > 0).$$

Then we have the kernels

$$h(xy) = e^{-\rho(xy)^\eta}, \quad k_0(x,y) = e^{-\rho(\frac{y}{x})^\eta}$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_0(\sigma) &= \int_0^\infty t^{\sigma-1} e^{-\rho t^\eta} dt \\ &= \frac{1}{\eta\rho^{\sigma/\eta}}\int_0^\infty e^{-u} u^{\frac{\sigma}{\eta}-1} du = \frac{1}{\eta\rho^{\sigma/\eta}}\Gamma\left(\frac{\sigma}{\eta}\right) \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{1}{\eta\rho^{\sigma/\eta}}\Gamma\left(\frac{\sigma}{\eta}\right).$$

Then for (a)–(c), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{\sigma}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

Example 3. (a) Set

$$h(t) = k_0(1,t) = \operatorname{csc} h(\rho t^{\eta}) = \frac{2}{e^{\rho t^{\eta}} - e^{-\rho t^{\eta}}} \quad (\rho > 0, 0 < \eta < \sigma)$$

where $\operatorname{csc} h(\cdot)$ stands for the hyperbolic cosecant function (cf. [36]). Then we have the kernels

$$h(xy) = \frac{2}{e^{\rho(xy)^{\eta}} - e^{-\rho(xy)^{\eta}}}, \quad k_0(x,y) = \frac{2}{e^{\rho(\frac{y}{x})^{\eta}} - e^{-\rho(\frac{y}{x})^{\eta}}}.$$

By the Lebesgue term by term integration theorem, we obtain the constant factors

$$\begin{aligned}
k(\sigma) = k_0(\sigma) &= \int_0^{\infty} \frac{2t^{\sigma-1}dt}{e^{\rho t^{\eta}} - e^{-\rho t^{\eta}}} = \int_0^{\infty} \frac{2t^{\sigma-1}dt}{e^{\rho t^{\eta}}(1 - e^{-2\rho t^{\eta}})} \\
&= 2\int_0^{\infty} t^{\sigma-1} \sum_{k=0}^{\infty} e^{-(2k+1)\rho t^{\eta}} dt = 2\sum_{k=0}^{\infty} \int_0^{\infty} t^{\sigma-1} e^{-(2k+1)\rho t^{\eta}} dt \\
&= \frac{2}{\eta \rho^{\sigma/\eta}} \sum_{k=0}^{\infty} \frac{1}{(2k+1)^{\sigma/\eta}} \int_0^{\infty} e^{-u} u^{\frac{\sigma}{\eta}-1} du \\
&= \frac{2}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \sum_{k=0}^{\infty} \frac{1}{(2k+1)^{\sigma/\eta}} \\
&= \frac{2}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left[\sum_{k=1}^{\infty} \frac{1}{k^{\sigma/\eta}} - \sum_{k=1}^{\infty} \frac{1}{(2k)^{\sigma/\eta}}\right] \\
&= \frac{2}{\eta \rho^{\sigma/\eta}} \left(1 - \frac{1}{2^{\sigma/\eta}}\right) \Gamma\left(\frac{\sigma}{\eta}\right) \zeta\left(\frac{\sigma}{\eta}\right) \in \mathbf{R}_+,
\end{aligned}$$

where $\zeta\left(\frac{\sigma}{\eta}\right) = \sum_{k=1}^{\infty} \frac{1}{k^{\sigma/\eta}}$ ($\zeta(\cdot)$ is Riemann zeta function). In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{2}{\eta \rho^{\sigma/\eta}} \left(1 - \frac{1}{2^{\sigma/\eta}}\right) \Gamma\left(\frac{\sigma}{\eta}\right) \zeta\left(\frac{\sigma}{\eta}\right).$$

(b) Set

$$\begin{aligned}
h(t) = k_0(1,t) &= e^{-\rho t^{\eta}} \cot h(\rho t^{\eta}) = e^{-\rho t^{\eta}} \frac{e^{\rho t^{\eta}} + e^{-\rho t^{\eta}}}{e^{\rho t^{\eta}} - e^{-\rho t^{\eta}}} \\
&= \frac{1 + e^{-2\rho t^{\eta}}}{e^{\rho t^{\eta}} - e^{-\rho t^{\eta}}} = \frac{e^{-\rho t^{\eta}} + e^{-3\rho t^{\eta}}}{1 - e^{-2\rho t^{\eta}}} \quad (\rho > 0, 0 < \eta < \sigma).
\end{aligned}$$

Let $\cot h(\cdot)$ stand for the hyperbolic cotangent function (cf. [36]). Then we have the kernels

$$h(xy) = \frac{1 + e^{-2\rho(xy)^\eta}}{e^{\rho(xy)^\eta} - e^{-\rho(xy)^\eta}}, \quad k_0(x, y) = \frac{1 + e^{-2\rho(\frac{x}{y})^\eta}}{e^{\rho(\frac{y}{x})^\eta} - e^{-\rho(\frac{y}{x})^\eta}}.$$

By the Lebesgue term by term integration theorem, we obtain the constant factors

$$\begin{aligned} k(\sigma) = k_0(\sigma) &= \int_0^\infty \frac{(e^{-\rho t^\eta} + e^{-3\rho t^\eta})t^{\sigma-1}}{1 - e^{-2\rho t^\eta}} dt \\ &= \int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty (e^{-(2k+1)\rho t^\eta} + e^{-(2k+3)\rho t^\eta}) dt \\ &= \frac{1}{\eta \rho^{\sigma/\eta}} \sum_{k=0}^\infty \left[\frac{1}{(2k+1)^{\sigma/\eta}} + \frac{1}{(2k+3)^{\sigma/\eta}} \right] \int_0^\infty e^{-u} u^{\frac{\sigma}{\eta}-1} du \\ &= \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left[2 \sum_{k=0}^\infty \frac{1}{(2k+1)^{\sigma/\eta}} - 1 \right] \\ &= \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left[\left(2 - \frac{1}{2^{(\sigma/\eta)-1}}\right) \zeta\left(\frac{\sigma}{\eta}\right) - 1 \right] \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left[\left(2 - \frac{1}{2^{(\sigma/\eta)-1}}\right) \zeta\left(\frac{\sigma}{\eta}\right) - 1 \right].$$

Then for (a)–(b), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{\sigma-\eta}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

(c) Set

$$h(t) = k_0(1, t) = \sec h(\rho t^\eta) = \frac{2}{e^{\rho t^\eta} + e^{-\rho t^\eta}} \quad (\rho, \eta, \sigma > 0).$$

Let $\sec h(\cdot)$ stand for the hyperbolic secant function (cf. [36]). Then we have the kernels

$$h(xy) = \frac{2}{e^{\rho(xy)^\eta} + e^{-\rho(xy)^\eta}}, \quad k_0(x, y) = \frac{2}{e^{\rho(\frac{y}{x})^\eta} + e^{-\rho(\frac{y}{x})^\eta}}.$$

By the Lebesgue term by term integration theorem, we obtain the constant factors

$$\begin{aligned}
k(\sigma) = k_0(\sigma) &= \int_0^\infty \frac{2t^{\sigma-1}dt}{e^{\rho t^\eta} + e^{-\rho t^\eta}} = \int_0^\infty \frac{2t^{\sigma-1}dt}{e^{\rho t^\eta}(1 + e^{-2\rho t^\eta})} \\
&= 2\int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty (-1)^k e^{-(2k+1)\rho t^\eta} dt \\
&= 2\int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty [e^{-(4k+1)\rho t^\eta} - e^{-(4k+3)\rho t^\eta}] dt \\
&= 2\sum_{k=0}^\infty \int_0^\infty t^{\sigma-1} [e^{-(4k+1)\rho t^\eta} - e^{-(4k+3)\rho t^\eta}] dt \\
&= 2\sum_{k=0}^\infty (-1)^k \int_0^\infty t^{\sigma-1} e^{-(2k+1)\rho t^\eta} dt \\
&= \frac{2}{\eta\rho^{\sigma/\eta}} \sum_{k=0}^\infty \frac{(-1)^k}{(2k+1)^{\sigma/\eta}} \int_0^\infty e^{-u} u^{\frac{\sigma}{\eta}-1} du \\
&= \frac{2}{\eta\rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \xi\left(\frac{\sigma}{\eta}\right) \in \mathbf{R}_+,
\end{aligned}$$

where

$$\xi(\frac{\sigma}{\eta}) = \sum_{k=0}^\infty \frac{(-1)^k}{(2k+1)^{\sigma/\eta}}.$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{2}{\eta\rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \xi\left(\frac{\sigma}{\eta}\right).$$

(d) Set

$$\begin{aligned}
h(t) = k_0(1,t) &= e^{-\rho t^\eta} \tan h(\rho t^\eta) = e^{-\rho t^\eta} \frac{e^{\rho t^\eta} - e^{-\rho t^\eta}}{e^{\rho t^\eta} + e^{-\rho t^\eta}} \\
&= \frac{1 - e^{-2\rho t^\eta}}{e^{\rho t^\eta} + e^{-\rho t^\eta}} = \frac{e^{-\rho t^\eta} - e^{-3\rho t^\eta}}{1 + e^{-2\rho t^\eta}} (\rho, \eta, \sigma > 0).
\end{aligned}$$

Let $\tan h(\cdot)$ stand for the hyperbolic tangent function (cf. [36]). Then we have the kernels

$$h(xy) = \frac{1 - e^{-2\rho(xy)^\eta}}{e^{\rho(xy)^\eta} + e^{-\rho(xy)^\eta}}, \quad k_0(x, y) = \frac{1 - e^{-2\rho(\frac{y}{x})^\eta}}{e^{\rho(\frac{y}{x})^\eta} + e^{-\rho(\frac{y}{x})^\eta}}.$$

By the Lebesgue term by term integration theorem, we obtain the constant factors

$$\begin{aligned}
k(\sigma) = k_0(\sigma) &= \int_0^\infty \frac{(e^{-\rho t^\eta} - e^{-3\rho t^\eta})t^{\sigma-1}}{1 + e^{-2\rho t^\eta}} dt \\
&= \int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty (-1)^k e^{-(2k+1)\rho t^\eta} dt - \int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty (-1)^k e^{-(2k+3)\rho t^\eta} dt \\
&= \int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty [e^{-(4k+1)\rho t^\eta} - e^{-(4k+3)\rho t^\eta}] dt \\
&\quad - \int_0^\infty t^{\sigma-1} \sum_{k=0}^\infty [e^{-(4k+3)\rho t^\eta} - e^{-(4k+5)\rho t^\eta}] dt \\
&= \sum_{k=0}^\infty \left\{ \int_0^\infty t^{\sigma-1} [e^{-(4k+1)\rho t^\eta} - e^{-(4k+3)\rho t^\eta}] dt \right. \\
&\quad \left. - \int_0^\infty t^{\sigma-1} [e^{-(4k+3)\rho t^\eta} - e^{-(4k+5)\rho t^\eta}] dt \right\} \\
&= \frac{1}{\eta \rho^{\sigma/\eta}} \sum_{k=0}^\infty (-1)^k \left[\frac{1}{(2k+1)^{\sigma/\eta}} - \frac{1}{(2k+3)^{\sigma/\eta}} \right] \int_0^\infty e^{-u} u^{\frac{\sigma}{\eta}-1} du \\
&= \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left[2 \sum_{k=0}^\infty \frac{(-1)^k}{(2k+1)^{\sigma/\eta}} - 1 \right] \\
&= \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left(2\xi\left(\frac{\sigma}{\eta}\right) - 1 \right) \in \mathbf{R}_+.
\end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{1}{\eta \rho^{\sigma/\eta}} \Gamma\left(\frac{\sigma}{\eta}\right) \left(2\xi\left(\frac{\sigma}{\eta}\right) - 1 \right).$$

Then for (c)–(d), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{\sigma}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

Lemma 2. *Let* $\mathbf{C}$ *stand for the set of complex numbers. If* $\mathbf{C}_\infty = \mathbf{C} \cup \{\infty\}$,

$$z_k \in \mathbf{C}\backslash\{z|\operatorname{Re} z \geq 0,\ \operatorname{Im} z = 0\}\ \ (k = 1,\ 2, \ldots,\ n)$$

are different points, the function $f(z)$ *is analytic in* $\mathbf{C}_\infty$ *except for* $z_i,\ i = 1, 2, \ldots, n$, *and* $z = \infty$ *is a zero point of* $f(z)$ *whose order is not less than 1, then for* $\alpha \in \mathbf{R}$, *we have*

$$\int_0^\infty f(x)x^{\alpha-1}dx = \frac{2\pi i}{1 - e^{2\pi\alpha i}} \sum_{k=1}^{n} \operatorname{Re} s[f(z)z^{\alpha-1}, z_k], \tag{62}$$

where $0 < \operatorname{Im} \ln z = \arg z < 2\pi$. *In particular, if* $z_k,\ k = 1, \ldots,\ n$, *are all poles of order 1, setting*

$$\varphi_k(z) = (z - z_k)f(z)\ \ (\varphi_k(z_k) \neq 0),$$

then

$$\int_0^\infty f(x)x^{\alpha-1}dx = \frac{\pi}{\sin \pi\alpha} \sum_{k=1}^{n} (-z_k)^{\alpha-1}\varphi_k(z_k). \tag{63}$$

Proof. In view of the theorem (cf. [37], p. 118), we obtain (62). We have

$$\begin{aligned} 1 - e^{2\pi\alpha i} &= 1 - \cos 2\pi\alpha - i \sin 2\pi\alpha \\ &= -2i \sin \pi\alpha(\cos \pi\alpha + i \sin \pi\alpha) = -2ie^{i\pi\alpha} \sin \pi\alpha. \end{aligned}$$

In particular, since

$$f(z)z^{\alpha-1} = \frac{1}{z - z_k}\varphi_k(z)z^{\alpha-1},$$

it is obvious that

$$\operatorname{Re} s[f(z)z^{\alpha-1}, -a_k] = {z_k}^{\alpha-1}\varphi_k(z_k) = -e^{i\pi\alpha}(-z_k)^{\alpha-1}\varphi_k(z_k).$$

Then by (62), we obtain (63). □

Example 4. (a) Set

$$h(t) = k_\lambda(1, t) = \frac{1}{\prod_{k=1}^{s}(a_k + t^{\lambda/s})}$$

where $s \in \mathbf{N},\ 0 < a_1 < \cdots < a_s,\ \mu,\ \sigma > 0,\ \mu + \sigma = \lambda$. Then we have the kernels

$$h(xy) = \frac{1}{\prod_{k=1}^{s}[a_k + (xy)^{\lambda/s}]},\ \ k_\lambda(x,y) = \frac{1}{\prod_{k=1}^{s}(a_k x^{\lambda/s} + y^{\lambda/s})}.$$

For

$$f(z) = \frac{1}{\prod_{k=1}^{s}(z + a_k)},\ \ z_k = -a_k,$$

by (63), we get

$$\varphi_k(z_k) = (z + a_k)\frac{1}{\prod_{i=1}^{s}(z + a_i)}\Big|_{z=-a_k} = \prod_{j=1(j\neq k)}^{s}\frac{1}{a_j - a_k},$$

and obtain the constant factors

$$\begin{aligned} k(\sigma) = k_\lambda(\sigma) &= \int_0^\infty \frac{t^{\sigma-1}dt}{\prod_{k=1}^{s}(a_k + t^{\lambda/s})} = \frac{s}{\lambda}\int_0^\infty \frac{u^{(s\sigma/\lambda)-1}du}{\prod_{k=1}^{s}(u + a_k)} \\ &= \frac{\pi s}{\lambda \sin \pi(s\sigma/\lambda)}\sum_{k=1}^{s} a_k^{s\sigma/\lambda}\prod_{j=1(j\neq k)}^{s}\frac{1}{a_j - a_k} \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{\pi s}{\lambda \sin \pi(s\sigma/\lambda)}\sum_{k=1}^{s} a_k^{s\sigma/\lambda}\prod_{j=1(j\neq k)}^{s}\frac{1}{a_j - a_k}.$$

In particular, (i) if $s = 1, a_1 = a, h(t) = k_\lambda(1,t) = \frac{1}{a+t^\lambda}(a, \mu, \sigma > 0\ \mu + \sigma = \lambda)$, then we have the kernels $h(xy) = \frac{1}{a+(xy)^\lambda}, k_\lambda(x,y) = \frac{1}{ax^\lambda + y^\lambda}$, and

$$||T_1|| = ||T_2|| = \frac{\pi}{\lambda \sin \pi(\sigma/\lambda)}a^{\frac{\sigma}{\lambda}-1};$$

(ii) if $s = 2,\ a_1 = a,\ a_2 = b$,

$$h(t) = k_\lambda(1,t) = \frac{1}{(a + t^{\lambda/2})(b + t^{\lambda/2})}\ \ (0 < a < b,\ \mu,\ \sigma > 0,\ \mu + \sigma = \lambda),$$

then we have the kernels

$$h(xy) = \frac{1}{[a + (xy)^{\lambda/2}][b + (xy)^{\lambda/2}]},\ \ k_\lambda(x,y) = \frac{1}{(ax^{\lambda/2} + y^{\lambda/2})(ax^{\lambda/2} + y^{\lambda/2})},$$

and

$$||T_1|| = ||T_2|| = \frac{2\pi}{\lambda \sin \pi(2\sigma/\lambda)} \frac{1}{b-a}(a^{\frac{2\sigma}{\lambda}-1} - b^{\frac{2\sigma}{\lambda}-1}).$$

(b) Set

$$h(t) = k_\lambda(1,t) = \frac{1}{t^\lambda + 2ct^{\lambda/2}\cos\gamma + c^2}$$

($c > 0$, $|\gamma| < \frac{\pi}{2}$, μ, $\sigma > 0$, μ, $\sigma = \lambda$). Then we have the kernels

$$h(xy) = \frac{1}{(xy)^\lambda + 2c(xy)^{\lambda/2}\cos\gamma + c^2},$$

$$k_\lambda(x,y) = \frac{1}{y^\lambda + 2c(xy)^{\lambda/2}\cos\gamma + c^2x^\lambda}.$$

By (63), we can find

$$\begin{aligned} k(\sigma) = k_\lambda(\sigma) &= \int_0^\infty \frac{t^{\sigma-1}}{t^\lambda + 2ct^{\lambda/2}\cos\gamma + c^2}dt \\ &= \frac{2}{\lambda}\int_0^\infty \frac{u^{(2\sigma/\lambda)-1}du}{u^2 + 2cu\cos\gamma + c^2} = \frac{2}{\lambda}\int_0^\infty \frac{u^{(2\sigma/\lambda)-1}du}{(u+ce^{i\gamma})(u+ce^{-i\gamma})} \\ &= \frac{2\pi}{\lambda\sin\pi(2\sigma/\lambda)}\left[\frac{(ce^{i\gamma})^{(2\sigma/\lambda)-1}}{c(e^{-i\gamma}-e^{i\gamma})} + \frac{(ce^{-i\gamma})^{(2\sigma/\lambda)-1}}{c(e^{i\gamma}-e^{-i\gamma})}\right] \\ &= \frac{2\pi\sin\gamma(1-2\sigma/\lambda)}{\lambda\sin\pi(2\sigma/\lambda)\sin\gamma}c^{\frac{2\sigma}{\lambda}-2} \in \mathbf{R}_+. \end{aligned}$$

In view of (49) and (57), we have

$$||T_1|| = ||T_2|| = \frac{2\pi\sin\gamma(1-2\sigma/\lambda)}{\lambda\sin\pi(2\sigma/\lambda)\sin\gamma}c^{\frac{2\sigma}{\lambda}-2}.$$

Then for (a)–(b), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 1–4. Setting $\delta_0 = \frac{\sigma}{2} > 0$, we can obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 1–4.

Remark 4. Setting $p = q = 2$, $\mu = \sigma = \frac{\lambda}{2}$ in Theorem 1 and Corollary 3, in view of Remark 3 and the above results, if $f(x)$, $g(y) \geq 0$, with

$$0 < \int_0^\infty x^{1-\lambda}f^2(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{1-\lambda} g^2(y)dy < \infty,$$

then we obtain the following eight couples of simpler equivalent inequalities with an independent parameter λ and the best possible constant factors:

(a) For $\lambda > 0$,

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{x^\lambda + y^\lambda} dxdy < \frac{\pi}{\lambda} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{64}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{1 + (xy)^\lambda} dxdy < \frac{\pi}{\lambda} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}; \tag{65}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(x+y)^\lambda} dxdy < B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right) \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{66}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(1+xy)^\lambda} dxdy < B\left(\frac{\lambda}{2}, \frac{\lambda}{2}\right) \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{67}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(\max\{x, y\})^\lambda} dxdy < \frac{4}{\lambda} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{68}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(\max\{xy, 1\})^\lambda} dxdy < \frac{4}{\lambda} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}; \tag{69}$$

$$\int_0^\infty \int_0^\infty \frac{|\ln(\frac{x}{y})| f(x)g(y)}{(\max\{x, y\})^\lambda} dxdy < \frac{8}{\lambda^2} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{70}$$

$$\int_0^\infty \int_0^\infty \frac{|\ln(xy)| f(x)g(y)}{(\max\{1, xy\})^\lambda} dxdy < \frac{8}{\lambda^2} \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}; \tag{71}$$

$$\int_0^\infty \int_0^\infty \frac{\ln(\frac{x}{y}) f(x)g(y)}{x^\lambda - y^\lambda} dxdy < \left(\frac{\pi}{\lambda}\right)^2 \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{72}$$

$$\int_0^\infty \int_0^\infty \frac{\ln(xy) f(x)g(y)}{(xy)^\lambda - 1} dxdy < \left(\frac{\pi}{\lambda}\right)^2 \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}; \tag{73}$$

(b) for $0 < \lambda < 1$,

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{|x-y|^\lambda} dxdy < 2B\left(1-\lambda, \frac{\lambda}{2}\right) \left(\int_0^\infty x^{1-\lambda} f^2(x)dx \int_0^\infty y^{1-\lambda} g^2(y)dy \right)^{\frac{1}{2}}, \tag{74}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{|1-xy|^\lambda}dxdy < 2B\left(1-\lambda, \frac{\lambda}{2}\right)\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}; \tag{75}$$

(c) for $\lambda < 0$,

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(\min\{x,y\})^\lambda}dxdy < \frac{-4}{\lambda}\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}, \tag{76}$$

$$\int_0^\infty \int_0^\infty \frac{f(x)g(y)}{(\min\{1,xy\})^\lambda}dxdy < \frac{-4}{\lambda}\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}; \tag{77}$$

$$\int_0^\infty \int_0^\infty \frac{|\ln(\frac{x}{y})|f(x)g(y)}{(\min\{x,y\})^\lambda}dxdy < \frac{8}{\lambda^2}\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}, \tag{78}$$

$$\int_0^\infty \int_0^\infty \frac{|\ln(xy)|f(x)g(y)}{(\min\{1,xy\})^\lambda}dxdy < \frac{8}{\lambda^2}\left(\int_0^\infty x^{1-\lambda}f^2(x)dx \int_0^\infty y^{1-\lambda}g^2(y)dy\right)^{\frac{1}{2}}. \tag{79}$$

3 Yang–Hilbert-Type Integral Inequalities in the Whole Plane

In this section, we study some Yang–Hilbert-type integral inequalities in the whole plane with parameters and the best constant factors. The equivalent forms, the reverses, the Hardy-type inequalities, the operator expressions, and some particular examples are also discussed.

3.1 Weight Functions and a Lemma

Definition 9. If $\delta \in \{-1, 1\}, \sigma \in \mathbf{R}$, $H(t)$ is a non-negative measurable function in $\mathbf{R}$, define the following weight functions:

$$\omega_\delta(\sigma, y) := |y|^\sigma \int_{-\infty}^\infty H(x^\delta y)|x|^{\delta\sigma-1}dx (y \in \mathbf{R}\backslash\{0\}), \tag{80}$$

$$\varpi_\delta(\sigma, x) := |x|^{\delta\sigma} \int_{-\infty}^\infty H(x^\delta y)|y|^{\sigma-1}dy (x \in \mathbf{R}\backslash\{0\}). \tag{81}$$

Setting $t = x^{\delta}y$ in (80), we obtain $x = y^{-\frac{1}{\delta}}t^{\frac{1}{\delta}}$, $dx = \frac{1}{\delta}y^{-\frac{1}{\delta}}t^{\frac{1}{\delta}-1}dt$ and

$$\begin{aligned}\omega_{\delta}(\sigma, y) &= |y|^{\sigma}\int_{-\infty}^{\infty} H(t)|y^{-\frac{1}{\delta}}t^{\frac{1}{\delta}}|^{\delta\sigma-1}|y|^{-\frac{1}{\delta}}|t|^{\frac{1}{\delta}-1}dt \\ &= K(\sigma) := \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt. \end{aligned} \tag{82}$$

Setting $t = x^{\delta}y$ in (81), we find $y = x^{-\delta}t$, $dy = x^{-\delta}dt$ and

$$\varpi_{\delta}(\sigma, x) = |x|^{\delta\sigma}\int_{-\infty}^{\infty} H(t)|x^{-\delta}t|^{\sigma-1}|x|^{-\delta}dt = K(\sigma). \tag{83}$$

Remark 5. We can still get

$$\begin{aligned}K(\sigma) &= \int_{-\infty}^{0} H(t)(-t)^{\sigma-1}dt + \int_{0}^{\infty} H(t)t^{\sigma-1}dt \\ &= \int_{0}^{\infty} H(-u)u^{\sigma-1}du + \int_{0}^{\infty} H(t)t^{\sigma-1}dt \\ &= \int_{0}^{\infty} (H(-t) + H(t))t^{\sigma-1}dt. \end{aligned} \tag{84}$$

If $H(t) = H(-t)$, then

$$K(\sigma) = 2\int_{0}^{\infty} H(t)t^{\sigma-1}dt,$$

and thus we obtain again cases of integrals in the first quadrant. In this section, we assume that $H(t) \neq H(-t)$.

Lemma 3. *If $p > 0$ $(p \neq 1)$, $\frac{1}{p} + \frac{1}{q} = 1, \sigma \in \mathbf{R}$, both $H(t)$ and $f(t)$ are non-negative measurable functions in $\mathbf{R}$, and $K(\sigma)$ is defined by (83), then, (i) for $p > 1$, we have the following inequality:*

$$\begin{aligned}J &:= \int_{-\infty}^{\infty} |y|^{p\sigma-1}\left(\int_{-\infty}^{\infty} H(x^{\delta}y)f(x)dx\right)^{p} dy \\ &\leq K^{p}(\sigma)\int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1}f^{p}(x)dx; \end{aligned} \tag{85}$$

(ii) for $0 < p < 1$, we have the reverse of (85).

Proof. (i) By Hölder's weighted inequality (cf. [33]) and (80), it follows that

$$\begin{aligned}
&\int_{-\infty}^{\infty} H(x^{\delta}y)f(x)dx \\
&= \int_{-\infty}^{\infty} H(x^{\delta}y)\left[\frac{|x|^{(1-\delta\sigma)/q}}{|y|^{(1-\sigma)/p}}f(x)\right]\left[\frac{|y|^{(1-\sigma)/p}}{|x|^{(1-\delta\sigma)/q}}\right]dx \\
&\le \left[\int_{-\infty}^{\infty} H(x^{\delta}y)\frac{|x|^{(1-\delta\sigma)p/q}}{|y|^{1-\sigma}}f^{p}(x)dx\right]^{\frac{1}{p}} \\
&\quad\times\left[\int_{-\infty}^{\infty} H(x^{\delta}y)\frac{|y|^{(1-\sigma)q/p}}{|x|^{1-\delta\sigma}}dx\right]^{\frac{1}{q}} \\
&= \frac{(\omega_{\delta}(\sigma,y))^{\frac{1}{q}}}{|y|^{\sigma-\frac{1}{p}}}\left[\int_{-\infty}^{\infty} H(x^{\delta}y)\frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}}f^{p}(x)dx\right]^{\frac{1}{p}}. \qquad (86)
\end{aligned}$$

Then, by (82) and Fubini's theorem (cf. [34]), we get

$$\begin{aligned}
J &\le K^{p-1}(\sigma)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x^{\delta}y)\frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}}f^{p}(x)dxdy \\
&= K^{p-1}(\sigma)\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} H(x^{\delta}y)\frac{|x|^{(1-\delta\sigma)(p-1)}}{|y|^{1-\sigma}}dy\right]f^{p}(x)dx \\
&= K^{p-1}(\sigma)\int_{-\infty}^{\infty}\varpi_{\delta}(\sigma,x)|x|^{p(1-\delta\sigma)-1}f^{p}(x)dx. \qquad (87)
\end{aligned}$$

By (83), we obtain (85).

(ii) For $0 < p < 1$, by the reverse of Hölder's weighted inequality (cf. [33]), we can similarly derive the reverses of (86) and (87). Then we obtain the reverse of (85).

This completes the proof of the lemma.

□

3.2 Equivalent Inequalities with the Best Possible Constant Factors

Theorem 5. *Suppose that* $p > 1, \frac{1}{p}+\frac{1}{q} = 1, \sigma \in \mathbf{R}, H(t) \ge 0$, *and*

$$K(\sigma) = \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt \in \mathbf{R}_{+}.$$

If $f(x),\ g(y) \ge 0$, *such that*

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1}f^{p}(x)dx < \infty$$

and

$$0 < \int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we obtain the following equivalent inequalities:

$$\begin{aligned} I &:= \int_{-\infty}^\infty \int_{-\infty}^\infty H(x^\delta y) f(x) g(y) dxdy \\ &< K(\sigma) \left[\int_{-\infty}^\infty |x|^{p(1-\delta\sigma)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \end{aligned} \tag{88}$$

$$\begin{aligned} J &= \int_{-\infty}^\infty |y|^{p\sigma-1} \left(\int_{-\infty}^\infty H(x^\delta y) f(x)dx \right)^p dy \\ &< K^p(\sigma) \int_{-\infty}^\infty |x|^{p(1-\delta\sigma)-1} f^p(x)dx, \end{aligned} \tag{89}$$

where the constant factors $K(\sigma)$ and $K^p(\sigma)$ are the best possible.

In particular, for $\delta = 1$, we have

$$\begin{aligned} I &:= \int_{-\infty}^\infty \int_{-\infty}^\infty H(xy) f(x) g(y) dxdy \\ &< K(\sigma) \left[\int_{-\infty}^\infty |x|^{p(1-\sigma)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_0^\infty y^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \end{aligned} \tag{90}$$

$$\begin{aligned} J &= \int_{-\infty}^\infty |y|^{p\sigma-1} \left(\int_{-\infty}^\infty H(xy) f(x)dx \right)^p dy \\ &< K^p(\sigma) \int_{-\infty}^\infty |x|^{p(1-\sigma)-1} f^p(x)dx. \end{aligned} \tag{91}$$

Proof. We shall initially prove that (86) preserves the form of strict inequality for any $y \in \mathbf{R}\backslash\{0\}$. Otherwise, there exist two constants A and B, such that they are not both zero, and (cf. [33])

$$A \frac{|x|^{(1-\delta\sigma)p/q}}{|y|^{1-\sigma}} f^p(x) = B \frac{|y|^{(1-\sigma)q/p}}{|x|^{1-\sigma}} \quad \text{a. e. in } \mathbf{R}.$$

If $A = 0$, then $B = 0$, which is impossible. Suppose that $A \neq 0$. Then it follows that

$$|x|^{p(1-\delta\sigma)-1} f^p(x) = |y|^{(1-\sigma)q} \frac{B}{A|x|} \quad \text{a. e. in } \mathbf{R},$$

which contradicts the fact that

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x)dx < \infty,$$

in virtue of

$$\int_{-\infty}^{\infty} \frac{1}{|x|} dx = \infty.$$

Hence, both (86) and (87) preserve the forms of strict inequality, and thus we obtain (89).

By Hölder's inequality (cf. [33]), we find

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \left(|y|^{\sigma-\frac{1}{p}} \int_{-\infty}^{\infty} H(x^{\delta}y) f(x)dx \right) (|y|^{\frac{1}{p}-\sigma} g(y))dy \\ &\le J^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}. \end{aligned} \tag{92}$$

Then by (89), we have (88). On the other hand, assuming that (88) is valid, we set

$$g(y) := |y|^{p\sigma-1} \left(\int_{-\infty}^{\infty} H(x^{\delta}y) f(x)dx \right)^{p-1}, \ y \in \mathbf{R}.$$

Then we get

$$J = \int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy.$$

By (85), in view of

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x)dx < \infty,$$

it follows that $J < \infty$.

If $J = 0$, then (89) is trivially valid; if $J > 0$, then by (88), we have

$$\begin{aligned} 0 &< \int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy = J = I \\ &< K(\sigma) \left[\int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \end{aligned}$$

$$J^{\frac{1}{p}} = \left[\int_{-\infty}^{\infty} y^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{p}} < K(\sigma) \left[\int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x)dx \right]^{\frac{1}{p}},$$

and then (89) follows, which is equivalent to (88).

For any $n \in \mathbf{N}$, we define two sets

$$E_\delta := \{x \in \mathbf{R}; |x|^\delta \geq 1\},\ \ E_\delta^+ := \{x \in \mathbf{R}_+; x^\delta \geq 1\},$$

and the functions $f_n(x)$, $g_n(y)$ as follows:

$$f_n(x) := \begin{cases} 0,\ x \in \mathbf{R}\backslash E_\delta \\ |x|^{\delta(\sigma-\frac{1}{np})-1},\ x \in E_\delta \end{cases} \qquad g_n(y) := \begin{cases} |y|^{\sigma+\frac{1}{nq}-1},\ y \in [-1,1] \\ 0,\ \ y \in \mathbf{R}\backslash[-1,1] \end{cases}$$

Then we find

$$\begin{aligned} L_n &:= \left[\int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f_n^p(x)dx\right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g_n^q(y)dy\right]^{\frac{1}{q}} \\ &= \left(\int_{E_\delta} |x|^{-\frac{\delta}{n}-1}dx\right)^{\frac{1}{p}} \left(\int_{-1}^{1} |y|^{\frac{1}{n}-1}dy\right)^{\frac{1}{q}} \\ &= \left(2\int_{E_\delta^+} x^{-\frac{\delta}{n}-1}dx\right)^{\frac{1}{p}} \left(2\int_0^1 y^{\frac{1}{n}-1}dy\right)^{\frac{1}{q}} = 2n. \end{aligned}$$

Setting $Y = -y$, we obtain

$$\begin{aligned} I(x) &:= \int_{-1}^{1} H(x^\delta y)|y|^{\sigma+\frac{1}{nq}-1}dy \\ &= \int_{-1}^{1} H((-x)^\delta Y)|Y|^{\sigma+\frac{1}{nq}-1}dY = I(-x), \end{aligned}$$

and then $I(x)$ is an even function. For $x > 0$, we find

$$\begin{aligned} I(x) &= \int_{-1}^{0} H(x^\delta y)(-y)^{\sigma+\frac{1}{nq}-1}dy + \int_0^1 H(x^\delta y)y^{\sigma+\frac{1}{nq}-1}dy \\ &= x^{-\delta\sigma-\frac{\delta}{nq}} \int_0^{x^\delta} (H(-t) + H(t))t^{\sigma+\frac{1}{nq}-1}dt. \end{aligned}$$

By the above results and Fubini's theorem, it follows that

$$\begin{aligned} I_n &:= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x^\delta y)f_n(x)g_n(y)dxdy \\ &= \int_{E_\delta} |x|^{\delta(\sigma-\frac{1}{np})-1} \left(\int_{-1}^{1} H(x^\delta y)|y|^{\sigma+\frac{1}{nq}-1}dy\right)dx \end{aligned}$$

$$
\begin{aligned}
&= \int_{E_\delta} |x|^{\delta(\sigma-\frac{1}{np})-1} I(x)dx = 2\int_{E_\delta^+} x^{\delta(\sigma-\frac{1}{np})-1} I(x)dx \\
&= 2\int_{E_\delta^+} x^{-\frac{\delta}{n}-1} \left(\int_0^1 (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt \right) dx \\
&\quad +2\int_{E_\delta^+} x^{-\frac{\delta}{n}-1} \left(\int_1^{x^\delta} (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt \right) dx \\
&= 2n \left(\int_0^1 (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt \right) \\
&\quad +2\int_1^\infty \left(\int_{\{x>0; x^\delta \geq t\}} x^{-\frac{\delta}{n}-1} dx \right) (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt \\
&= 2n \left[\int_0^1 (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt + \int_1^\infty (H(-t)+H(t)) t^{\sigma-\frac{1}{np}-1} dt \right].
\end{aligned}
$$

If there exists a positive number $k \leq K(\sigma)$, such that (88) is still valid when replacing $K(\sigma)$ by k, then in particular, it follows that

$$\frac{1}{2n} I_n < k \frac{1}{2n} L_n,$$

and

$$\int_0^1 (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt + \int_1^\infty (H(-t)+H(t)) t^{\sigma-\frac{1}{np}-1} dt < k.$$

Since both

$$\{(H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1}\}_{n=1}^\infty \ (t \in (0,1])$$

and

$$\{(H(-t)+H(t)) t^{\sigma-\frac{1}{np}-1}\}_{n=1}^\infty \ (t \in (1,\infty))$$

are non-negative and increasing, then by Levi's theorem (cf. [34]), it follows that

$$
\begin{aligned}
K(\sigma) &= \int_0^1 (H(-t)+H(t)) t^{\sigma-1} dt + \int_1^\infty (H(-t)+H(t)) t^{\sigma-1} dt \\
&= \lim_{n\to\infty} \left(\int_0^1 (H(-t)+H(t)) t^{\sigma+\frac{1}{nq}-1} dt + \int_1^\infty (H(-t)+H(t)) t^{\sigma-\frac{1}{np}-1} dt \right) \\
&\leq k,
\end{aligned}
$$

and thus $k = K(\sigma)$ is the best possible constant factor of (88).

The constant factor in (89) is still the best possible. Otherwise, we would reach a contradiction by (92). This completes the proof of the theorem. □

Theorem 6. *Replacing $p > 1$ by $0 < p < 1$ in Theorem 1, we obtain the equivalent reverses of (88) and (89). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,*

$$K(\tilde{\sigma}) = \int_{-\infty}^{\infty} H(t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (88) and (89) are the best possible.

Proof. By Lemma 3 and the reverse of Hölder's inequality, we get the reverses of (88), (89), and (92). Similarly, we can set $g(y)$ as in Theorem 5, and prove that the reverses of (88) and (89) are equivalent.

For $n > \frac{2}{\delta_0|q|}$ ($n \in \mathbf{N}$), we set $f_n(x)$ and $g_n(y)$ as in Theorem 1. If there exists a positive number $k \geq K(\sigma)$, such that the reverse of (88) is valid when replacing $K(\sigma)$ by k, then it follows that

$$\frac{1}{2n}I_n > k\frac{1}{2n}L_n,$$

and

$$\int_0^1 (H(-t) + H(t))t^{\sigma+\frac{1}{nq}-1}dt + \int_1^{\infty} (H(-t) + H(t))t^{\sigma-\frac{1}{np}-1}dt > k. \tag{93}$$

Since

$$\{(H(-t) + H(t))t^{\sigma-\frac{1}{np}-1}\}_{n=1}^{\infty} \ (t \in (1, \infty))$$

is still a non-negative and increasing sequence, then by Levi's theorem, it follows that

$$\lim_{n\to\infty} \int_1^{\infty} (H(-t) + H(t))t^{\sigma-\frac{1}{np}-1}dt = \int_1^{\infty} (H(-t) + H(t))t^{\sigma-1}dt.$$

Due to the fact that

$$0 \leq (H(-t) + H(t))t^{\sigma+\frac{1}{nq}-1} \leq (H(-t) + H(t))t^{(\sigma-\frac{\delta_0}{2})-1}$$

($t \in (0, 1], n > \frac{2}{\delta_0|q|}$), and

$$0 \leq \int_0^1 (H(-t) + H(t))t^{(\sigma-\frac{\delta_0}{2})-1}dt \leq K\left(\sigma - \frac{\delta_0}{2}\right) < \infty,$$

then, by Lebesgue's dominated convergence theorem (cf. [34]), it follows that

$$\lim_{n\to\infty}\int_0^1 (H(-t)+H(t))t^{\sigma+\frac{1}{nq}-1}dt = \int_0^1 (H(-t)+H(t))t^{\sigma-1}dt.$$

In view of the above results and (93), we have

$$\begin{aligned} K(\sigma) &= \int_0^1 (H(-t)+H(t))t^{\sigma-1}dt + \int_1^\infty (H(-t)+H(t))t^{\sigma-1}dt \\ &= \lim_{n\to\infty}\left(\int_0^1 (H(-t)+H(t))t^{\sigma+\frac{1}{nq}-1}dt\right. \\ &\quad \left. + \int_1^\infty (H(-t)+H(t))t^{\sigma-\frac{1}{np}-1}dt\right) \geq k, \end{aligned}$$

and then $k = K(\sigma)$ is the best possible constant factor in the reverse of (88).

Similarly, we can prove that the constant factor in the reverse of (89) is the best possible by using the reverse of (92). □

3.3 *Yang–Hilbert-Type Integral Inequalities in the Whole Plane with Multi-Variables*

Theorem 7. *Suppose that* $p > 1$, $\frac{1}{p}+\frac{1}{q} = 1$, $H(t) \geq 0$, $K(\sigma) \in \mathbf{R}_+$, $\delta \in \{-1, 1\}$, $-\infty \leq a_i < b_i \leq \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = -\infty$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\delta\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2}\int_{a_1}^{b_1} H(v_1^\delta(x)v_2(y))f(x)g(y)dxdy \\ &< K(\sigma)\left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\delta\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx\right]^{\frac{1}{p}} \\ &\quad \times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy\right]^{\frac{1}{q}}, \end{aligned} \tag{94}$$

$$\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{a_1}^{b_1} H(v_1^{\delta}(x)v_2(y)) f(x)dx \right)^p dy$$
$$< K^p(\sigma) \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\delta\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \tag{95}$$

where the constant factors $K(\sigma)$ and $K^p(\sigma)$ are the best possible.

Proof. Setting $x = v_1(s)$, $y = v_2(t)$ in (88), we get $dx = v_1'(s)ds$, $dy = v_2'(t)dt$, and

$$I = \int_{a_2}^{b_2} \int_{a_1}^{b_1} H(v_1^{\delta}(s)v_2(t)) f(v_1(s)) g(v_2(t)) v_1'(s) v_2'(t) ds dt$$
$$= \int_{a_2}^{b_2} \int_{a_1}^{b_1} H(v_1^{\delta}(s)v_2(t)) (f(v_1(s))v_1'(s)) (g(v_2(t))v_2'(t)) ds dt,$$
$$I_1 := \int_{-\infty}^{\infty} |x|^{p(1-\delta\sigma)-1} f^p(x)dx = \int_{a_1}^{b_1} |v_1(s)|^{p(1-\delta\sigma)-1} f^p(v_1(s))v_1'(s)ds,$$
$$I_2 := \int_{-\infty}^{\infty} y^{q(1-\sigma)-1} g^q(y)dy = \int_{a_2}^{b_2} |v_2(t)|^{q(1-\sigma)-1} g^q(v_2(t))v_2'(t)dt.$$

If we set

$$F(s) = f(v_1(s))v_1'(s) \text{ and } G(t) = g(v_2(t))v_2'(t),$$

we obtain

$$f^p(v_1(s)) = (v_1'(s))^{-p} F^p(s), g^q(v_2(t)) = (v_2'(t))^{-q} G^q(t),$$

and then it follows that

$$I = \int_{a_2}^{b_2} \int_{a_1}^{b_1} H(v_1^{\delta}(s)v_2(t)) F(s) G(t) ds dt,$$
$$I_1 = \int_{a_1}^{b_1} \frac{|v_1(s)|^{p(1-\delta\sigma)-1}}{(v_1'(s))^{p-1}} F^p(s)ds,$$
$$I_2 = \int_{a_2}^{b_2} \frac{|v_2(t)|^{q(1-\sigma)-1}}{(v_2'(t))^{q-1}} G^q(t)dt.$$

Substitution of the above results to (88), with

$$s = x,\ t = y,\ F(s) = f(x),\ \text{and}\ G(t) = g(y),$$

we obtain (94). Similarly, we derive (95). On the other hand, if we set

$$v_1(x) = x,\ v_2(y) = y,\ a_i = -\infty,\ b_i = \infty$$

in (94), we get (88). Hence, the inequalities (94) and (88) are equivalent. It is evident that the inequalities (95) and (89) are equivalent. Hence, the inequalities (94) and (95) are equivalent. Since the constant factors in (88) and (89) are the best possible, it follows that the constant factors in (94) and (95) are also the best possible. This completes the proof of the theorem. □

Theorem 8. *Replacing $p > 1$ by $0 < p < 1$ in Theorem 7, we obtain the equivalent reverses of (94) and (95). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,*

$$K(\tilde{\sigma}) = \int_{-\infty}^{\infty} H(t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (94) and (95) are the best possible.

Remark 6. We list the following $v_i(s)$ $(i = 1, 2)$ that satisfy the conditions of Theorems 7 and 8:

(a) $v_i(s) = s^\gamma,\ s \in (-\infty, \infty)$ $(\gamma \in \{a; a = \frac{1}{2k-1}, 2k+1\ (k \in \mathbf{N})\})$, satisfying $v_i'(s) = \gamma s^{\gamma-1} > 0$;
(b) $v_i(s) = \tan^\gamma s,\ s \in (-\frac{\pi}{2}, \frac{\pi}{2})$ $(\gamma \in \{a; a = \frac{1}{2k-1}, 2k+1\ (k \in \mathbf{N})\})$, satisfying $v_i'(s) = \gamma \tan^{\gamma-1} s \sec^2 s > 0$;
(c) $v_i(s) = \ln^\gamma s,\ s \in (0, \infty)$ $(\gamma \in \{a; a = \frac{1}{2k-1}, 2k+1\ (k \in \mathbf{N})\})$, satisfying $v_i'(s) = \frac{\gamma}{s} \ln^{\gamma-1} s > 0$;
(d) $v_i(s) = (e^{|s|} - 1)\mathrm{sgn}(s),\ s \in (-\infty, \infty)$, satisfying $v_i'(s) = e^{|s|} > 0$.

Definition 10. If $\lambda \in \mathbf{R}$, $K_\lambda(x, y)$ is a non-negative measurable function in $\mathbf{R}^2$, satisfying

$$K_\lambda(tx, ty) = |t|^{-\lambda} K_\lambda(x, y),$$

for any $t \in \mathbf{R}\backslash\{0\}$, $x,\ y \in \mathbf{R}$, then $K_\lambda(x, y)$ is said to be the homogeneous function of degree $-\lambda$ in $\mathbf{R}^2$.

In particular, by Theorems 7 and 8, with $\delta = 1$, we obtain the following integral inequalities with the non-homogeneous kernel:

Corollary 12. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1, \sigma \in \mathbf{R}, H(t) \geq 0,$

$$K(\sigma) = \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt \in \mathbf{R}_+,$$

$-\infty \leq a_i < b_i \leq \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = -\infty$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x)$, $g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2}\int_{a_1}^{b_1} H(v_1(x)v_2(y))f(x)g(y)dxdy \\ &< K(\sigma)\left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx\right]^{\frac{1}{p}} \\ &\times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy\right]^{\frac{1}{q}}, \end{aligned} \tag{96}$$

$$\begin{aligned} &\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{a_1}^{b_1} H(v_1(x)v_2(y))f(x)dx\right)^p dy \\ &< K^p(\sigma) \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \end{aligned} \tag{97}$$

where the constant factors $K(\sigma)$ *and* $K^p(\sigma)$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in Corollary 12, we derive the equivalent reverses of (96) and (97). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K(\tilde{\sigma}) = \int_{-\infty}^{\infty} H(t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (96) and (97) are the best possible.

In particular, for $\delta = -1$ in Theorems 7 and 8, setting $H(t) = K_\lambda(1,t)$ (cf. Definition 10), we get

$$H\left(\frac{v_2(y)}{v_1(x)}\right) = K_\lambda\left(1, \frac{v_2(y)}{v_1(x)}\right) = |v_1(x)|^\lambda K_\lambda(v_1(x), v_2(y)).$$

Replacing $f(x)$ by $|v_1(x)|^{-\lambda} f(x)$, it follows that $|v_1(x)|^{p(1+\sigma)-1} f^p(x)$ is replaced by

$$|v_1(x)|^{p(1+\sigma)-1}[|v_1(x)|^{-\lambda} f(x)]^p = |v_1(x)|^{p(1-\mu)-1} f^p(x),$$

and we have the following integral inequalities with the homogeneous kernel:

Corollary 13. *Let $p > 1, \frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $K_\lambda(x,y)$ is a homogeneous function in $\mathbf{R}^2$ of degree $-\lambda$,*

$$K_\lambda(\sigma) := \int_{-\infty}^{\infty} K_\lambda(1,t)|t|^{\sigma-1} dt \in \mathbf{R}_+,$$

$0 \le a_i < b_i \le \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = -\infty, v_i(b_i^-) = \infty$ $(i = 1,2)$. If $f(x), g(y) \ge 0$, such that

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y) dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2}\int_{a_1}^{b_1} K_\lambda(v_1(x), v_2(y)) f(x) g(y) dx dy \\ &< K_\lambda(\sigma)\left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx\right]^{\frac{1}{p}} \\ &\quad \times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y) dy\right]^{\frac{1}{q}}, \end{aligned} \tag{98}$$

$$\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{a_1}^{b_1} K_\lambda(v_1(x), v_2(y)) f(x) dx \right)^p dy$$
$$< K_\lambda^p(\sigma) \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x) dx, \tag{99}$$

where the constant factors $K_\lambda(\sigma)$ and $K_\lambda^p(\sigma)$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above cases, we obtain the equivalent reverses of (98) and (99). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K_\lambda(\tilde{\sigma}) = \int_{-\infty}^{\infty} K_\lambda(1, t)|t|^{\tilde{\sigma}-1} dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (98) and (99) are the best possible.

Setting $a_i = -\infty$, $b_i = \infty$ $(i = 1, 2)$, $v_1(x) = x$, $v_2(y) = y$ in Corollary 13, we obtain the following corollary:

Corollary 14. *Let $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $K_\lambda(x, y)$ is a homogeneous function in $\mathbf{R}^2$ of degree $-\lambda$,*

$$K_\lambda(\sigma) = \int_{-\infty}^{\infty} K_\lambda(1, t) t^{\sigma-1} dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_0^{\infty} x^{p(1-\mu)-1} f^p(x) dx < \infty$$

and

$$0 < \int_0^{\infty} g^{q(1-\sigma)-1}(y) dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_\lambda(x, y) f(x) g(y) dx dy$$
$$< K_\lambda(\sigma) \left[\int_{-\infty}^{\infty} x^{p(1-\mu)-1} f^p(x) dx \right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} y^{q(1-\sigma)-1} g^q(y) dy \right]^{\frac{1}{q}}, \tag{100}$$

$$\int_{-\infty}^{\infty} y^{p\sigma-1} \left(\int_{-\infty}^{\infty} K_\lambda(x, y) f(x) dx \right)^p dy < K_\lambda^p(\sigma) \int_0^{\infty} x^{p(1-\mu)-1} f^p(x) dx, \tag{101}$$

where the constant factors $K_\lambda(\sigma)$ and $K_\lambda^p(\sigma)$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above cases, we get the equivalent reverses of (100) and (101). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K_\lambda(\tilde{\sigma}) = \int_{-\infty}^{\infty} K_\lambda(1,t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (100) and (101) are the best possible.

Remark 7. It is evident that (90) and (100) are equivalent for $H(t) = K_\lambda(1,t)$. The same holds for (91) and (101).

3.4 Hardy-Type Integral Inequalities in the Whole Plane

In the following two sections, if the constant factors in the inequalities (operator inequalities) are related to $K^{(1)}(\sigma)$ (or $K_\lambda^{(1)}(\sigma)$), then we shall call them Hardy-type inequalities (operator) of the first kind; if the constant factors in the inequalities (operator inequalities) are related to $K^{(2)}(\sigma)$ (or $K_\lambda^{(2)}(\sigma)$), then we shall call them Hardy-type inequalities (operator) of the second kind.

If $H(t) = 0$ ($|t| > 1$), then $H(xy) = 0$ ($|x| > \frac{1}{|y|} > 0$), and

$$K(\sigma) = \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt = \int_{-1}^{1} H(t)|t|^{\sigma-1}dt.$$

Set

$$K^{(1)}(\sigma) := \int_{-1}^{1} H(t)|t|^{\sigma-1}dt = \int_{0}^{1} (H(-t) + H(t))t^{\sigma-1}dt. \tag{102}$$

Then, by Theorems 7 and 8 ($\delta = 1$), we have the following Hardy-type integral inequalities of the first kind, with non-homogeneous kernel in the whole plane:

Corollary 15. *Suppose that* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $H(t) \geq 0, \sigma \in \mathbf{R}$,

$$K^{(1)}(\sigma) = \int_{-1}^{1} H(t)|t|^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_{-\infty}^{\infty} |x|^{p(1-\sigma)-1} f^p(x)dx < \infty$$

and

$$0 < \int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{-\infty}^{\infty} \left(\int_{-\frac{1}{|y|}}^{\frac{1}{|y|}} H(xy)f(x)dx \right) g(y)dy = \int_{0}^{\infty} \left(\int_{-\frac{1}{|x|}}^{\frac{1}{|x|}} H(xy)g(y)dy \right) f(x)dx$$

$$< K^{(1)}(\sigma) \left[\int_{-\infty}^{\infty} |x|^{p(1-\sigma)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \quad (103)$$

$$\int_{-\infty}^{\infty} |y|^{p\sigma-1} \left(\int_{-\frac{1}{|y|}}^{\frac{1}{|y|}} H(xy)f(x)dx \right)^p dy < (K^{(1)}(\sigma))^p \int_{-\infty}^{\infty} |x|^{p(1-\sigma)-1} f^p(x)dx; \quad (104)$$

where the constant factors $K^{(1)}(\sigma)$ *and* $(K^{(1)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (103) and (104). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K^{(1)}(\tilde{\sigma}) = \int_{-1}^{1} H(t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (103) and (104) are the best possible.

If we have $H(t) = 0$ ($|t| > 1$) in Corollary 12, then

$$H(v_1(x)v_2(y)) = 0 \left(|v_1(x)| > \frac{1}{|v_2(y)|} > 0 \right),$$

and thus we derive the following general results:

Corollary 16. *Suppose that* $p > 1, \frac{1}{p} + \frac{1}{q} = 1$, $H(t) \geq 0, \sigma \in \mathbf{R}$,

$$K^{(1)}(\sigma) = \int_{-1}^{1} H(t)|t|^{\sigma-1}dt \in \mathbf{R}_+,$$

$-\infty \leq a_i < b_i \leq \infty$, $v_i'(s) > 0$ ($s \in (a_i, b_i)$), $v_i(a_i^+) = -\infty$, $v_i(b_i^-) = \infty$ ($i = 1, 2$). *If* $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we obtain the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2} \left(\int_{v_1^{-1}(\frac{-1}{|v_2(y)|})}^{v_1^{-1}(\frac{1}{|v_2(y)|})} H(v_1(x)v_2(y))f(x)dx \right) g(y)dy \\
&< K^{(1)}(\sigma) \left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx \right]^{\frac{1}{p}} \\
&\times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy \right]^{\frac{1}{q}}, \qquad (105)
\end{aligned}$$

$$\begin{aligned}
&\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{v_1^{-1}(\frac{-1}{|v_2(y)|})}^{v_1^{-1}(\frac{1}{|v_2(y)|})} H(v_1(x)v_2(y))f(x)dx \right)^p dy \\
&< (K^{(1)}(\sigma))^p \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \qquad (106)
\end{aligned}$$

where the constant factors $K^{(1)}(\sigma)$ and $(K^{(1)}(\sigma))^p$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we get the equivalent reverses of (105) and (106). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K^{(1)}(\tilde{\sigma}) = \int_{-1}^{1} H(t)|t|^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (105) and (106) are the best possible.

If $H(t) = 0$ $(0 < |t| < 1)$, then $H(xy) = 0$ $(0 < |x| < \frac{1}{|y|})$, and

$$\begin{aligned}
K(\sigma) &= \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt = \int_{-\infty}^{-1} H(t)(-t)^{\sigma-1}dt + \int_{1}^{\infty} H(t)t^{\sigma-1}dt \\
&= \int_{1}^{\infty} (H(-t) + H(t))t^{\sigma-1}dt. \qquad (107)
\end{aligned}$$

If we set

$$E_y := \left\{ x \in \mathbf{R}; x \geq \frac{1}{|y|}, \text{or } x \leq \frac{-1}{|y|} \right\},$$

and

$$K^{(2)}(\sigma) := \int_{E_1} H(t)|t|^{\sigma-1}dt = \int_1^\infty (H(-t)+H(t))t^{\sigma-1}dt,$$

then, by Theorems 7 and 8 ($\delta = 1$), we obtain the following Hardy-type integral inequalities of the second kind with the non-homogeneous kernel in the whole plane:

Corollary 17. *Let* $p > 1$, $\frac{1}{p}+\frac{1}{q}=1$, $H(t) \geq 0, \sigma \in \mathbf{R}$,

$$K^{(2)}(\sigma) = \int_1^\infty (H(-t)+H(t))t^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), g(y) \geq 0$,

$$0 < \int_{-\infty}^\infty |x|^{p(1-\sigma)-1}f^p(x)dx < \infty$$

and

$$0 < \int_{-\infty}^\infty |y|^{q(1-\sigma)-1}g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{-\infty}^\infty \left(\int_{E_y} H(xy)f(x)dx\right) g(y)dy$$
$$< K^{(2)}(\sigma)\left[\int_{-\infty}^\infty |x|^{p(1-\sigma)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^\infty |y|^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \quad (108)$$

$$\int_{-\infty}^\infty |y|^{p\sigma-1}\left(\int_{E_y} H(xy)f(x)dx\right)^p dy < (K^{(2)}(\sigma))^p \int_{-\infty}^\infty |x|^{p(1-\sigma)-1}f^p(x)dx; \quad (109)$$

where the constant factors $K^{(2)}(\sigma)$ *and* $(K^{(2)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we get the equivalent reverses of (108) and (109). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K^{(2)}(\tilde{\sigma}) = \int_1^\infty (H(-t)+H(t))t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (108) and (109) are the best possible.

If we have $H(t) = 0$ $(0 < |t| < 1)$ in Corollary 12, then

$$H(v_1(x)v_2(y)) = 0 \left(0 < |v_1(x)| < \frac{1}{|v_2(y)|}\right).$$

Setting

$$\tilde{E}_y := \left\{x \in (a_1, b_1);\ x \geq v_1^{-1}\left(\frac{1}{|v_2(y)|}\right) \text{ or } x \leq v_1^{-1}\left(\frac{-1}{|v_2(y)|}\right)\right\},$$

we obtain the following general results:

Corollary 18. *Let* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $H(t) \geq 0, \sigma \in \mathbf{R}$,

$$K^{(2)}(\sigma) = \int_1^{\infty} (H(-t) + H(t))t^{\sigma-1}dt \in \mathbf{R}_+,$$

$-\infty \leq a_i < b_i \leq \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = -\infty$, $v_i(b_i^-) = \infty$ $(i = 1, 2)$. *If* $f(x), g(y) \geq 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned}
&\int_{a_2}^{b_2} \left(\int_{\tilde{E}_y} H(v_1(x)v_2(y))f(x)dx\right) g(y)dy \\
&< K^{(2)}(\sigma) \left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx\right]^{\frac{1}{p}} \\
&\quad \times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy\right]^{\frac{1}{q}}, \qquad (110)
\end{aligned}$$

$$\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{\tilde{E}_y} H(v_1(x)v_2(y))f(x)dx \right)^p dy$$
$$< (K^{(2)}(\sigma))^p \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\sigma)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \tag{111}$$

where the constant factors $K^{(2)}(\sigma)$ and $(K^{(2)}(\sigma))^p$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we get the equivalent reverses of (110) and (111). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K^{(2)}(\tilde{\sigma}) = \int_1^\infty (H(-t) + h(t))t^{\tilde{\sigma}-1}dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (110) and (111) are the best possible.

Similarly, if $K_\lambda(1,t) = 0$ $(|t| > 1)$, then

$$K_\lambda(x,y) = |x|^{-\lambda} K_\lambda\left(1, \frac{y}{x}\right) = 0 \ (|y| > |x|).$$

By Corollary 14, setting

$$F_y := \{x \in \mathbf{R}; x \geq |y| \ \text{ or } \ x \leq -|y|\},$$

we obtain the following Hardy-type integral inequalities of the first kind, with the homogeneous kernel in the whole plane:

Corollary 19. *Let $p > 1, \frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $K_\lambda(x,y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}^2$,*

$$K_\lambda^{(1)}(\sigma) = \int_{-1}^1 K_\lambda(1,t)|t|^{\sigma-1}dt \in \mathbf{R}_+.$$

If $f(x), \ g(y) \geq 0$, such that

$$0 < \int_{-\infty}^\infty |x|^{p(1-\mu)-1} f^p(x)dx < \infty$$

and

$$0 < \int_{-\infty}^\infty |y|^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{-\infty}^{\infty}\left(\int_{F_y} K_\lambda(x,y)f(x)dx\right)g(y)dy$$
$$< K_\lambda^{(1)}(\sigma)\left[\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^p(x)dx\right]^{\frac{1}{p}}\left[\int_{-\infty}^{\infty}|y|^{q(1-\sigma)-1}g^q(y)dy\right]^{\frac{1}{q}}, \quad (112)$$

$$\int_{-\infty}^{\infty}|y|^{p\sigma-1}\left(\int_{F_y} K_\lambda(x,y)f(x)dx\right)^p dy < (K_\lambda^{(1)}(\sigma))^p\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^p(x)dx, \quad (113)$$

where the constant factors $K_\lambda^{(1)}(\sigma)$ and $(K_\lambda^{(1)}(\sigma))^p$ are the best possible.

Replacing $p>1$ by $0<p<1$ in the above inequalities, we get the equivalent reverses of (112) and (113). If there exists a constant $\delta_0>0$, such that for any $\tilde{\sigma}\in(\sigma-\delta_0,\sigma]$,

$$K_\lambda^{(1)}(\tilde{\sigma})=\int_{-1}^{1}K_\lambda(1,t)|t|^{\tilde{\sigma}-1}dt\in\mathbf{R}_+,$$

then the constant factors in the reverses of (112) and (113) are the best possible.

If $K_\lambda(1,t)=0$ ($|t|>1$) in Corollary 12, then

$$K_\lambda(v_1(x),v_2(y))=0\ (0<|v_1(x)|<|v_2(y)|).$$

Setting

$$\tilde{F}_y:=\{x\in(a_1,b_1);\ x\geq v_1^{-1}(|v_2(y)|)\ \text{or}\ \ x\leq v_1^{-1}(-|v_2(y)|)\},$$

we obtain the following general results:

Corollary 20. *Let $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, $\sigma,\ \mu\in\mathbf{R}$, $\sigma+\mu=\lambda$, $K_\lambda(x,y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}^2$,*

$$K_\lambda^{(1)}(\sigma)=\int_{-1}^{1}K_\lambda(1,t)|t|^{\sigma-1}dt\in\mathbf{R}_+,$$

$-\infty\leq a_i<b_i\leq\infty$, $v_i'(s)>0$ ($s\in(a_i,b_i)$), $v_i(a_i^+)=-\infty$, $v_i(b_i^-)=\infty$ ($i=1,2$). If $f(x),\ g(y)\geq 0$, such that

$$0<\int_{a_1}^{b_1}\frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx<\infty$$

and

$$0<\int_{a_2}^{b_2}\frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}}g^q(y)dy<\infty,$$

then we have the following equivalent inequalities:

$$\int_{a_2}^{b_2}\left(\int_{\tilde{F}_y} K_\lambda(v_1(x),v_2(y))f(x)dx\right)g(y)dy$$
$$< K_\lambda^{(1)}(\sigma)\left[\int_{a_1}^{b_1}\frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx\right]^{\frac{1}{p}}$$
$$\times\left[\int_{a_2}^{b_2}\frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}}g^q(y)dy\right]^{\frac{1}{q}}, \tag{114}$$

$$\int_{a_2}^{b_2}\frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}}\left(\int_{\tilde{F}_y} K_\lambda(v_1(x),v_2(y))f(x)dx\right)^p dy$$
$$< (K_\lambda^{(1)}(\sigma))^p\int_{a_1}^{b_1}\frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}}f^p(x)dx, \tag{115}$$

where the constant factors $K_\lambda^{(1)}(\sigma)$ and $(K_\lambda^{(1)}(\sigma))^p$ are the best possible.

Replacing $p>1$ by $0<p<1$ in the above inequalities, we obtain the equivalent reverses of (114) and (115). If there exists a constant $\delta_0>0$, such that for any $\tilde{\sigma}\in(\sigma-\delta_0,\sigma]$,

$$K_\lambda^{(1)}(\tilde{\sigma})=\int_{-1}^{1}K_\lambda(1,t)|t|^{\tilde{\sigma}-1}dt\in\mathbf{R}_+,$$

then the constant factors in the reverses of (114) and (115) are the best possible.

Similarly, if $K_\lambda(1,t)=0$ $(0<|t|<1)$ in Corollary 14, then

$$K_\lambda(x,y)=0\ (|x|>|y|>0).$$

The following Hardy-type integral inequalities of the second kind with the homogeneous kernel hold true:

Corollary 21. *Let $p>1$, $\frac{1}{p}+\frac{1}{q}=1$, $\mu,\sigma\in\mathbf{R}$, $\mu+\sigma=\lambda$, $K_\lambda(x,y)$ is a homogeneous function of degree $-\lambda$ in $\mathbf{R}^2$,*

$$K_\lambda^{(2)}(\sigma)=\int_1^\infty(K_\lambda(1,-t)+K_\lambda(1,t))t^{\sigma-1}dt\in\mathbf{R}_+.$$

If $f(x),g(y)\geq 0$,

$$0<\int_{-\infty}^{\infty}|x|^{p(1-\mu)-1}f^p(x)dx<\infty$$

and

$$0 < \int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\int_{-\infty}^{\infty} \left(\int_{-|y|}^{|y|} K_\lambda(x,y) f(x)dx \right) g(y)dy = \int_{-\infty}^{\infty} \left(\int_{-|x|}^{|x|} K_\lambda(x,y) g(y)dy \right) f(x)dx$$

$$< K_\lambda^{(2)}(\sigma) \left[\int_{-\infty}^{\infty} |x|^{p(1-\mu)-1} f^p(x)dx \right]^{\frac{1}{p}} \left[\int_{-\infty}^{\infty} |y|^{q(1-\sigma)-1} g^q(y)dy \right]^{\frac{1}{q}}, \quad (116)$$

$$\int_{-\infty}^{\infty} |y|^{p\sigma-1} \left(\int_{-|y|}^{|y|} K_\lambda(x,y) f(x)dx \right)^p dy < (K_\lambda^{(2)}(\sigma))^p \int_{-\infty}^{\infty} |x|^{p(1-\mu)-1} f^p(x)dx; \quad (117)$$

where the constant factors $K_\lambda^{(2)}(\sigma)$ *and* $(K_\lambda^{(2)}(\sigma))^p$ *are the best possible.*

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we obtain the equivalent reverses of (116) and (117). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K_\lambda^{(2)}(\tilde{\sigma}) = \int_1^{\infty} (K_\lambda(1,-t) + K_\lambda(1,t)) t^{\tilde{\sigma}-1} dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (116) and (117) are the best possible.

If $K_\lambda(1,t) = 0$ $(0 < |t| < 1)$ in Corollary 12, then

$$K_\lambda(v_1(x), v_2(y)) = 0 \ (|v_1(x)| > |v_2(y)| > 0),$$

we have the following general results:

Corollary 22. *Let* $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$, $K_\lambda(x,y)$ *is a homogeneous function of degree* $-\lambda$ *in* $\mathbf{R}^2$,

$$K_\lambda^{(2)}(\sigma) = \int_1^{\infty} (K_\lambda(1,-t) + K_\lambda(1,t)) t^{\sigma-1} dt \in \mathbf{R}_+,$$

$-\infty \le a_i < b_i \le \infty$, $v_i'(s) > 0$ $(s \in (a_i, b_i))$, $v_i(a_i^+) = -\infty$, $v_i(b_i^-) = \infty$ $(i = 1,2)$. *If* $f(x)$, $g(y) \ge 0$, *such that*

$$0 < \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x)dx < \infty$$

and

$$0 < \int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy < \infty,$$

then we have the following equivalent inequalities:

$$\begin{aligned} &\int_{a_2}^{b_2} \left(\int_{v_1^{-1}(-|v_2(y)|)}^{v_1^{-1}(|v_2(y)|)} K_\lambda(v_1(x), v_2(y)) f(x)dx \right) g(y)dy \\ &< K_\lambda^{(2)}(\sigma) \left[\int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x)dx \right]^{\frac{1}{p}} \\ &\times \left[\int_{a_2}^{b_2} \frac{|v_2(y)|^{q(1-\sigma)-1}}{(v_2'(y))^{q-1}} g^q(y)dy \right]^{\frac{1}{q}}, \end{aligned} \tag{118}$$

$$\begin{aligned} &\int_{a_2}^{b_2} \frac{v_2'(y)}{|v_2(y)|^{1-p\sigma}} \left(\int_{v_1^{-1}(-|v_2(y)|)}^{v_1^{-1}(|v_2(y)|)} K_\lambda(v_1(x), v_2(y)) f(x)dx \right)^p dy \\ &< (K_\lambda^{(2)}(\sigma))^p \int_{a_1}^{b_1} \frac{|v_1(x)|^{p(1-\mu)-1}}{(v_1'(x))^{p-1}} f^p(x)dx, \end{aligned} \tag{119}$$

where the constant factors $K_\lambda^{(2)}(\sigma)$ and $(K_\lambda^{(2)}(\sigma))^p$ are the best possible.

Replacing $p > 1$ by $0 < p < 1$ in the above inequalities, we have the equivalent reverses of (118) and (119). If there exists a constant $\delta_0 > 0$, such that for any $\tilde{\sigma} \in (\sigma - \delta_0, \sigma]$,

$$K_\lambda^{(2)}(\tilde{\sigma}) = \int_1^\infty (K_\lambda(1, -t) + K_\lambda(1, t)) t^{\tilde{\sigma}-1} dt \in \mathbf{R}_+,$$

then the constant factors in the reverses of (118) and (119) are the best possible.

3.5 Yang–Hilbert-Type Operators and Hardy-Type Operators in the Whole Plane

Let $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, $\mu, \sigma \in \mathbf{R}$, $\mu + \sigma = \lambda$. We define the following functions:

$$\varphi(x) := |x|^{p(1-\sigma)-1}, \psi(y) := |y|^{q(1-\sigma)-1}, \phi(x) := |x|^{p(1-\mu)-1} (x, y \in \mathbf{R}),$$

wherefrom, $\psi^{1-p}(y) = |y|^{p\sigma-1}$.

We define also the following real normed linear space:

$$L_{p,\varphi}(\mathbf{R}) := \left\{ f : ||f||_{p,\varphi} := \left\{ \int_{-\infty}^{\infty} \varphi(x)|f(x)|^p dx \right\}^{\frac{1}{p}} < \infty \right\},$$

wherefrom,

$$L_{p,\psi^{1-p}}(\mathbf{R}) = \left\{ h : ||h||_{p,\psi^{1-p}} := \left\{ \int_{-\infty}^{\infty} \psi^{1-p}(y)|h(y)|^p dy \right\}^{\frac{1}{p}} < \infty \right\},$$

$$L_{p,\phi}(\mathbf{R}) = \left\{ g : ||g||_{p,\phi} := \left\{ \int_{-\infty}^{\infty} \phi(x)|g(x)|^p dx \right\}^{\frac{1}{p}} < \infty \right\}.$$

(a) In view of Theorem 5 ($\delta = 1$), for $f \in L_{p,\varphi}(\mathbf{R})$,

$$H_1(y) := \int_{-\infty}^{\infty} H(xy)|f(x)|dx \quad (y \in \mathbf{R}_+),$$

by (91), we have

$$||H_1||_{p,\psi^{1-p}} := \left(\int_{-\infty}^{\infty} \psi^{1-p}(y) H_1^p(y) dy \right)^{\frac{1}{p}} < K(\sigma)||f||_{p,\varphi} < \infty. \tag{120}$$

Definition 11. Define Yang–Hilbert-type integral operator with the non-homogeneous kernel in the whole plane

$$T_1 : L_{p,\varphi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\varphi}(\mathbf{R})$, there exists a unique representation

$$T_1 f = H_1 \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_1 f(y) = H_1(y),$$

for any $y \in \mathbf{R}$.

In view of (120), it follows that

$$||T_1 f||_{p,\psi^{1-p}} = ||H_1||_{p,\psi^{1-p}} \le K(\sigma)||f||_{p,\varphi}.$$

Therefore, the operator T_1 is bounded and it satisfies the following relation

$$||T_1|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R})} \frac{||T_1 f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \leq K(\sigma).$$

Since the constant factor $K(\sigma)$ in (120) is the best possible, we have

$$||T_1|| = K(\sigma) = \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1}dt. \tag{121}$$

If we define the formal inner product of $T_1 f$ and g as

$$\begin{aligned}(T_1 f, g) &:= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} H(xy) f(x) dx \right) g(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(xy) f(x) g(y) dx dy,\end{aligned}$$

then we can rewrite (90) and (91) as follows:

$$(T_1 f, g) < ||T_1|| \cdot ||f||_{p,\varphi} ||g||_{q,\psi}, \quad ||T_1 f||_{p,\psi^{1-p}} < ||T_1|| \cdot ||f||_{p,\varphi}.$$

(b) In view of Corollary 15, for $f \in L_{p,\varphi}(\mathbf{R})$, setting

$$H_1^{(1)}(y) := \int_{\frac{-1}{|y|}}^{\frac{1}{|y|}} H(xy)|f(x)|dx (y \in \mathbf{R}\backslash\{0\}),$$

by (104), we obtain

$$||H_1^{(1)}||_{p,\psi^{1-p}} = \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)(H_1^{(1)}(y))^p dy \right)^{\frac{1}{p}} < K^{(1)}(\sigma)||f||_{p,\varphi} < \infty. \tag{122}$$

Definition 12. Let us define the Hardy-type integral operator of the first kind, with the non-homogeneous kernel in the whole plane

$$T_1^{(1)} : L_{p,\varphi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\varphi}(\mathbf{R})$, there exists a unique representation

$$T_1^{(1)} f = H_1^{(1)} \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_1^{(1)} f(y) = H_1^{(1)}(y),$$

for any $y \in \mathbf{R}$.

In view of (122), it follows that

$$||T_1^{(1)}f||_{p,\psi^{1-p}} = ||H_1^{(1)}||_{p,\psi^{1-p}} \leq K^{(1)}(\sigma)||f||_{p,\varphi}.$$

Then, the operator $T_1^{(1)}$ is bounded satisfying

$$||T_1^{(1)}|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R})} \frac{||T_1^{(1)}f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \leq K^{(1)}(\sigma).$$

Since the constant factor $K^{(1)}(\sigma)$ in (122) is the best possible, we have

$$||T_1^{(1)}|| = K^{(1)}(\sigma) = \int_{-1}^{1} H(t)|t|^{\sigma-1}dt. \tag{123}$$

Setting the formal inner product of $T_1^{(1)}f$ and g as

$$(T_1^{(1)}f, g) = \int_{-\infty}^{\infty}\left(\int_{\frac{-1}{|y|}}^{\frac{1}{|y|}} H(xy)f(x)dx\right)g(y)dy,$$

we can rewrite (103) and (104) as follows:

$$(T_1^{(1)}f, g) < ||T_1^{(1)}|| \cdot ||f||_{p,\varphi}||g||_{q,\psi}, \quad ||T_1^{(1)}f||_{p,\psi^{1-p}} < ||T_1^{(1)}|| \cdot ||f||_{p,\varphi}. \tag{124}$$

(c) In view of Corollary 17, for $f \in L_{p,\varphi}(\mathbf{R})$, setting

$$H_1^{(2)}(y) := \int_{E_y} H(xy)|f(x)|dx \ \ (y \in \mathbf{R}),$$

by (109), we have

$$\begin{aligned} ||H_1^{(2)}||_{p,\psi^{1-p}} &= \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)(H_1^{(2)}(y))^p dy\right)^{\frac{1}{p}} \\ &< K^{(2)}(\sigma)||f||_{p,\varphi} < \infty. \end{aligned} \tag{125}$$

Definition 13. Let us define the Hardy-type integral operator of the second kind with the non-homogeneous kernel in the whole plane

$$T_1^{(2)} \ : \ L_{p,\varphi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\varphi}(\mathbf{R})$, there exists a unique representation

$$T_1^{(2)} f = H_1^{(2)} \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_1^{(2)} f(y) = H_1^{(2)}(y),$$

for any $y \in \mathbf{R}$.

In view of (125), it follows that

$$||T_1^{(2)} f||_{p,\psi^{1-p}} = ||H_1^{(2)}||_{p,\psi^{1-p}} \le K^{(2)}(\sigma)||f||_{p,\varphi}.$$

Thus, the operator $T_1^{(2)}$ is bounded satisfying

$$||T_1^{(2)}|| = \sup_{f(\neq\theta)\in L_{p,\varphi}(\mathbf{R})} \frac{||T_1^{(2)} f||_{p,\psi^{1-p}}}{||f||_{p,\varphi}} \le K^{(2)}(\sigma).$$

Since the constant factor $K^{(2)}(\sigma)$ in (125) is the best possible, we have

$$||T_1^{(2)}|| = K^{(2)}(\sigma) = \int_1^\infty (H(-t) + H(t)) t^{\sigma-1} dt. \tag{126}$$

Setting the formal inner product of $T_1^{(2)} f$ and g as

$$(T_1^{(2)} f, g) = \int_{-\infty}^{\infty} \left(\int_{E_y} H(xy) f(x) dx \right) g(y) dy,$$

we can rewrite (108) and (109) as follows:

$$(T_1^{(2)} f, g) < ||T_1^{(2)}|| \cdot ||f||_{p,\varphi} ||g||_{q,\psi}, \quad ||T_1^{(2)} f||_{p,\psi^{1-p}} < ||T_1^{(2)}|| \cdot ||f||_{p,\varphi}. \tag{127}$$

(d) In view of Corollary 14, for $f \in L_{p,\phi}(\mathbf{R})$,

$$H_2(y) := \int_{-\infty}^{\infty} K_\lambda(x, y) |f(x)| dx \quad (y \in \mathbf{R}),$$

by (101), we have

$$\begin{aligned} ||H_2||_{p,\psi^{1-p}} &= \left(\int_{-\infty}^{\infty} \psi^{1-p}(y) H_2^p(y) dy \right)^{\frac{1}{p}} \\ &< K_\lambda(\sigma) ||f||_{p,\phi} < \infty. \end{aligned} \tag{128}$$

Definition 14. We define the Yang–Hilbert-type integral operator with the homogeneous kernel in the whole plane

$$T_2 \ : \ L_{p,\phi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\phi}(\mathbf{R})$, there exists a unique representation

$$T_2 f = H_2 \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_2 f(y) = H_2(y),$$

for any $y \in \mathbf{R}$.

By (128), it follows that

$$||T_2 f||_{p,\psi^{1-p}} = ||H_2||_{p,\psi^{1-p}} \le K_\lambda(\sigma)||f||_{p,\phi}.$$

Hence, the operator T_2 is bounded satisfying

$$||T_2|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R})} \frac{||T_2 f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \le K_\lambda(\sigma).$$

Since the constant factor $K_\lambda(\sigma)$ in (128) is the best possible, we have

$$||T_2|| = K_\lambda(\sigma) = \int_{-\infty}^{\infty} K_\lambda(1,t)|t|^{\sigma-1}dt. \tag{129}$$

Setting the formal inner product of $T_2 f$ and g as

$$\begin{aligned}(T_2 f, g) &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} K_\lambda(x,y)f(x)dx\right)g(y)dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K_\lambda(x,y)f(x)g(y)dxdy,\end{aligned}$$

we can rewrite (100) and (101) as follows:

$$(T_2 f, g) < ||T_2|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \ \ ||T_2 f||_{p,\psi^{1-p}} < ||T_2|| \cdot ||f||_{p,\phi}.$$

(e) By Corollary 19, for $f \in L_{p,\phi}(\mathbf{R})$,

$$H_2^{(1)}(y) := \int_{F_y} K_\lambda(x,y)|f(x)|dx \ \ (y \in \mathbf{R}),$$

combined with (113), we obtain

$$||H_2^{(1)}||_{p,\psi^{1-p}} = \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)(H_2^{(1)}(y))^p dy\right)^{\frac{1}{p}} < K_\lambda^{(1)}(\sigma)||f||_{p,\phi} < \infty. \tag{130}$$

Definition 15. We define the Hardy-type integral operator of the first kind, with the homogeneous kernel in the whole plane

$$T_2^{(1)} : L_{p,\phi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\phi}(\mathbf{R})$, there exists a unique representation

$$T_2^{(1)} f = H_2^{(1)} \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_2^{(1)} f(y) = H_2^{(1)}(y),$$

for any $y \in \mathbf{R}$.

In view of (130), it follows that

$$||T_2^{(1)} f||_{p,\psi^{1-p}} = ||H_2^{(1)}||_{p,\psi^{1-p}} \le K_\lambda^{(1)}(\sigma)||f||_{p,\phi}.$$

Therefore, the operator $T_2^{(1)}$ is bounded satisfying

$$||T_2^{(1)}|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R})} \frac{||T_2^{(1)} f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \le K_\lambda^{(1)}(\sigma).$$

Since the constant factor $K_\lambda^{(1)}(\sigma)$ in (130) is the best possible, we have

$$||T_2^{(1)}|| = K_\lambda^{(1)}(\sigma) = \int_{-1}^{1} K_\lambda(1,t)|t|^{\sigma-1} dt. \tag{131}$$

Setting the formal inner product of $T_2^{(1)} f$ and g as

$$(T_2^{(1)} f, g) = \int_{-\infty}^{\infty} \left(\int_{F_y} K_\lambda(x,y) f(x) dx\right) g(y) dy,$$

we can rewrite (112) and (113) as follows:

$$(T_2^{(1)} f, g) < ||T_2^{(1)}|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \quad ||T_2^{(1)} f||_{p,\psi^{1-p}} < ||T_2^{(1)}|| \cdot ||f||_{p,\phi}.$$

(f) In view of Corollary 21, for $f \in L_{p,\phi}(\mathbf{R}_+)$,

$$H_2^{(2)}(y) := \int_{-|y|}^{|y|} K_\lambda(x,y)|f(x)|dx \ (y \in \mathbf{R}),$$

by (117), we have

$$||H_2^{(2)}||_{p,\psi^{1-p}} := \left(\int_{-\infty}^{\infty} \psi^{1-p}(y)(H_2^{(2)}(y))^p dy\right)^{\frac{1}{p}}$$
$$< K_\lambda^{(2)}(\sigma)||f||_{p,\phi} < \infty. \tag{132}$$

Definition 16. We define the Hardy-type integral operator of the second kind, with the homogeneous kernel in the whole plane

$$T_2^{(2)} : L_{p,\phi}(\mathbf{R}) \to L_{p,\psi^{1-p}}(\mathbf{R})$$

as follows:

For any $f \in L_{p,\phi}(\mathbf{R})$, there exists a unique representation

$$T_2^{(2)} f = H_2^{(2)} \in L_{p,\psi^{1-p}}(\mathbf{R}),$$

satisfying

$$T_2^{(2)} f(y) = H_2^{(2)}(y),$$

for any $y \in \mathbf{R}$.

By (132), it follows that

$$||T_2^{(2)} f||_{p,\psi^{1-p}} = ||H_2^{(2)}||_{p,\psi^{1-p}} \le K_\lambda^{(2)}(\sigma)||f||_{p,\phi}$$

and then the operator $T_2^{(2)}$ is bounded satisfying

$$||T_2^{(2)}|| = \sup_{f(\neq\theta)\in L_{p,\phi}(\mathbf{R})} \frac{||T_2^{(2)} f||_{p,\psi^{1-p}}}{||f||_{p,\phi}} \le K_\lambda^{(2)}(\sigma).$$

Since the constant factor $K_\lambda^{(2)}(\sigma)$ in (132) is the best possible, we have

$$||T_2^{(2)}|| = K_\lambda^{(2)}(\sigma) = \int_1^\infty (K_\lambda(1,-t) + K_\lambda(1,t))t^{\sigma-1}dt. \tag{133}$$

Setting the formal inner product of $T_2^{(2)} f$ and g as

$$(T_2^{(2)} f, g) = \int_{-\infty}^{\infty} \left(\int_{-|y|}^{|y|} K_\lambda(x,y) f(x) dx\right) g(y) dy,$$

we can rewrite (116) and (117) as follows:

$$(T_2^{(2)}f, g) < ||T_2^{(2)}|| \cdot ||f||_{p,\phi}||g||_{q,\psi}, \quad ||T_2^{(2)}f||_{p,\psi^{1-p}} < ||T_2^{(2)}|| \cdot ||f||_{p,\phi}.$$

Remark 8. (a) If $K_\lambda(x,y)$ is a symmetric function satisfying $K_\lambda(y,x) = K_\lambda(x,y)$, then by setting

$$H(t) =: K_\lambda(1,t)\arctan|t|^\beta \ (\beta \in \mathbf{R}),$$

and $\mu = \sigma = \frac{\lambda}{2}$ in (121), we obtain

$$\begin{aligned} ||T_1|| &= \int_{-\infty}^{\infty} H(t)|t|^{\sigma-1} = \int_{-\infty}^{\infty} K_\lambda(1,t)\arctan|t|^\beta |t|^{\sigma-1}dt \\ &= \frac{\pi}{4}K_\lambda\left(\frac{\lambda}{2}\right), \end{aligned} \tag{134}$$

where

$$K_\lambda\left(\frac{\lambda}{2}\right) = \int_{-\infty}^{\infty} K_\lambda(1,t)|t|^{\sigma-1}dt.$$

In fact, we obtain

$$\begin{aligned} &\int_0^\infty K_\lambda(1,t)(\arctan t^\beta)t^{\frac{\lambda}{2}-1}dt \\ &= \int_0^1 K_\lambda(1,t)(\arctan t^\beta)t^{\frac{\lambda}{2}-1}dt + \int_1^\infty K_\lambda(1,u)(\arctan u^\beta)u^{\frac{\lambda}{2}-1}du \\ &= \int_0^1 K_\lambda(1,t)(\arctan t^\beta)t^{\frac{\lambda}{2}-1}dt + \int_0^1 K_\lambda(t,1)(\arctan t^{-\beta})t^{\frac{\lambda}{2}-1}dt \\ &= \int_0^1 K_\lambda(1,t)(\arctan t^\beta + \arctan t^{-\beta})t^{\frac{\lambda}{2}-1}dt \\ &= \frac{\pi}{2}\int_0^1 K_\lambda(1,t)t^{\frac{\lambda}{2}-1}dt = \frac{\pi}{4}\int_0^\infty K_\lambda(1,t)t^{\frac{\lambda}{2}-1}dt. \end{aligned}$$

Similarly, we get

$$\begin{aligned} &\int_{-\infty}^0 K_\lambda(1,t)\arctan(-t)^\beta(-t)^{\sigma-1}dt \\ &= \int_0^\infty K_\lambda(1,-u)(\arctan u^\beta)u^{\sigma-1}du = \frac{\pi}{4}\int_0^\infty K_\lambda(1,-t)t^{\frac{\lambda}{2}-1}dt. \end{aligned}$$

Then we have

$$||T_1|| = \frac{\pi}{4}\int_0^\infty (K_\lambda(1,-t) + K_\lambda(1,t))t^{\frac{\lambda}{2}-1}dt = \frac{\pi}{4}K_\lambda\left(\frac{\lambda}{2}\right).$$

(b) If we replace $H(t)$ by

$$h(|t|^\gamma + t^\gamma \cos\alpha)(\gamma \in \{b; b = \frac{1}{2k-1}, 2k+1\ (k \in \mathbf{N})\}, \alpha \in (0, \pi))$$

in (121), where $h(t)$ is a non-negative measurable function in $\mathbf{R}_+$, satisfying

$$k\left(\frac{\sigma}{\gamma}\right) = \int_0^\infty h(t) t^{\frac{\sigma}{\gamma}-1} dt \in \mathbf{R}_+,$$

it follows that

$$\begin{aligned} ||T_1|| &= \int_{-\infty}^{\infty} h(|t|^\gamma + t^\gamma \cos\alpha)|t|^{\sigma-1} \\ &= \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}} + \left(\csc\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}}\right] k\left(\frac{\sigma}{\gamma}\right), \end{aligned} \tag{135}$$

In particular, setting $h(t) = k_\lambda(1, t)$ $(t \in \mathbf{R}_+)$, it follows that

$$\begin{aligned} ||T_1|| &= \int_{-\infty}^{\infty} k_\lambda(1, |t|^\gamma + t^\gamma \cos\alpha)|t|^{\sigma-1} \\ &= \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}} + \left(\csc\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}}\right] k_\lambda\left(\frac{\sigma}{\gamma}\right), \end{aligned} \tag{136}$$

where

$$k_\lambda\left(\frac{\sigma}{\gamma}\right) = \int_0^\infty k_\lambda(1, t) t^{\frac{\sigma}{\gamma}-1} dt \in \mathbf{R}_+.$$

In fact, setting $u = t^\gamma(1 + \cos\alpha)$, we get

$$\begin{aligned} &\int_0^\infty h(|t|^\gamma + t^\gamma \cos\alpha)|t|^{\sigma-1} dt \\ &= \int_0^\infty h(t^\gamma(1 + \cos\alpha)) t^{\sigma-1} dt = \frac{1}{\gamma 2^{\sigma/\gamma}} \left(\sec\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}} k\left(\frac{\sigma}{\gamma}\right); \end{aligned}$$

Moreover, setting $u = -t$, we have

$$\begin{aligned} &\int_{-\infty}^0 h(|t|^\gamma + t^\gamma \cos\alpha)|t|^{\sigma-1} dt = \int_{-\infty}^0 h(-t^\gamma(1 - \cos\alpha))(-t)^{\sigma-1} dt \\ &= \int_0^\infty h(u^\gamma(1 - \cos\alpha)) u^{\sigma-1} du = \frac{1}{\gamma 2^{\sigma/\gamma}} \left(\csc\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}} k\left(\frac{\sigma}{\gamma}\right), \end{aligned}$$

and then (135) follows.

(c) Replacing $K_\lambda(1,t)$ by $k_\lambda(1,|t|^\gamma + t^\gamma\cos\alpha)$ in (134), where $k_\lambda(x,y)$ is a homogeneous function in $\mathbf{R}_+$, satisfying

$$k_\lambda\left(\frac{\sigma}{\gamma}\right) = \int_0^\infty k_\lambda(1,t)t^{\frac{\sigma}{\gamma}-1}dt \in \mathbf{R}_+,$$

in view of (136), we obtain

$$\begin{aligned}||T_1|| &= \int_{-\infty}^\infty k_\lambda(1,|t|^\gamma + t^\gamma\cos\alpha)\arctan|t|^\beta|t|^{\frac{\lambda}{2}-1}dt \\ &= \frac{\pi}{\gamma 2^{2+\sigma/\gamma}}\left[\left(\sec\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}} + \left(\csc\frac{\alpha}{2}\right)^{\frac{2\alpha}{\gamma}}\right]k_\lambda\left(\frac{\lambda}{2\gamma}\right). \end{aligned} \tag{137}$$

3.6 Some Examples

Example 5. (a) Set

$$H(t) = K_\lambda(1,t) = \frac{1}{|1+t|^\lambda}\ (\mu,\sigma > 0, \mu+\sigma = \lambda < 1).$$

Then we have the kernels

$$H(xy) = \frac{1}{|1+xy|^\lambda},\ \ K_\lambda(x,y) = \frac{1}{|x+y|^\lambda}$$

and obtain the constant factors

$$\begin{aligned}K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^\infty \frac{|t|^{\sigma-1}}{|1+t|^\lambda}dt \\ &= \int_0^\infty \frac{t^{\sigma-1}}{|1-t|^\lambda}dt + \int_0^\infty \frac{t^{\sigma-1}}{(1+t)^\lambda}dt \\ &= B(1-\lambda,\sigma) + B(1-\lambda,\mu) + B(\mu,\sigma) \in \mathbf{R}_+.\end{aligned}$$

By (121) and (129), we have (cf. [16])

$$||T_1|| = ||T_2|| = B(1-\lambda,\sigma) + B(1-\lambda,\mu) + B(\mu,\sigma). \tag{138}$$

(b) Set

$$H(t) = K_\lambda(1,t) = \frac{|\ln|t|^\beta|}{|1+t|^\lambda},$$

where $\beta > -1, \mu, \sigma > 0, \mu + \sigma = \lambda < 1 + \beta$. Then we have the kernels

$$H(xy) = \frac{|\ln |xy|^{\beta}|}{|1 + xy|^{\lambda}}, \quad K_{\lambda}(x, y) = \frac{|\ln |x/y|^{\beta}|}{|x + y|^{\lambda}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_{\lambda}(\sigma) &= \int_{-\infty}^{\infty} \frac{|\ln |t|^{\beta}||t|^{\sigma-1}}{|1 + t|^{\lambda}} dt \\ &= \int_{0}^{\infty} \frac{|\ln t^{\beta}| t^{\sigma-1}}{|1 - t|^{\lambda}} dt + \int_{0}^{\infty} \frac{|\ln t^{\beta}| t^{\sigma-1}}{(1 + t)^{\lambda}} dt \\ &= \int_{0}^{1} (-\ln t)^{\beta} \left[\frac{1}{(1 - t)^{\lambda}} + \frac{1}{(1 + t)^{\lambda}} \right] (t^{\mu-1} + t^{\sigma-1}) dt. \end{aligned}$$

There exists a constant $\delta_0 > 0$, such that $\sigma > \delta_0, \ \mu > \delta_0$. Since

$$\lim_{t \to 0^+} t^{\delta_0} \frac{(-\ln t)^{\beta}}{(1 - t)^{\beta}} = 0, \quad \lim_{t \to 1^-} t^{\delta_0} \frac{(-\ln t)^{\beta}}{(1 - t)^{\beta}} = 1,$$

there exists a constant $L > 0$, such that

$$t^{\delta_0} \frac{(-\ln t)^{\beta}}{(1 - t)^{\beta}} \le L \ (0 < t < 1)$$

and

$$\begin{aligned} 0 < K(\sigma) &= \int_{0}^{1} (-\ln t)^{\beta} \left[\frac{1}{(1 - t)^{\lambda}} + \frac{1}{(1 + t)^{\lambda}} \right] (t^{\mu-1} + t^{\sigma-1}) dt \\ &\le 2 \int_{0}^{1} \frac{(-\ln t)^{\beta}}{(1 - t)^{\lambda}} (t^{\mu-1} + t^{\sigma-1}) dt \le 2L \int_{0}^{1} \frac{t^{\mu-\delta_0-1} + t^{\sigma-\delta_0-1}}{(1 - t)^{\lambda-\beta}} dt \\ &= 2L[B(\beta + 1 - \lambda, \mu - \delta_0) + B(\beta + 1 - \lambda, \sigma - \delta_0)] < \infty. \end{aligned}$$

Therefore, $K(\sigma) = K_{\lambda}(\sigma) \in \mathbf{R}_{+}$.
Since

$$\binom{-\lambda}{2k} = \binom{\lambda + 2k - 1}{2k} > 0,$$

then by Lebesgue's term by term theorem, it follows that

$$\begin{aligned} K(\sigma) &= \int_0^1 (-\ln t)^\beta \sum_{k=0}^{\infty} \binom{-\lambda}{k} [(-1)^k + 1](t^{k+\mu-1} + t^{k+\sigma-1})dt \\ &= 2\int_0^1 (-\ln t)^\beta \sum_{k=0}^{\infty} \binom{-\lambda}{2k} (t^{k+\mu-1} + t^{k+\sigma-1})dt \\ &= 2\sum_{k=0}^{\infty} \binom{-\lambda}{2k} \int_0^1 (-\ln t)^\beta (t^{2k+\mu-1} + t^{2k+\sigma-1})dt \\ &= 2\Gamma(\beta+1) \sum_{k=0}^{\infty} \binom{-\lambda}{2k} \left[\frac{1}{(2k+\mu)^\beta} + \frac{1}{(2k+\sigma)^\beta} \right]. \end{aligned}$$

In view of (121) and (129), we have

$$||T_1|| = ||T_2|| = 2\Gamma(\beta+1) \sum_{k=0}^{\infty} \binom{-\lambda}{2k} \left[\frac{1}{(2k+\mu)^\beta} + \frac{1}{(2k+\sigma)^\beta} \right]. \tag{139}$$

(c) Set

$$H(t) = K_\lambda(1,t) = \frac{(\max\{1,|t|\})^\beta}{|1+t|^{\lambda+\beta}} \ (\beta < 1, \mu, \sigma > 0, \mu + \sigma = \lambda < 1 - \beta).$$

Then we have the kernels

$$H(xy) = \frac{(\max\{1,|xy|\})^\beta}{|1+xy|^{\lambda+\beta}}, \quad K_\lambda(x,y) = \frac{(\max\{|x|,|y|\})^\beta}{|x+y|^{\lambda+\beta}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty} \frac{(\max\{1,|t|\})^\beta}{|1+t|^{\lambda+\beta}} |t|^{\sigma-1} dt \\ &= \int_0^{\infty} \frac{(\max\{1,t\})^\beta}{|1-t|^{\lambda+\beta}} t^{\sigma-1} dt + \int_0^{\infty} \frac{(\max\{1,t\})^\beta}{(1+t)^{\lambda+\beta}} t^{\sigma-1} dt \\ &= \int_0^1 \left[\frac{1}{(1-t)^{\lambda+\beta}} + \frac{1}{(1+t)^{\lambda+\beta}} \right] (t^{\mu-1} + t^{\sigma-1}) dt \\ &= B(1-\lambda-\beta,\mu) + B(1-\lambda-\beta,\sigma) + \int_0^1 \frac{t^{\mu-1} + t^{\sigma-1}}{(1+t)^{\lambda+\beta}} dt \in \mathbf{R}_+. \end{aligned}$$

By Taylor's formula, we still can obtain

$$\int_0^1 \frac{t^{\mu-1}+t^{\sigma-1}}{(1+t)^{\lambda+\beta}}dt = \int_0^1 \sum_{k=0}^{\infty} \binom{-\lambda-\beta}{k} (t^{k+\mu-1}+t^{k+\sigma-1})dt$$

$$= \int_0^1 \sum_{k=0}^{\infty} (-1)^k \binom{\lambda+\beta+k-1}{k} (t^{k+\mu-1}+t^{k+\sigma-1})dt$$

$$= \int_0^1 \sum_{k=0}^{\infty} \left[\binom{\lambda+\beta+2k-1}{2k} - \binom{\lambda+\beta+2k}{2k+1} t \right] (t^{2k+\mu-1}+t^{2k+\sigma-1})dt.$$

Since we find

$$\binom{\lambda+\beta+2k-1}{2k} - \binom{\lambda+\beta+2k}{2k+1} t = \binom{\lambda+\beta+2k-1}{2k} - \frac{(\lambda+\beta+2k)t}{2k+1} \binom{\lambda+\beta+2k-1}{2k}$$

$$= \left[1 - \frac{(\lambda+\beta+2k)t}{2k+1} \right] \binom{\lambda+\beta+2k-1}{2k},$$

there exists a number $k_0 \in \mathbf{N}_0 = \mathbf{N} \cup \{0\}$, such that $\lambda+\beta+2k_0 > 0$, and for any $s \in \mathbf{N}$,

$$\binom{\lambda+\beta+2(k_0+s)-1}{2(k_0+s)} - \binom{\lambda+\beta+2(k_0+s)+1}{2(k_0+s)+1} t$$

$$= \left[1 - \frac{(\lambda+\beta+2k_0+2s)t}{2(k_0+s)+1} \right] \binom{\lambda+\beta+2(k_0+s)-1}{2(k_0+s)}$$

$$= \left[1 - \frac{(\lambda+\beta+2k_0+2s)t}{2(k_0+s)+1} \right]$$

$$\times \frac{\lambda+\beta+2k_0+2s-1}{2k_0+2s} \cdots \frac{\lambda+\beta+2k_0}{2k_0+1} \binom{\lambda+\beta+2k_0-1}{2k_0}.$$

For $t \in (0,1]$, we get

$$1 - \frac{(\lambda+\beta+2k_0+2s)t}{2(k_0+s)+1} \geq 1 - \frac{\lambda+\beta+2k_0+2s}{2(k_0+s)+1}$$

$$= \frac{1-\lambda-\beta}{2(k_0+s)+1} > 0.$$

Then it follows that for any $s \in \mathbf{N}$,

$$\operatorname{sgn}\left(\binom{\lambda+\beta+2(k_0+s)-1}{2(k_0+s)} - \binom{\lambda+\beta+2(k_0+s)+1}{2(k_0+s)+1} t \right) = \operatorname{sgn}\binom{\lambda+\beta+2k_0-1}{2k_0}.$$

Hence by Lebesgue term by term integration theorem, we have

$$\begin{aligned} K(\sigma) &= B(1-\lambda-\beta,\mu)+B(1-\lambda-\beta,\sigma) \\ &\quad +\sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\int_0^1 (t^{k+\mu-1}+t^{k+\sigma-1})dt \\ &= B(1-\lambda-\beta,\mu)+B(1-\lambda-\beta,\sigma) \\ &\quad +\sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\left(\frac{1}{k+\mu}+\frac{1}{k+\sigma}\right). \end{aligned}$$

In view of (121) and (129), we have

$$\begin{aligned} ||T_1|| = ||T_2|| &= B(1-\lambda-\beta,\beta+\mu)+B(1-\lambda-\beta,\beta+\sigma) \\ &\quad +\sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\left(\frac{1}{k+\mu}+\frac{1}{k+\sigma}\right). \end{aligned} \tag{140}$$

For (a)–(c), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorem 5–8. Setting $\delta_0 = \frac{\sigma}{2} > 0$, we also obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorem 5–8.

(d) Set

$$H(t) = K_\lambda(1,t) = \frac{(\min\{1,|t|\})^\beta}{|1+t|^{\lambda+\beta}},$$

with $\beta > -1, \mu, \sigma > -\beta, \mu+\sigma = \lambda < 1-\beta$. Then we have the kernels

$$H(xy) = \frac{(\min\{1,|xy|\})^\beta}{|1+xy|^{\lambda+\beta}},\quad K_\lambda(x,y) = \frac{(\min\{|x|,|y|\})^\beta}{|x+y|^{\lambda+\beta}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty}\frac{(\min\{1,|t|\})^\beta}{|1+t|^{\lambda+\beta}}|t|^{\sigma-1}dt \\ &= \int_0^\infty \frac{(\min\{1,t\})^\beta}{|1-t|^{\lambda+\beta}}t^{\sigma-1}dt+\int_0^\infty \frac{(\min\{1,t\})^\beta}{(1+t)^{\lambda+\beta}}t^{\sigma-1}dt \\ &= \int_0^1\left[\frac{1}{(1-t)^{\lambda+\beta}}+\frac{1}{(1+t)^{\lambda+\beta}}\right](t^{\beta+\mu-1}+t^{\beta+\sigma-1})dt \\ &= B(1-\lambda-\beta,\beta+\mu)+B(1-\lambda-\beta,\beta+\sigma) \\ &\quad +\int_0^1 \frac{t^{\beta+\mu-1}+t^{\beta+\sigma-1}}{(1+t)^{\lambda+\beta}}dt \in \mathbf{R}_+. \end{aligned}$$

Similarly to the method followed in (c), we find

$$\int_0^1 \frac{t^{\beta+\mu-1}+t^{\beta+\sigma-1}}{(1+t)^{\lambda+\beta}}dt = \sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\int_0^1 (t^{k+\beta+\mu-1}+t^{k+\beta+\sigma-1})dt$$
$$= \sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\left(\frac{1}{k+\beta+\mu}+\frac{1}{k+\beta+\sigma}\right).$$

By the above results, (121) and (129), we have

$$||T_1|| = ||T_2|| = B(1-\lambda-\beta,\beta+\mu)+B(1-\lambda-\beta,\beta+\sigma)$$
$$+\sum_{k=0}^{\infty}\binom{-\lambda-\beta}{k}\left(\frac{1}{k+\beta+\mu}+\frac{1}{k+\beta+\sigma}\right). \tag{141}$$

Then in (d), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorem 5–8. Setting $\delta_0 = \frac{\sigma+\beta}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorem 5–8.

Example 6. Set

$$H(t) = K_2(1,t) = \frac{1}{1+2bt+(ct)^2}\ (|b|<|c|, \mu=\sigma=1).$$

Then we have the kernels

$$H(xy) = \frac{1}{1+2bxy+(cxy)^2},\ \ K_\lambda(x,y) = \frac{1}{x^2+2bxy+(cy)^2}$$

and obtain the constant factors

$$K(\sigma) = K_2(\sigma) = \int_{-\infty}^{\infty}\frac{1}{1+2bt+(ct)^2}dt$$
$$= \frac{2}{\sqrt{4c^2-4b^2}}\arctan\frac{2c^2t+2b}{\sqrt{4c^2-4b^2}}\Big|_{-\infty}^{\infty} = \frac{\pi}{\sqrt{c^2-b^2}} \in \mathbf{R}_+.$$

In view of (121) and (129), we have (cf. [13])

$$||T_1|| = ||T_2|| = \frac{\pi}{\sqrt{c^2-b^2}}.$$

In particular, for $c=1,\ b=\cos\alpha\ (0<\alpha<\pi)$, we have

$$||T_1|| = ||T_2|| = \frac{\pi}{\sin\alpha}.$$

Example 7. (a) Set

$$H(t) = K_2(1,t) = \min_{i\in\{1,2\}} \frac{1}{1+2t\cos\alpha_i + t^2},$$

with $0 < \alpha_1 \le \alpha_2 < \pi, 0 < \sigma < 2$. Then we have the kernels

$$H(xy) = \min_{i\in\{1,2\}} \frac{1}{1+2xy\cos\alpha_i + (xy)^2},\ \ K_2(x,y) = \min_{i\in\{1,2\}} \frac{1}{x^2+2xy\cos\alpha_i + y^2}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_2(\sigma) &= \int_{-\infty}^{\infty} \min_{i\in\{1,2\}} \frac{1}{1+2t\cos\alpha_i + t^2}|t|^{\sigma-1}dt \\ &= \int_0^{\infty} \min_{i\in\{1,2\}} \frac{t^{\sigma-1}}{1+2t\cos\alpha_i + t^2}dt + \int_0^{\infty} \min_{i\in\{1,2\}} \frac{t^{\sigma-1}}{1-2t\cos\alpha_i + t^2}dt \\ &= \int_0^{\infty} \frac{t^{\sigma-1}}{1+2t\cos\alpha_1 + t^2}dt + \int_0^{\infty} \frac{t^{\sigma-1}}{1+2t\cos(\pi-\alpha_2) + t^2}dt. \end{aligned}$$

Set

$$f(z) = \frac{1}{1+2z\cos\alpha_1 + z^2}.$$

Then

$$z_1 = -e^{i\alpha_1},\ \ z_2 = -e^{-i\alpha_1}$$

are the poles of order 1. Setting

$$\varphi_1(z) = (z-z_1)f(z) = \frac{1}{z-z_2},\ \varphi_2(z) = (z-z_2)f(z) = \frac{1}{z-z_1},$$

by (63), we have

$$\begin{aligned} \int_0^{\infty} \frac{t^{\sigma-1}}{1+2t\cos\alpha_1 + t^2}dt &= \int_0^{\infty} f(t)t^{\sigma-1}dt \\ &= \frac{\pi}{\sin\pi\sigma}[(-z_1)^{\sigma-1}\varphi_1(z_1) + (-z_2)^{\sigma-1}\varphi_2(z_2)] \\ &= \frac{\pi}{\sin\pi\sigma}\left[\frac{e^{i\alpha_1(\sigma-1)}}{-e^{i\alpha_1}+e^{-i\alpha_1}} + \frac{e^{-i\alpha_1(\sigma-1)}}{-e^{-i\alpha_1}+e^{i\alpha_1}}\right] \\ &= \frac{\pi\sin\alpha_1(1-\sigma)}{\sin\pi\sigma\sin\alpha_1}. \end{aligned}$$

Similarly, it follows that

$$\int_0^\infty \frac{t^{\sigma-1}dt}{1+2t\cos(\pi-\alpha_2)+t^2} = \frac{\pi\sin(\pi-\alpha_2)(1-\sigma)}{\sin\pi\sigma\sin(\pi-\alpha_2)} = \frac{\pi\sin(\pi-\alpha_2)(1-\sigma)}{\sin\pi\sigma\sin\alpha_2},$$

and then

$$K(\sigma) = \frac{\pi}{\sin\pi\sigma}\left[\frac{\sin\alpha_1(1-\sigma)}{\sin\alpha_1} + \frac{\sin(\pi-\alpha_2)(1-\sigma)}{\sin\alpha_2}\right] \in \mathbf{R}_+.$$

In view of (121) and (129), we have (cf. [21])

$$||T_1|| = ||T_2|| = \frac{\pi}{\sin\pi\sigma}\left[\frac{\sin\alpha_1(1-\sigma)}{\sin\alpha_1} + \frac{\sin(\pi-\alpha_2)(1-\sigma)}{\sin\alpha_2}\right]. \tag{142}$$

Then in (a), we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 5–8. Setting $\delta_0 = \frac{\sigma}{2} > 0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 5–8.

In particular, if $\alpha_1 = \alpha_2 = \alpha \in (0,\pi)$, then

$$H(t) = K_2(1,t) = \frac{1}{1+2t\cos\alpha+t^2}$$

and

$$||T_1|| = ||T_2|| = \frac{\pi}{\sin\pi\sigma\sin\alpha}[\sin\alpha(1-\sigma)+\sin(\pi-\alpha)(1-\sigma)].$$

(b) Set

$$H(t) = K_0(1,t) = \min_{i\in\{1,2\}}\frac{\min\{1,|t|\}}{\sqrt{1+2t\cos\alpha_i+t^2}},$$

with $0 < \alpha_1 \le \alpha_2 < \pi, \sigma = \mu = 0$. Then we have the kernels

$$H(xy) = \min_{i\in\{1,2\}}\frac{\min\{1,|xy|\}}{\sqrt{1+2xy\cos\alpha_i+(xy)^2}},$$

$$K_0(x,y) = \min_{i\in\{1,2\}}\frac{\min\{|x|,|y|\}}{\sqrt{x^2+2xy\cos\alpha_i+y^2}}$$

and obtain the constant factors

$$\begin{aligned}K(0) = K_0(0) &= \int_{-\infty}^{\infty} \min_{i\in\{1,2\}} \frac{\min\{1,|t|\}}{\sqrt{1+2t\cos\alpha_i+t^2}}|t|^{-1}dt \\ &= \int_0^{\infty} \min_{i\in\{1,2\}} \frac{\min\{1,t\}t^{-1}}{\sqrt{1+2t\cos\alpha_i+t^2}}dt \\ &\quad + \int_0^{\infty} \min_{i\in\{1,2\}} \frac{\min\{1,t\}t^{-1}}{\sqrt{1-2t\cos\alpha_i+t^2}}dt \\ &= \int_0^{\infty} \frac{\min\{1,t\}t^{-1}dt}{\sqrt{1+2t\cos\alpha_1+t^2}} + \int_0^{\infty} \frac{\min\{1,t\}t^{-1}dt}{\sqrt{1+2t\cos(\pi-\alpha_2)+t^2}} \\ &= 2\left[\int_0^1 \frac{dt}{\sqrt{1+2t\cos\alpha_1+t^2}} + \int_0^1 \frac{dt}{\sqrt{1+2t\cos(\pi-\alpha_2)+t^2}}\right].\end{aligned}$$

We get

$$\begin{aligned}&\int_0^1 \frac{dt}{\sqrt{1+2t\cos\alpha_1+t^2}} \\ &\quad = \ln\left(2t+2\cos\alpha_1+2\sqrt{1+2t\cos\alpha_1+t^2}\right)|_0^1 = \ln\left(1+\sec\frac{\alpha_1}{2}\right),\end{aligned}$$

and by the same way,

$$\int_0^1 \frac{dt}{\sqrt{1+2t\cos(\pi-\alpha_2)+t^2}} = \ln\left(1+\sec\frac{\pi-\alpha_2}{2}\right) = \ln\left(1+\csc\frac{\alpha_2}{2}\right).$$

Then it follows that

$$K(0) = K_0(0) = 2\ln\left(1+\sec\frac{\alpha_1}{2}\right)\left(1+\csc\frac{\alpha_2}{2}\right).$$

By (121) and (129), we have (cf. [20])

$$||T_1|| = ||T_2|| = 2\ln\left(1+\sec\frac{\alpha_1}{2}\right)\left(1+\csc\frac{\alpha_2}{2}\right). \tag{143}$$

In particular, if $\alpha_1 = \alpha_2 = \alpha \in (0,\pi)$, then

$$H(t) = K_0(1,t) = \frac{\min\{1,|t|\}}{\sqrt{1+2t\cos\alpha+t^2}}$$

and

$$||T_1|| = ||T_2|| = 2\ln\left(1+\sec\frac{\alpha}{2}\right)\left(1+\csc\frac{\alpha}{2}\right).$$

Example 8. Set

$$H(t) = K_0(1,t) = \Big| \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} \Big|,$$

with $0 < \alpha_1 \le \alpha_2 < \pi,\ \mu = -\sigma \in (0,1)$. Then we have the kernels

$$H(xy) = \Big| \ln \frac{1+2xy\cos\alpha_1+(xy)^2}{1+2xy\cos\alpha_2+(xy)^2} \Big|,$$

$$K_0(x,y) = \Big| \ln \frac{x^2+2xy\cos\alpha_1+y^2}{x^2+2xy\cos\alpha_2+y^2} \Big|$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_0(\sigma) &= \int_{-\infty}^{\infty} \Big| \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} \Big| \cdot |t|^{\sigma-1} dt \\ &= \int_0^{\infty} t^{\sigma-1} \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} dt \\ &\quad + \int_0^{\infty} t^{\sigma-1} \ln \frac{1-2t\cos\alpha_2+t^2}{1-2t\cos\alpha_1+t^2} dt \\ &= \int_0^{\infty} t^{\sigma-1} \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} dt \\ &\quad + \int_0^{\infty} t^{\sigma-1} \ln \frac{1+2t\cos(\pi-\alpha_2)+t^2}{1+2t\cos(\pi-\alpha_1)+t^2} dt. \end{aligned}$$

We find

$$\begin{aligned} I_1 := \int_0^{\infty} t^{\sigma-1} \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} dt &= \frac{1}{\sigma} \int_0^{\infty} \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} dt^{\sigma} \\ = \frac{1}{\sigma} \Big[t^{\sigma} \ln \frac{1+2t\cos\alpha_1+t^2}{1+2t\cos\alpha_2+t^2} \Big|_0^{\infty} & \\ -2\int_0^{\infty} t^{\sigma} \Big(\frac{\cos\alpha_1+t}{1+2t\cos\alpha_1+t^2} - \frac{\cos\alpha_2+t}{1+2t\cos\alpha_2+t^2} \Big) dt \Big] & \\ = \frac{-2}{\sigma} \Big[\int_0^{\infty} \frac{(\cos\alpha_1+t)\, t^{\sigma}}{1+2t\cos\alpha_1+t^2} dt - \int_0^{\infty} \frac{(\cos\alpha_2+t)\, t^{\sigma}}{1+2t\cos\alpha_2+t^2} dt \Big] & \\ = \frac{-2}{\sigma} (I_1^{(1)} - I_1^{(2)}), & \\ I_1^{(i)} = \int_0^{\infty} \frac{(\cos\alpha_i+t)t^{(\sigma+1)-1}}{1+2t\cos\alpha_i+t^2} dt \ (i=1,2). & \end{aligned}$$

By (63), we have

$$I_1^{(i)} = \frac{\pi}{\sin \pi(\sigma+1)} \left(e^{i\alpha_i \sigma} \frac{\cos \alpha_i - e^{i\alpha_i}}{-e^{i\alpha_i} + e^{-i\alpha_i}} + e^{-i\alpha_i \sigma} \frac{\cos \alpha_i - e^{-i\alpha_i}}{-e^{-i\alpha_i} + e^{i\alpha_i}} \right)$$
$$= \frac{\pi \cos \alpha_i \sigma}{\sin \pi(\sigma+1)} \ (i = 1, 2),$$

and then

$$I_1 = \frac{-2}{\sigma} \left[\frac{\pi \cos \alpha_1 \sigma}{\sin \pi(\sigma+1)} - \frac{\pi \cos \alpha_2 \sigma}{\sin \pi(\sigma+1)} \right]$$
$$= \frac{-2\pi(\cos \alpha_1 \sigma - \cos \alpha_2 \sigma)}{\sigma \sin \pi(\sigma+1)}$$
$$= \frac{4\pi}{\sigma \sin \pi(\sigma+1)} \sin \frac{\alpha_1 + \alpha_2}{2} \sigma \sin \frac{\alpha_1 - \alpha_2}{2} \sigma.$$

Similarly, we have

$$I_2 := \int_0^\infty t^{\sigma-1} \ln \frac{1 + 2t \cos(\pi - \alpha_2) + t^2}{1 + 2t \cos(\pi - \alpha_1) + t^2} dt$$
$$= \frac{4\pi}{\sigma \sin \pi(\sigma+1)} \sin \left(\pi - \frac{\alpha_1 + \alpha_2}{2} \right) \sigma \sin \frac{\alpha_1 - \alpha_2}{2} \sigma,$$

and then

$$K(\sigma) = K_0(\sigma) = \frac{4\pi}{\sigma \sin \pi(\sigma+1)} \sin \frac{\alpha_1 + \alpha_2}{2} \sigma$$
$$\times \left[\sin \frac{\alpha_1 - \alpha_2}{2} \sigma + \sin \left(\pi - \frac{\alpha_1 + \alpha_2}{2} \right) \sigma \right]$$
$$= \frac{-4\pi \sin \frac{\sigma}{2}(\alpha_1 - \alpha_2)}{\sigma \cos \pi(\sigma/2)} \cos \frac{\sigma}{2}(\alpha_1 + \alpha_2 - \pi) \in \mathbf{R}_+.$$

In view of (121) and (129), we have (cf. [18])

$$||T_1|| = ||T_2|| = \frac{-4\pi \sin \frac{\sigma}{2}(\alpha_1 - \alpha_2)}{\sigma \cos \pi(\sigma/2)} \cos \frac{\sigma}{2}(\alpha_1 + \alpha_2 - \pi). \tag{144}$$

Then we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 5–8. Setting $\delta_0 = \frac{-\sigma}{2} > 0$, we still can obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 5–8.

Remark 9. Since $K(0^-) = 2\pi(\alpha_2 - \alpha_1) \in \mathbf{R}_+$, then (144) is valid for $\sigma \in (-1, 0]$.

Example 9. (a) For

$$\gamma \in \left\{ a;\ a = \frac{1}{2k-1}, 2k+1 (k \in \mathbf{N}) \right\},$$

we set

$$H(t) = K_\lambda(1,t) = \min_{i\in\{1,2\}} \frac{1}{(1 + t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma}},$$

where $0 < \alpha_1 \le \alpha_2 < \pi,\ \mu,\ \sigma > 0,\ \mu + \sigma = \lambda$. Then we have the kernels

$$H(xy) = \min_{i\in\{1,2\}} \frac{1}{(1 + (xy)^\gamma \cos\alpha_i + |xy|^\gamma)^{\lambda/\gamma}},$$

$$K_\lambda(x,y) = \min_{i\in\{1,2\}} \frac{1}{(|x|^\gamma + y^\gamma \mathrm{sgn}(x) \cos\alpha_i + |y|^\gamma)^{\lambda/\gamma}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty} \min_{i\in\{1,2\}} \frac{1}{(1 + t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma}} |t|^{\sigma-1} dt \\ &= \int_0^\infty \frac{t^{\sigma-1} dt}{[1 + t^\gamma (1 + \cos\alpha_1)]^{\lambda/\gamma}} + \int_0^\infty \frac{t^{\sigma-1} dt}{[1 + t^\gamma (1 - \cos\alpha_2)]^{\lambda/\gamma}} \\ &= \frac{1}{\gamma} \left[\frac{1}{(1+\cos\alpha_1)^{\lambda/\gamma}} + \frac{1}{(1-\cos\alpha_2)^{\lambda/\gamma}} \right] \int_0^\infty \frac{u^{(\sigma/\gamma)-1} du}{(1+u)^{\lambda/\gamma}} \\ &= \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc\frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] B\left(\frac{\mu}{\gamma}, \frac{\sigma}{\gamma} \right). \end{aligned}$$

In view of (121) and (129), we have

$$||T_1|| = ||T_2|| = \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc\frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] B\left(\frac{\mu}{\gamma}, \frac{\sigma}{\gamma} \right). \tag{145}$$

In particular, if $\alpha_1 = \alpha_2 = \alpha \in (0, \pi)$, then it follows that

$$H(t) = K_\lambda(1,t) = \frac{1}{(1 + t^\gamma \cos\alpha + |t|^\gamma)^{\lambda/\gamma}},$$

and

$$||T_1|| = ||T_2|| = \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc\frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} \right] B\left(\frac{\mu}{\gamma}, \frac{\sigma}{\gamma} \right).$$

(b) For

$$\gamma \in \left\{ a; a = \frac{1}{2k-1}, 2k+1 (k \in \mathbf{N}) \right\},$$

we set

$$H(t) = K_\lambda(1,t) = \min_{i\in\{1,2\}} \frac{1}{|1 - t^\gamma \cos\alpha_i - |t|^\gamma|^{\lambda/\gamma}},$$

where $0 < \alpha_1 \le \alpha_2 < \pi$, $\mu, \sigma > 0$, $\mu + \sigma = \lambda < \gamma$. Then we have the kernels

$$H(xy) = \min_{i\in\{1,2\}} \frac{1}{|1 - (xy)^\gamma \cos\alpha_i - |xy|^\gamma|^{\lambda/\gamma}},$$

$$K_\lambda(x,y) = \min_{i\in\{1,2\}} \frac{1}{||x|^\gamma - y^\gamma \mathrm{sgn}(x) \cos\alpha_i - |y|^\gamma|^{\lambda/\gamma}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty} \min_{i\in\{1,2\}} \frac{1}{|1 - t^\gamma \cos\alpha_i - |t|^\gamma|^{\lambda/\gamma}} |t|^{\sigma-1} dt \\ &= \int_0^\infty \min_{i\in\{1,2\}} \frac{1}{|1 - t^\gamma (1 + \cos\alpha_i)|^{\lambda/\gamma}} t^{\sigma-1} dt \\ &\quad + \int_0^\infty \min_{i\in\{1,2\}} \frac{1}{|1 - t^\gamma (1 - \cos\alpha_i)|^{\lambda/\gamma}} t^{\sigma-1} dt \\ &= \frac{1}{\gamma} \left[\int_0^\infty \frac{1}{(1+\cos\alpha_1)^{\lambda/\gamma}} \int_0^\infty \frac{u^{(\sigma/\gamma)-1} du}{|1-u|^{\lambda/\gamma}} \right] \\ &\quad + \frac{1}{\gamma} \left[\int_0^\infty \frac{1}{(1-\cos\alpha_2)^{\lambda/\gamma}} \int_0^\infty \frac{u^{(\sigma/\gamma)-1} du}{|1-u|^{\lambda/\gamma}} \right] \\ &= \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc\frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \\ &\quad \times \left[B\left(1 - \frac{\lambda}{\gamma}, \frac{\mu}{\gamma} \right) + B\left(1 - \frac{\lambda}{\gamma}, \frac{\sigma}{\gamma} \right) \right]. \end{aligned}$$

In view of (121) and (129), we have

$$\begin{aligned} ||T_1|| = ||T_2|| &= \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec\frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc\frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \\ &\quad \times \left[B\left(1 - \frac{\lambda}{\gamma}, \frac{\mu}{\gamma} \right) + B\left(1 - \frac{\lambda}{\gamma}, \frac{\sigma}{\gamma} \right) \right]. \end{aligned} \tag{146}$$

In particular, if $\alpha_1 = \alpha_2 = \alpha \in (0, \pi)$, then we have

$$H(t) = K_\lambda(1, t) = \frac{1}{|1 - t^\gamma \cos\alpha - |t|^\gamma|^{\lambda/\gamma}},$$

and

$$||T_1|| = ||T_2|| = \frac{1}{\gamma 2^{\sigma/\gamma}} \left[\left(\sec \frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc \frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \times \left[B\left(1 - \frac{\lambda}{\gamma}, \frac{\mu}{\gamma}\right) + B\left(1 - \frac{\lambda}{\gamma}, \frac{\sigma}{\gamma}\right) \right].$$

(c) For

$$\gamma \in \left\{ a; a = \frac{1}{2k-1}, 2k+1 \ (k \in \mathbf{N}) \right\},$$

we set

$$H(t) = K_\lambda(1, t) = \min_{i \in \{1,2\}} \frac{\ln(t^\gamma \cos\alpha_i + |t|^\gamma)}{(t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma} - 1},$$

with $0 < \alpha_1 \le \alpha_2 < \pi,\ \mu,\ \sigma > 0,\ \mu + \sigma = \lambda$. Then we have the kernels

$$H(xy) = \min_{i \in \{1,2\}} \frac{\ln((xy)^\gamma \cos\alpha_i + |xy|^\gamma)}{((xy)^\gamma \cos\alpha_i + |xy|^\gamma)^{\lambda/\gamma} - 1},$$

$$K_\lambda(x, y) = \min_{i \in \{1,2\}} \frac{\ln((\frac{y}{x})^\gamma \cos\alpha_i + |\frac{y}{x}|^\gamma)}{(y^\gamma \mathrm{sgn}(x) \cos\alpha_i + |y|^\gamma)^{\lambda/\gamma} - |x|^\lambda}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty} \min_{i \in \{1,2\}} \frac{\ln(t^\gamma \cos\alpha_i + |t|^\gamma)}{(t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma} - 1} |t|^{\sigma-1} dt \\ &= \int_0^\infty \min_{i \in \{1,2\}} \frac{\ln[t^\gamma (1 + \cos\alpha_i)]}{[t^\gamma (1 + \cos\alpha_i)]^{\lambda/\gamma} - 1} t^{\sigma-1} dt \\ &\quad + \int_0^\infty \min_{i \in \{1,2\}} \frac{\ln[t^\gamma (1 - \cos\alpha_i)]}{[t^\gamma (1 - \cos\alpha_i)]^{\lambda/\gamma} - 1} t^{\sigma-1} dt \\ &= \frac{\gamma}{\lambda^2} \left[\int_0^\infty \frac{1}{(1 + \cos\alpha_1)^{\sigma/\gamma}} \int_0^\infty \frac{\ln u}{u - 1} u^{\frac{\sigma}{\lambda} - 1} du \right] \\ &\quad + \frac{\gamma}{\lambda^2} \left[\int_0^\infty \frac{1}{(1 - \cos\alpha_2)^{\sigma/\gamma}} \int_0^\infty \frac{\ln u}{u - 1} u^{\frac{\sigma}{\lambda} - 1} du \right] \\ &= \frac{\gamma}{\lambda^2 2^{\sigma/\gamma}} \left[\left(\sec \frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc \frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \left[\frac{\pi}{\sin \pi\,(\sigma/\lambda)} \right]^2. \end{aligned}$$

By (121) and (129), we have

$$||T_1|| = ||T_2|| = \frac{\gamma}{\lambda^2 2^{\sigma/\gamma}} \left[\left(\sec \frac{\alpha_1}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc \frac{\alpha_2}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \left[\frac{\pi}{\sin \pi(\sigma/\lambda)} \right]^2.$$

In particular, if $\alpha_1 = \alpha_2 = \alpha \in (0, \pi)$, then we have

$$H(t) = K_\lambda(1, t) = \frac{\ln(t^\gamma \cos\alpha + |t|^\gamma)}{(t^\gamma \cos\alpha + |t|^\gamma)^{\lambda/\gamma} - 1},$$

and

$$||T_1|| = ||T_2|| = \frac{\gamma}{\lambda^2 2^{\sigma/\gamma}} \left[\left(\sec \frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} + \left(\csc \frac{\alpha}{2} \right)^{\frac{2\sigma}{\gamma}} \right] \left[\frac{\pi}{\sin \pi\,(\sigma/\lambda)} \right]^2.$$

(d) For

$$\gamma \in \left\{ a; a = \frac{1}{2k-1}, 2k + 1 (k \in \mathbf{N}) \right\},$$

we set

$$H(t) = K_\lambda(1, t) = \min_{i \in \{1,2\}} \frac{1}{\max\{1, (t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma}\}},$$

where $0 < \alpha_1 \le \alpha_2 < \pi,\ \mu,\ \sigma > 0,\ \mu + \sigma = \lambda$. Then we have the kernels

$$H(xy) = \min_{i \in \{1,2\}} \frac{1}{\max\{1, [(xy)^\gamma \cos\alpha_i + |xy|^\gamma]^{\lambda/\gamma}\}},$$

$$K_\lambda(x, y) = \min_{i \in \{1,2\}} \frac{1}{\max\{|x|, (y^\gamma \mathrm{sgn}(x) \cos\alpha_i + |y|^\gamma)^{\lambda/\gamma}\}}$$

and obtain the constant factors

$$\begin{aligned} K(\sigma) = K_\lambda(\sigma) &= \int_{-\infty}^{\infty} \min_{i \in \{1,2\}} \frac{1}{\max\{1, (t^\gamma \cos\alpha_i + |t|^\gamma)^{\lambda/\gamma}\}} |t|^{\sigma-1} dt \\ &= \int_0^\infty \min_{i \in \{1,2\}} \frac{1}{\max\{1, [t^\gamma (1 + \cos\alpha_i)]^{\lambda/\gamma}\}} t^{\sigma-1} dt \\ &\quad + \int_0^\infty \min_{i \in \{1,2\}} \frac{1}{\max\{1, [t^\gamma (1 - \cos\alpha_i)]^{\lambda/\gamma}\}} t^{\sigma-1} dt \\ &= \frac{1}{\lambda} \left[\int_0^\infty \min_{i \in \{1,2\}} \frac{1}{(1 + \cos\alpha_i)^{\sigma/\gamma}} \frac{1}{\max\{1, u\}} u^{\frac{\sigma}{\lambda}-1} du \right] \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\lambda}\left[\int_0^{\infty}\min_{i\in\{1,2\}}\frac{1}{(1-\cos\alpha_i)^{\sigma/\gamma}}\frac{1}{\max\{1,u\}}u^{\frac{\sigma}{\lambda}-1}du\right]\\
&=\frac{1}{\lambda}\left[\int_0^{\infty}\frac{1}{(1+\cos\alpha_1)^{\sigma/\gamma}}\int_0^{\infty}\frac{1}{\max\{1,u\}}u^{\frac{\sigma}{\lambda}-1}du\right]\\
&\quad+\frac{1}{\lambda}\left[\int_0^{\infty}\frac{1}{(1-\cos\alpha_2)^{\sigma/\gamma}}\int_0^{\infty}\frac{1}{\max\{1,u\}}u^{\frac{\sigma}{\lambda}-1}du\right]\\
&=\frac{1}{2^{\sigma/\gamma}}\left[\left(\sec\frac{\alpha_1}{2}\right)^{\frac{2\sigma}{\gamma}}+\left(\csc\frac{\alpha_2}{2}\right)^{\frac{2\sigma}{\gamma}}\right]\frac{\lambda}{\mu\sigma}.
\end{aligned}
$$

By (121) and (129), we have

$$
||T_1||=||T_2||=\frac{1}{2^{\sigma/\gamma}}\left[\left(\sec\frac{\alpha_1}{2}\right)^{\frac{2\sigma}{\gamma}}+\left(\csc\frac{\alpha_2}{2}\right)^{\frac{2\sigma}{\gamma}}\right]\frac{\lambda}{\mu\sigma}. \tag{147}
$$

In particular, if $\alpha_1=\alpha_2=\alpha\in(0,\pi)$, then we have

$$
H(t)=K_\lambda(1,t)=\frac{1}{\max\{1,(t^\gamma\cos\alpha+|t|^\gamma)^{\lambda/\gamma}\}},
$$

and

$$
||T_1||=||T_2||=\frac{1}{2^{\sigma/\gamma}}\left[\left(\sec\frac{\alpha}{2}\right)^{\frac{2\sigma}{\gamma}}+\left(\csc\frac{\alpha}{2}\right)^{\frac{2\sigma}{\gamma}}\right]\frac{\lambda}{\mu\sigma}.
$$

Then, for (a)–(d) we can obtain the equivalent inequalities with the kernels and the best possible constant factors in Theorems 5–8. Setting $\delta_0=\frac{\sigma}{2}>0$, we can still obtain the equivalent reverse inequalities with the kernels and the best possible constant factors in Theorems 5–8.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 61370186), and 2013 Knowledge Construction Special Foundation Item of Guangdong Institution of Higher Learning College and University (No. 2013KJCX0140).

The authors wish to express their thanks to Professors Tserendorj Batbold, Mario Krnic and Jichang Kuang for their careful reading of the manuscript and their valuable suggestions.

References

1. Schur, I.: Bernerkungen sur Theorie der beschrankten Billnearformen mit unendlich vielen Veränderlichen. J. Reine. Angew. Math. **140**, 1–28 (1911)
2. Carleman, T.: Sur les equations integrals singulieres a noyau reel et symetrique. Uppsala (1923)
3. Zhang, K.W.: A bilinear inequality. J. Math. Anal. Appl. **271**, 288–296 (2002)
4. Hardy, G.H.: Note on a theorem of Hilbert concerning series of positive terms. Proc. Lond. Math. Soc. **23**(2), Records of Proc. xlv–xlvi (1925)

5. Hardy, G.H., Littlewood, J.E., Pólya, G.: Inequalities. Cambridge University Press, Cambridge (1934)
6. Mitrinović, D.S., Pečarić, J.E., Fink, A.M.: Inequalities Involving Functions and Their Integrals and Deivatives. Kluwer Academic, Boston (1991)
7. Yang, B.C.: On Hilbert's integral inequality. J. Math. Anal. Appl. **220**, 778–785 (1998)
8. Yang, B.C.: A note on Hilbert's integral inequality. Chin. Q. J. Math. **13**(4), 83–86 (1998)
9. Wang, D.X., Guo, D.R.: Special Functions. Science Press, Beijing (1979)
10. Yang, B.C.: On an extension of Hilbert's integral inequality with some parameters. Aust. J. Math. Anal. Appl. **1**(1), Art., 11, 1–8 (2004)
11. Yang, B.C.: The Norm of Operator and Hilbert-Type Inequalities. Science Press, Beijing (2009)
12. Yang, B.C.: Hilbert-Type Integral Inequalities, Bentham Science, Sharjah (2009)
13. Yang, B.C.: Hilbert-type integral operator: norms and inequalities. In: Pardalos, P.M., Georgiev, P.G., Srivastava, H.M. (eds.) Nonlinear Analysis: Stability, Approximation, and Inequalities. Springer, New York (2012)
14. Yang, B.C.: On Hilbert-type integral inequalities and their operator expressions. J. Guangdong Univ. Educ. **33**(5), 1–17 (2014)
15. Yang, B.C.: A new Hilbert-type integral inequality. Soochow J. Math. **33**(4), 849–859 (2007)
16. Yang, B.C.: A new Hilbert-type integral inequality with some parameters. J. Jilin Univ. (Sci. Ed.) **46**(6), 1085–1090 (2008)
17. He, B., Yang, B.C.: On a Hilbert-type integral inequality with the homogeneous kernel of 0-degree and the hypergeometrc function. Math. Pract. Theory **40**(18), 105–211 (2010)
18. Zeng, Z., Xie, Z.T.: On a new Hilbert-type integral inequality with the homogeneous kernel of degree 0 and the integral in whole plane. J. Inequal. Appl. **2010**, Article ID 256796, 9pp (2010)
19. Yang, B.C.: A reverse Hilbert-type integral inequality with some parameters. J. Xinxiang Univ. (Nat. Sci. Ed.) **27**(6), 1–4 (2010)
20. Wang, A.Z., Yang, B.C.: A new Hilbert-type integral inequality in whole plane with the non-homogeneous kernel. J. Inequal. Appl. **2011**, 123 (2011). doi:10.1186/1029-24X-2011-123
21. Xin, D.M., Yang, B.C.: A Hilbert-type integral inequality in whole plane with the homogeneous kernel of degree -2. J. Inequal. Appl. **2011**, Article ID 401428, 11pp (2011)
22. He, B., Yang, B.C.: On an inequality concerning a non-homogeneous kernel and the hypergeometric function. Tamsul Oxf. J. Inf. Math. Sci. **27**(1), 75–88 (2011)
23. Yang, B.C.: A reverse Hilbert-type integral inequality with a non-homogeneous kernel. J. Jilin Univ. (Sci. Ed.) **49**(3), 437–441 (2011)
24. Xie, Z.T., Zeng, Z., Sun, Y.F.: A new Hilbert-type inequality with the homogeneous kernel of degree -2. Adv. Appl. Math. Sci. **12**(7), 391–401 (2013)
25. Adiyasuren, V., Batbold, Ts.: On a generalization of a Hilbert-type integral inequality in the whole plane with a hypergeometric function. J. Inequal. Appl. **2013**, Article ID 189, 8pp (2013)
26. Zheng, Z, Raja Rama Gandhi, K., Xie, Z.T.: A new Hilbert-type inequality with the homogeneous kernel of degree -2 and with the integral. Bull. Math. Sci. Appl. **3**(1), 11–20 (2014)
27. Milovanović, G.V., Rassias, M.Th.: Some properties of a hypergeometric function which appear in an approximation problem. J. Glob. Optim. **57**, 1173–1192 (2013)
28. Rassias, M.Th., Yang, B.C.: On half-discrete Hilbert's inequality. Appl. Math. Comput. **220**, 75–93 (2013)
29. Rassias, M.Th., Yang, B.C.: A multidimensional half - discrete Hilbert - type inequality and the Riemann zeta function. Appl. Math. Comput. **225**, 263–277 (2013)
30. Rassias, M.Th., Yang, B.C.: On a multidimensional half - discrete Hilbert - type inequality related to the hyperbolic cotangent function. Appl. Math. Comput. **242**, 800–813 (2014)
31. Rassias, M.Th., Yang, B.C.: A multidimensional Hilbert - type integral inequality related to the Riemann zeta function. In: Daras, N.J. (ed.) Applications of Mathematics and Informatics in Science and Engineering, pp. 417–433. Springer, New York (2014)
32. Milovanović, G.V., Rassias, M.Th. (eds.): Analytic Number Theory, Approximation Theory and Special Functions. Springer, New York (2014)

33. Kuang, J.C.: Applied Inequalities. Shangdong Science and Technology Press, Jinan (2004)
34. Kuang, J.C.: Introduction to Real Analysis. Hunan Educiton Press, Changsha (1996)
35. Cheng, Q.X., et al.: Base on Real Functions and Functional Analysis, 3rd edn. Higher Education Press, Beijing (2010)
36. Zhong Y.Q.: On Complex Functions, 3rd edn. Higher Education Press, Beijing (2003)
37. Pan, Y.L., Wang, H.T., Wang, F.T.: Complex Functions. Science Press, Beijing (2006)

A Secure Communication Design Based on the Chaotic Logistic Map: An Experimental Realization Using Arduino Microcontrollers

Mauricio Zapateiro De la Hoz, Leonardo Acho, and Yolanda Vidal

Abstract Chaotic systems feature some characteristics that are being actively exploited in the field of communication systems. However, there are still some drawbacks to be solved before actual feasible implementation of these systems can be possible. One basic communication scheme found in the literature is the digital-based scheme that uses discrete dynamical systems. In this case, chaotic maps are frequently employed as pseudo-random bit generators used for encrypting the messages. In this chapter we present a digital-based communication system that uses the discrete logistic map which is a second-order polynomial map. The input message signals is modulated using a 1-bit analog-to-digital converter. Then, a logistic map is implemented in order to generate a digital binary version that is used to encrypt the message. In the receiver side, the binary-coded message is decrypted using a key signal that is sent through one of the communication channels. The proposed scheme is experimentally tested using Arduino shields which are simple yet powerful development kits that allows for the implementation of the communication system for testing purposes.

Keywords: Logistic map • Arduino • Secure communications

1 Introduction

Security and secrecy in communications are some of the most important concerns in nowadays societies. With the advent of worldwide networks and digital communication techniques, the cryptographic techniques that once were restricted to military and state affairs are now covering several domains such as banks, private

M. Zapateiro D. (✉)
Universidade Tecnológica do Paraná, Avenida Alberto Carazzai 1640,
86300-000 Cornélio Procópio, Paraná, Brazil
e-mail: hoz@utfpr.edu.br

L. Acho • Y. Vidal
Departament de Matemàtica Aplicada III, Universitat Politècnica de Catalunya,
Comte d'Urgell 187, 08036 Barcelona, Spain
e-mail: leonardo.acho@upc.edu; yolanda.vidal@upc.edu

N.J. Daras, M.Th. Rassias (eds.), *Computation, Cryptography, and Network Security*, DOI 10.1007/978-3-319-18275-9_28

Table 1 Comparison between chaos and cryptography properties

Chaos property	Cryptographic property
Ergodicity	Confusion
Sensitivity to initial conditions/control parameter	Diffusion with a small change in the text/secret key
Mixing property	Diffusion with a small change within one block of the plaintext
Deterministic dynamics	Deterministic pseudo-randomness
Structural complexity	Algorithm complexity

Table adapted from Table 1 as it appears in [2, 32]

companies, medical organizations, etc. This has led to a very active research field oriented to finding optimal solutions to the problem of communications security [5, 11, 31]. As a result, numerous cryptographic techniques that seek to preserve the privacy of the information transmitted have been designed. However, they are all vulnerable to some degree and thus efforts are still being made in order to find better solutions. One trend in this research field is the application of chaotic systems. The highly unpredictable and random-look nature of chaotic signals is the most attractive feature of deterministic chaotic systems that may lead to novel engineering applications [10]. Alvarez and Li [2] and Volos [32] summarize in a very concise way the comparison between chaos and cryptography that helps understand this last point as can be seen in Table 1.

Chaos thus has become very important for encryption/decryption purposes as will be seen in the next paragraphs. There are basically two main approaches to designing secure communication systems based on chaotic dynamics: analog and digital.

Analog communication systems based on chaos has become possible because of the possibility of synchronization. This is the possibility of using the output of a driving system (master) to control the response system (slave) in such a way that they both oscillate in a synchronized manner. This was discovered by Pecora and Carroll [24] and opened up the way to applying chaos in communication systems. A wide variety of synchronization schemes have been developed since then. For instance, Agiza and Yassen [1] demonstrated that synchronization was possible in two different systems: one of them composed of two identical Rossler chaotic systems and the other one composed of two identical Chen chaotic systems. Another example can be found in the work by Huang [8] who investigates the application of adaptive control techniques for chaos synchronization between the Lorenz–Stenflo (LS) system and a novel dynamical system called CYQY, as well as the synchronization between an LS system and a hyper-chaotic system. An interesting work in this field is that by Park [20] who accomplished synchronization between two different chaotic systems by means of nonlinear control laws. The author demonstrates that the two different systems could be controlled using nonlinear control techniques and proved the closed-loop stability by means of linear control theory.

As mentioned earlier, the synchronization of chaotic systems led to the design of communication systems in which chaotic oscillators are used to encrypt/decrypt the information. In these systems, the chaotic oscillator signals are used to encrypt the message and thus, these systems always require a synchronizing signal so that the chaotic signal (or signals) can be reconstructed in the receiver. This is how the message sent from the transmitter can be retrieved. Here are some examples. Zapateiro et al. [35] designed a chaotic communication system in which a binary signal is encrypted in the frequency of the sinusoidal term of a chaotic Duffing oscillator. Two chaotic signals of the oscillator are further encrypted with a Delta modulator before they are sent through the channel. In the receiver, a Lyapunov-based observer uses the chaotic signals for retrieving the sinusoidal term that contains the message. A novel frequency estimator is then used to obtain the binary signal. Furthermore, in a new proposal, Zapateiro et al. [36] investigated a modified Chua chaotic oscillator in which the nonlinear term of the original oscillator was changed for a smooth and bounded function that allows for easier analysis and synchronization with another oscillators. An application to secure communications using the modified oscillator was developed and its performance evaluated by numerical simulations. Fallahi and Leung [6] developed a chaotic communication system based on a chaos multiplication modulator that encrypts the signal. The chaotic signal is generated by using the Genesio-Tesi chaotic system. This scheme does not require the knowledge of the initial conditions of the transmitter. The authors also prove that the system security could not be broken with the existing methods at that time. Yang and Chua [34] proposed a secure communication system based on impulsive stabilization. In the transmitter, a chaotic oscillator and an embedded cryptographic scheme is implemented. The receiver consists of a chaotic oscillator and a cryptographic scheme, both identical to those of the transmitter. The transmitted signal consists of a sequence of time frames divided into two regions. The first region is a synchronization region, which uses synchronization impulses for synchronizing the chaotic systems in both the transmitter and the receiver. The second region contains the scrambled signal. The synchronization of the chaotic oscillators is performed by means of the theory of impulsive synchronization, which the authors developed in this work. In this system, the key signal is generated by the chaotic system.

On the other hand, digital chaos communication systems do not depend on chaos synchronization at all. Instead, they usually use one or more chaotic maps in which the initial conditions and the control parameters play the role of the secret key [2]. For instance, Lee et al. [13] proposed a chaotic cipher stream, a new scheme for generating pseudo-random numbers based on the composition of chaotic maps. The method consists of using one chaotic map to generate a sequence of pseudo-random bytes and then apply some permutation on them using another chaotic map. Liu and Sun [14] propose a new design of chaotic cryptosystems in which they use high dimensional chaotic maps along with some cryptography techniques to achieve a high security level. The high dimensionality of the map leads to a high complexity and effective byte confusion and diffusion of the output ciphertext at the time that the small key space problem is overcome. Patidar and Sud [22] proposed

a pseudo random bit generator based on two chaotic systems running side by side. This scheme increases complexity and thus the difficulty for an intruder to break it. In an image encryption application, Zhang et al. [38] used discrete exponential maps along with spatial S-box transform to design a key scheme resistant to statistic attack and grey code attack. The properties of confusion and diffusion were improved as shown by their experimental results. Pareek et al. [19] designed an image encryption scheme in which two logistic maps are used along with an 80-bit key to encrypt/decrypt the images. Eight different types of operation are used to encrypt the pixels of an image; the type of operation is chosen according the outcome of the logistic maps. The robustness of the secure communication scheme was proven by means a statistical, key sensitivity and key space analysis.

In this chapter, we present a digital chaos communication system in which a logistic map is used along with a 1-bit analog-to-digital converter (ADC) to encrypt a message. This technique is also known as Delta modulator and is one of the most simple and robust methods of ADC schemes requiring serial digital communications of analog signals [27]. The transmitter and receiver are implemented low cost, small but powerful microcontroller boards: Arduino Uno R3 [3]. The Arduino transmitter receives a message which is analog in nature and encrypts it using a logistic map and a 1-bit ADC. Then the Arduino receiver decrypts the message and converts to digital form which corresponds to the Delta-modulated signal. In order to obtain the analog version of the message signal, an analog circuitry performs 1-bit digital-to-analog conversion (DAC) and retrieves the message.

This chapter is organized as follows. Section 2 describes the problem to be treated and a scheme of the proposed solution. Section 3 is an introduction to the logistic map from its origin to its applications in secure communications. Section 4 presents the details of the implementation of the proposed technique. Finally, the conclusions are presented in Sect. 5.

2 Problem Statement

In this chapter we explore the secure communications problem by means of chaos techniques. The objective is to transmit a message $m(t)$ between two points. The communication system scheme is shown in Fig. 1 and it consists of the following elements:

- **Arduino transmitter.** This is the core of the transmitter. The Arduino board will take the message $m(t)$ through one of its analog input ports, convert it to a digital signal $m(k)$, and then encrypt it using a logistic map and a 1-bit ADC. This process generates two outputs: one is the encrypted message $me(k)$ and the other one is the signal $s(k)$ that is used for decryption purposes. Note that the Arduino will sample the input message $m(t)$ and convert it to $m(k)$, $k = nt$, $n = 0, 1, 2, \ldots$
- **Channels.** Two wired channels are used to send the encrypted and key signals to the receiver.

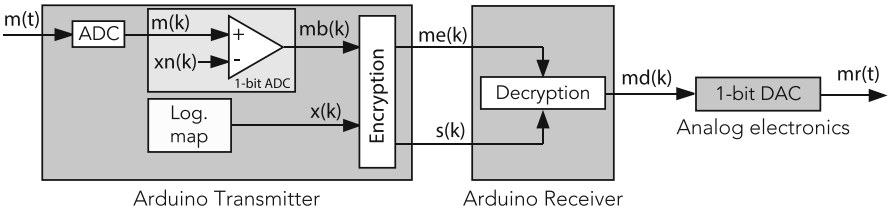


Fig. 1 Block diagram of the communication system

- **Arduino receiver.** This is one of the two main blocks in the receiver side. It takes the signals $me(k)$ and $s(k)$ to decrypt the 1-bit digital signal before it is converted to its analog form. The output is a digital signal $md(k)$ which corresponds to the signal $mb(k)$.
- **1-bit DAC.** This is the second block in the receiver. It is a 1-bit DAC consisting of an integrator, a filter and some amplifiers to retrieve the original message. Its output is a signal $mr(t) \approx m(t)$.

The details of these blocks will be outlined in the following sections of this chapter.

3 The Logistic Map

The Logistic Map has its origins in the works by the Belgian mathematician Pierre-François Verhulst in the first half of the eighteenth century. According to the biographies by Kint et al. [9] and Pastijn [21] and the references therein, Verhulst was a brilliant mathematician who excelled since very young: at age 18, he enrolled an exact sciences career at the University of Ghent and obtained his doctoral degree 3 years later with a dissertation on the reduction of binomial equations. After a break in which he lived in Italy and enrolled the Belgian army to participate in a battle against Holland, he became professor of the Royal Military Academy in Brussels.

Verhulst began to develop an interest in the application of mathematics to the political context, particularly, to the idea of how population growth could be modeled. After a few years of research and discussions, he published in 1845 an article entitled *Recherches mathématiques sur la loi d'acroissement de la population* [29] (Mathematical investigations about the law of population growth) in which he developed the idea of the logistic growth model (*la courbe logistique* as he named it). In this work, Verhulst highlights the importance of knowing the laws that rule the population progress and recognizes that there are numerous factors that influence the multiplication of the human race. Due to the difficulty of solving the problem in a general way, he proposed a calculus-based model that neglected the "accidental causes" because to his understanding, the statistics science was not developed enough by that time. In 1846 he further presented another article to the

Academy entitled *Deuxième memoire sur la loi d'acroissement de la population* (Second memory about the law of population growth). It was published in 1847 [30] and it was a critical revision of his previous work.

After Verhulst's dead in 1849, the logistic curve lost interest until 1920 when it was rediscovered [9]. In that year, a paper entitled *On the rate of Growth of the Population of the United States since 1790 and its Mathematical Representation* [23], Pearl and Reed studied the mathematical models that were used at the time to determine how the population size evolved along the time in the United States. They concluded that the existing models did not reflect too much accuracy and came up with an equation that better fitted the real data. That equation was exactly the same Verhulst's logistic curve though they initially ignored it. The logistic model then resurged to be widely used in natural sciences to represent the population dynamics of different species.

In 1972 the meteorologist Lorenz presented one of the pioneering works on chaotic dynamics. His work entitled *Does the flap of butterfly's wings in Brazil set off a tornado in Texas?* [15] was a research on how some meteorological phenomena could be modeled with a chaotic dynamic system. Then a series of works on chaotic systems began to be developed and the logistic model would soon come along with them. In 1976 May presented an article entitled *Simple mathematical models with very complicated dynamics* [16] in which he described how the simple logistic map, i.e., the discrete-time version of Verhulst's logistic curve, would lead to chaos. The importance of the logistic map as a simple chaotic system then begun.

The equation of the logistic model as it appears in Verhulst's works is:

$$\frac{M}{p}\frac{dp}{dt} = m - np \tag{1}$$

where p is the population, $m = l + nb$, b is the population corresponding at the moment that the study begins, l/M is a coefficient relative to the weakening of the population, and n is a constant. This is a first order differential equation. It is well known that chaotic systems must be at least of third order, however this is not true for discrete time systems. The logistic map, the discrete-time version of Verhulst's logistic model is indeed chaotic under certain conditions. Its equation is:

$$x_{n+1} = rx_n(1 - x_n), \;\; 0 \leq x \leq 1 \tag{2}$$

where r is a constant parameter. Figure 2 is the bifurcation diagram of the logistic map created by varying the parameter r from 2.5 to 4.0.

As can be seen in the bifurcation diagram, there are different regions that depend on the value of r. It is of particular interest when $r = 3$ because there it begins the period doubling that leads to the chaotic dynamics when $r \approx 3.5699\ldots$ until $r = 4.0$. Figure 3 shows the Lyapunov exponent of the logistic map as r is varied from 2.5 to 4.0. It can be seen that the Lyapunov exponent λ becomes positive for values of or greater than 3.56 approximately which is a strong indicator of chaos [33].

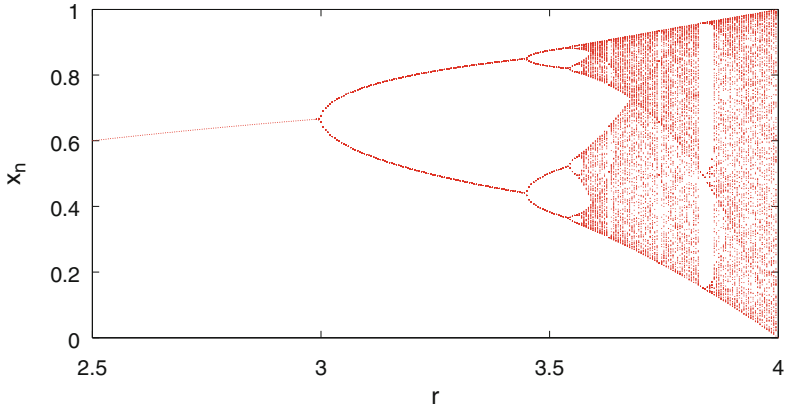


Fig. 2 Logistic map bifurcation diagram

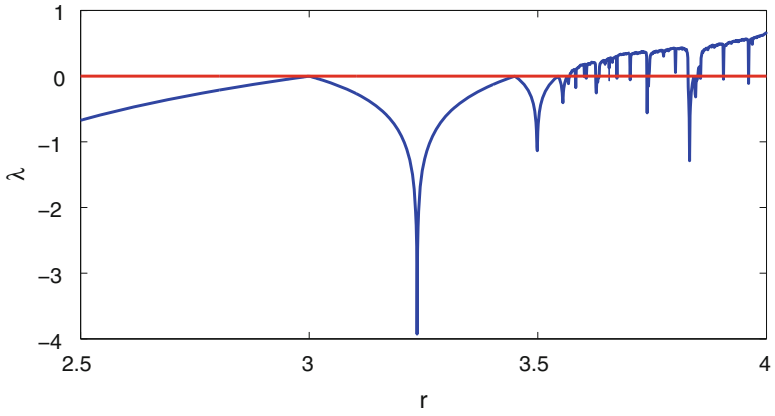


Fig. 3 Logistic map Lyapunov exponent

As was discussed earlier in Sect. 1, digital communication systems based on chaotic maps are being widely studied and the logistic map is no exception. Several works can be found in the literature in which the chaotic properties of the logistic maps are exploited in the design of cryptography techniques for improving secure communications. For example, Murillo-Escobar et al. [17] presented a symmetric text cipher in which they used a 128-bit secret key, two logistic maps with optimized pseudorandom sequences, plain text characteristics, and only one permutation diffusions round. Security analysis was performed to demonstrate its feasibility. Ursulean [28] studied the properties of the logistic map as a pseudo-random bit generator and carried out statistical tests to analyze its performance. Lawrence and Wolff [12] explored the generation of one or more binary-valued sequences from a standard logistic map, according to the continuous values being in one of two sub-intervals of the map's domain defined by cut-points, one applying to each binary

process and presented an application to secure communications. Zhang and Cao [37] proposed a technique for encrypting images in which a new modification of the logistic map is proposed. The modified logistic map is, according to the authors, a better choice for encryption due to the improved chaotic properties as a result of a much larger Lyapunov exponent. Singh and Sinha [26] proposed an opto-electronic communication system that uses a logistic map and pulse position modulation. In this scheme, the input signal (message) is added to a chaotic signal generated by a logistic map. Then it is modulated with a pulse position modulator (PPM). The modulated signal is then sent through the channel to the receiver in which the inverse operation is performed in order to retrieve the message. The authors experimentally tested this scheme with optical fiber with satisfactory results. He et al. [7] proposed a scheme in which the message is processed using a logistic map and the chaotic parameter modulation (CPM) technique. Then it is sent to the receiver where a nonlinear control factor is introduced in order to synchronize the transmitter and the receiver and thus retrieve the message. Chang [4] presented a communication system based on the asymptotic synchronization of modified logistic hyper-chaotic system. For that purpose, they proposed a modification of the logistic map in which is uniformly distributed in [0,1]. The difference with respect to the original logistic map is that the modified version does not exhibit windows. This has the advantage of a greater key space for communications. Volos [32] presented a chaotic random bit generator and implemented it in an Arduino board. The microcontroller runs side-by-side two logistic maps working in different chaotic regimes due to the different initial conditions and system parameters. Statistical tests were carried out to prove security against intruders. Pande and Zambreno [18] presented another experimental realization of a chaotic encryption scheme, this time using a Xilin Virtex 6 FPGA. They implemented a modified logistic map that improves the performance of the logistic map in terms of Lyapunov exponent and uniformity of the bifurcation diagram.

In the next sections, we will use a logistic map as part of an encryption/decryption scheme for transmitting information. In the next sections we will explain the details of the prototype of this communication system which is implemented in two Arduino Uno boards.

4 Experimental Implementation

4.1 Description of the Communication System

The communication system implemented in this work consists of a transmitter and a receiver whose cores are the Arduino Uno R3 microcontroller boards, shown in Fig. 4. These are low cost, simple but powerful microcontrollers based on the ATmega328 chip. They have 14 digital input/output pins (six of them can be used

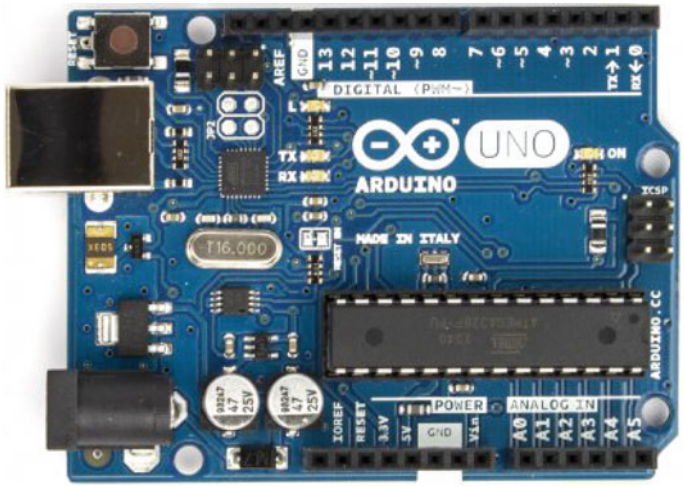


Fig. 4 Picture of an Arduino Uno R3 microcontroller. Picture taken from the Arduino website [3]

as PWM outputs), six analog inputs, a 16 MHz crystal oscillator, a USB connection, and a reset button. They can be programmed using a language similar to C++ called Wiring [3].

The flow diagram of the programs executed by each Arduino is shown in Fig. 5 in order to facilitate the description of the communication system algorithms.

The communication starts when a message $m(t)$ is produced by a function generator and sent to the analog input A0 of the Arduino transmitter. Arduino analog inputs only accepts unipolar signals in the range form 0 to 5 V. An embedded 10-bit ADC converts the input signal from analog to digital at a maximum rate of 10,000 samples per second. However, as can be seen in the flow diagram, the loop is repeated every 0.5 ms and thus, the message input is sampled at a rate of 2000 samples per second. In order to guarantee the timing, we made use of the SimpleTimer library [25]. Since the output of the ADC is a value between 0 and 1023 (the ADC resolution), an internal operation to bring it back to the range from 0 to 5 V is executed. The result is a sampled message signal $m(k)$.

The next step is the 1-bit ADC conversion. The ADC conversion scheme, also known as simple Delta modulation, shown in Fig. 6, consists of a comparator in the forward path and an integrator in the feedback path of a simple control loop. The modulated output $mb(k)$ is either true or false at any given time. The signal $m(k)$ is compared to another signal $xn(k)$ which is generated internally by the algorithm. $xn(k)$ is a digital implementation of an integrator, which is the base of the 1-bit ADC conversion [27]. This value is updated every loop of the Arduino program.

After one bit from the ADC is obtained, the logistic map is called to generate a value $x(k)$ and then proceed to the encryption. The encryption algorithm is then:

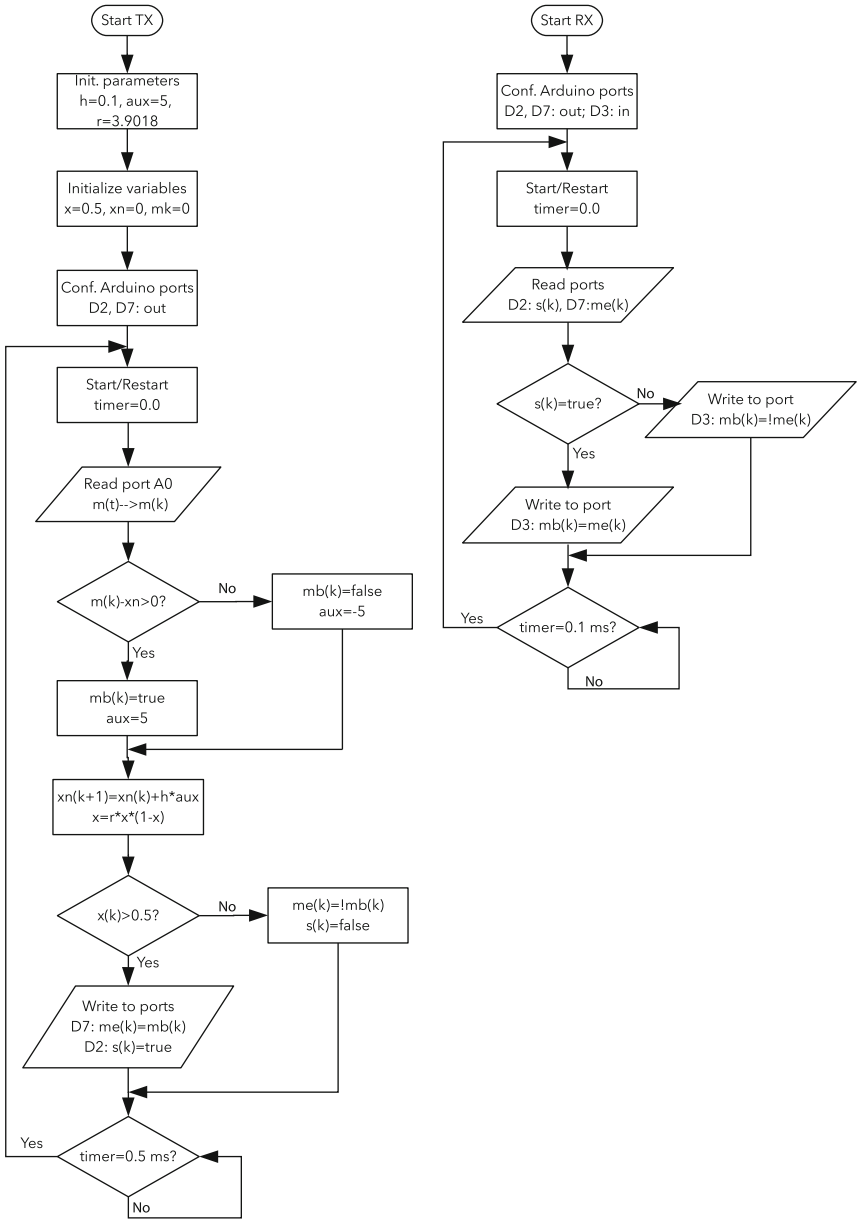


Fig. 5 Flow diagram of the Arduino codes. *Left*: transmitter. *Right*: receiver

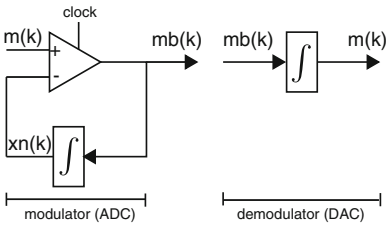


Fig. 6 Diagram of the 1-bit ADC/DAC converter also known as Delta modulator

```
if x(k) > 0.5 then
me(k) = mb(k)
s(k) = true
else
me(k) = !mb(k)  //Symbol ! means boolean negation
s(k) = false
end
```

where $me(k)$ is the encrypted message and $s(k)$ is the key. These signals are sent to the receiver through digital outputs D2 and D7. The key signal $s(k)$ can also be encrypted using, for example, Karnaugh maps, however this was not done in this work.

In the receiver, the signals $me(k)$ and $s(k)$ go directly to the Arduino inputs D7 and D2, respectively. The flow diagram of the receiver program is shown in Fig. 5 as well. The receiver decrypts the message by analyzing the key signal $s(k)$ by running the following algorithm:

```
if s(k) = 1 then
md(k)=me(k)
else
md(k)=!me(k)
end
```

where $md(k)$ is the decrypted signal. The receiver runs every loop in 0.5 ms. The output $md(k)$ is sent to the output pin D3 and it goes directly to the 1-bit DAC realized with analog electronics using operational amplifiers. As shown in Fig. 6, the DAC or Delta demodulation consists of an integrator. The signal is passed through different stages though as shown in the circuit diagram of Fig. 7. The circuit has three main blocks. The first one, composed of the amplifiers U1 and U2 is a unipolar to bipolar converter. Recall that the Arduino inputs must be unipolar so in the case that the original signals are bipolar they must recovered to its original form at the output of the Arduino. Thus the signal $m(k) \in [0, 5]$ V is converted to a signal $m(t) \in [-2.5, 2.5]$ V. The second block is composed of amplifiers U3 and U4. They are an integrator that performs the DAC and an amplifier to adjust the quality of its output. This signal is finally sent through a low-pass filter, an amplifier, and an inverter (amplifiers U5–U7) to get the final $mr(t)$ which should be approximately equal to $m(t)$.

The codes of the Arduino transmitter and receiver are shown in the Appendix.

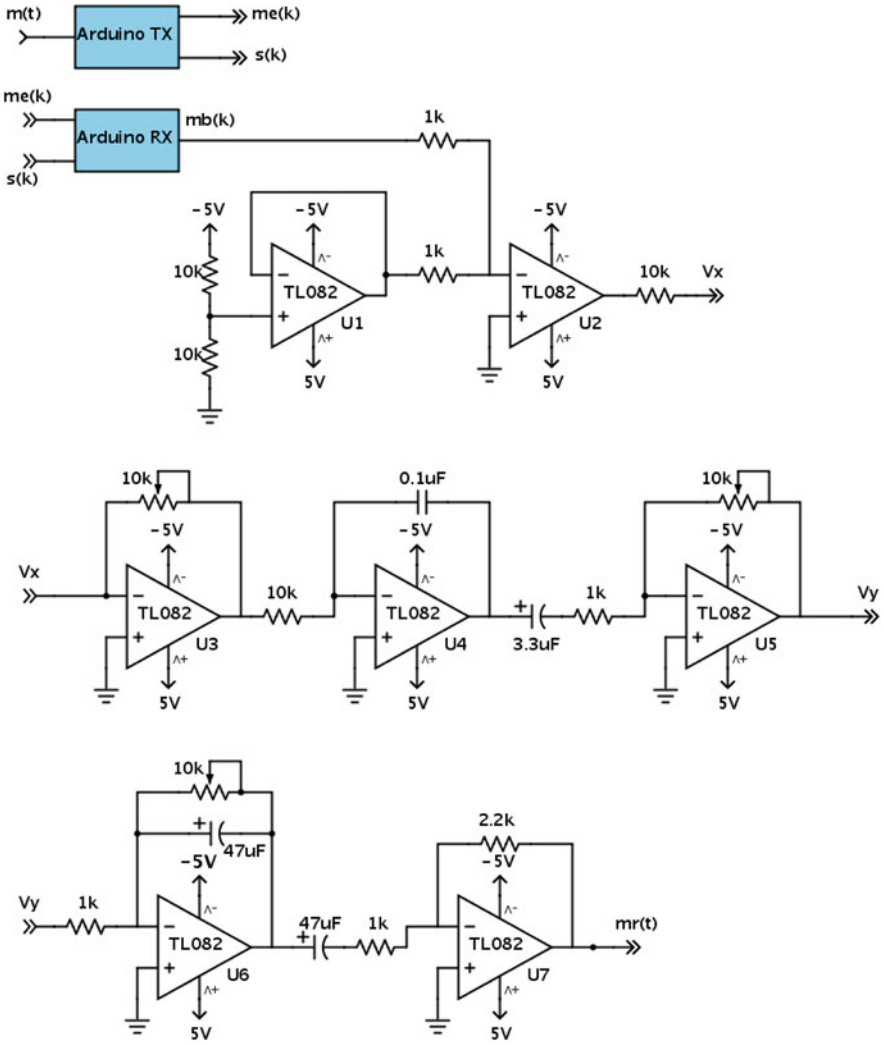


Fig. 7 Circuit diagram of the analog electronics in the receiver

4.2 Experimental Results

The communication system was implemented for experimental purposes. Figure 8 is a picture of the experiment in which we observe the two Arduino boards and a protoboard with the analog electronics. For the experiments, the logistic map was implemented with $r = 3.9018$ and an initial condition $x(0) = 0.5$. The sequence of numbers generated under these conditions is shown in Fig. 9.

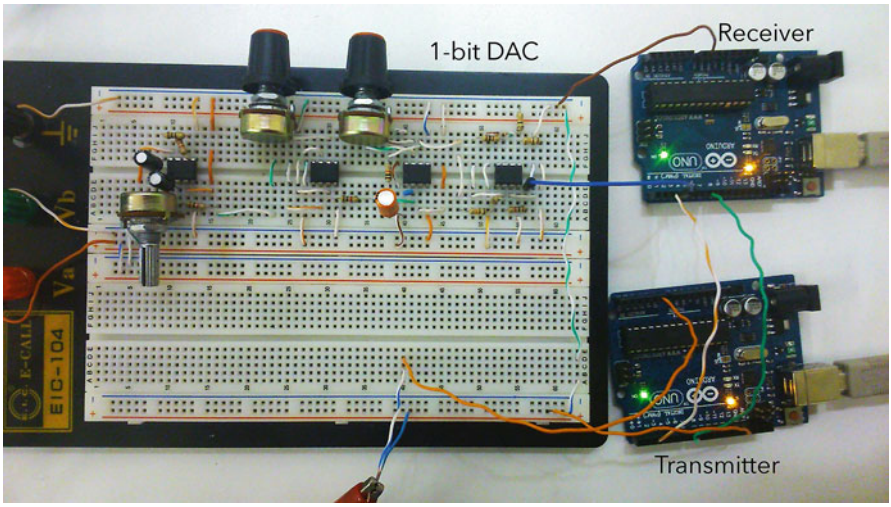


Fig. 8 Picture of the circuit

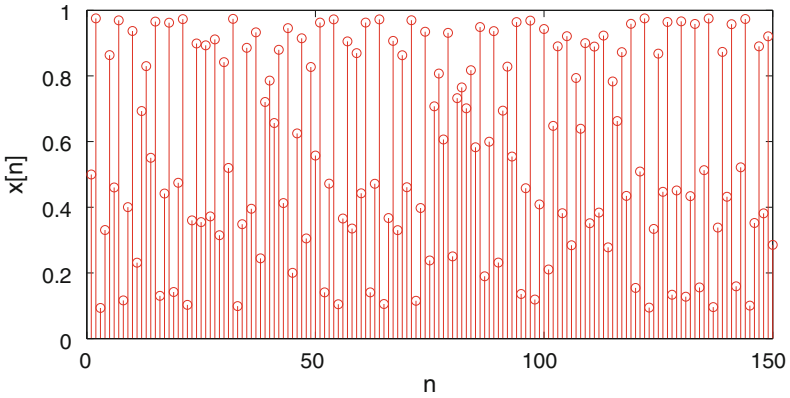


Fig. 9 Numbers generated by the logistic map with $r = 3.9018$ and $x(0) = 0.5$

Figures 10, 11, and 12 are screenshots of the oscilloscope corresponding to the first experiment. In this case, a 160 Hz sine wave, 5 V peak-to-peak amplitude, was used as a message signal. In Fig. 10 we see a comparison of the sent message $m(t)$ (in blue) and the retrieved message $mr(t)$ (in yellow). Figure 11 compares the sent message $m(t)$ (in blue) and the key signal $s(k)$ (in yellow). Figure 12 is a comparison on the sent message $m(t)$ (in blue) and the encrypted message $me(k)$ (in yellow).

In a second experiment, a 150 Hz triangular wave was used as a message. The screenshots of the oscilloscope are displayed in Figs. 13, 14, and 15. Figure 13 compares the sent message $m(t)$ to the retrieved message $mr(t)$. Figures 14 and 15

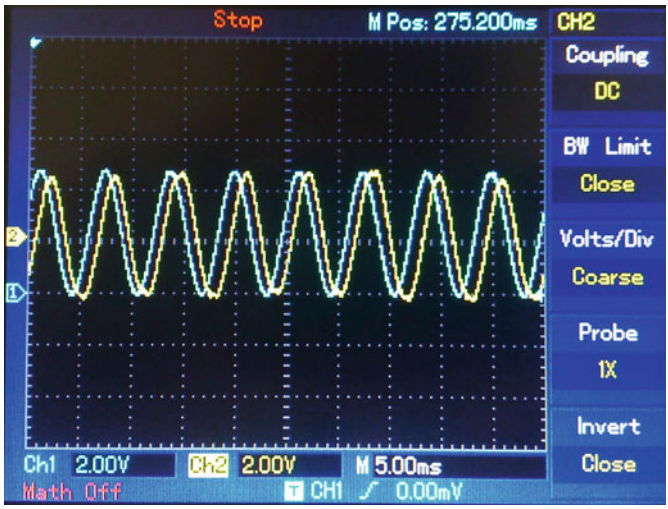


Fig. 10 160 Hz sine wave message. *Blue*: sent message. *Yellow*: retrieved message

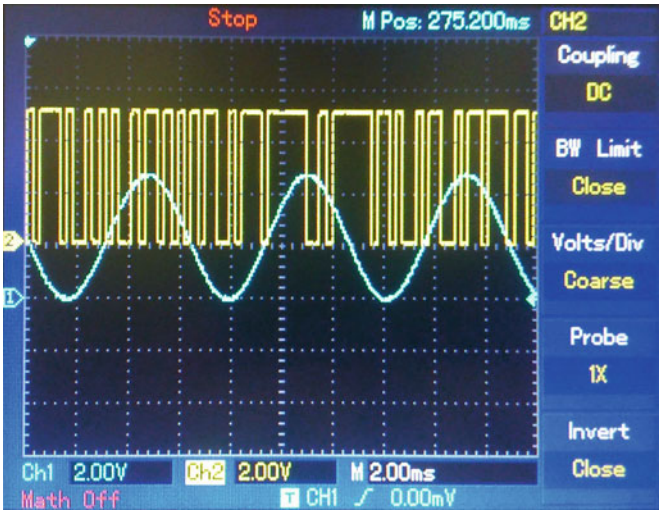


Fig. 11 160 Hz sine wave message. *Blue*: sent message. *Yellow*: key signal

are the key signal $s(k)$ and the encrypted message $me(k)$ compared to the sent message $m(t)$, respectively.

Finally, in Fig. 16 we can see a random-like message signal (in yellow) and its retrieved version (in blue). This signal was generated by making sounds through an electret microphone. For this experiment, it was necessary to reduce the execution time of every loop of the Arduino microcontrollers to 0.1 ms in order to account for the wider frequency spectrum of the signal.

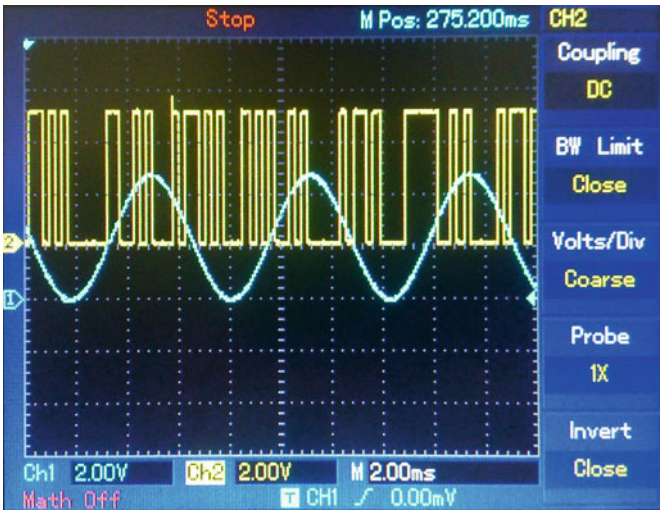


Fig. 12 160 Hz sine wave message. *Blue*: sent message. *Yellow*: encrypted message

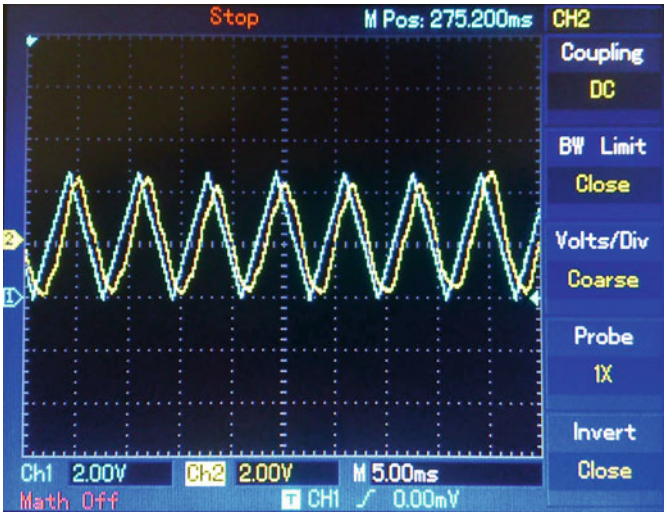


Fig. 13 160 Hz triangular wave message. *Blue*: sent message. *Yellow*: retrieved message

5 Conclusion

In this chapter we have reviewed the digital secure communication systems using the logistic map and proposed a new scheme based on it. The communication system proposed uses a 1-bit DAC (also known as Delta modulator) to modulate the message signal and a logistic map for encryption. The whole system was

Fig. 14 160 Hz triangular wave message. *Blue*: sent message. *Yellow*: key signal

Fig. 15 160 Hz triangular wave message. *Blue*: sent message. *Yellow*: encrypted message

implemented with Arduino Uno microcontroller boards that run the encryption and decryption algorithms in the transmitter and receiver, respectively. The results of experiments showed the feasibility of using the Arduino microprocessors for the task proposed. In future works, the key signal used to decrypt the message is going to be encrypted as well in order to increase the security of the transmission.

Fig. 16 Random signal. *Blue*: sent message. *Yellow*: encrypted message

Acknowledgements Mauricio Zapateiro is supported by the fellowship from CAPES/Programa Nacional de Pós-Doutorado from Brazil. This work was funded by the European Union (European Regional Development Fund) and the Spanish Ministry of Economy and Competitiveness through the research projects DPI2012-32375/FEDER, DPI2011-28033-C03-01 and DPI2014-58427-C2-1-R and by the government of Catalonia (Spain) through 2014SGR859.

Appendix

Arduino Transmitter Code

```
///TX code
#include <SimpleTimer.h>
SimpleTimer timer;
double x=0.5; //Logistic map threshold for encryption.
double h=0.1; //Digital integrator parameter.
double xn=0; //Digital integrator signal xn(k).
double r=3.9018;//Logistic map parameter r.
int mk; //k-th sample of message signal m(t).
int aux; //Digital integrator parameter.
int me; //Encrypted message me(k).

void setup(){
pinMode(2,OUTPUT);
    pinMode(7,OUTPUT);
```

```
    timer.setInterval(0.5,repeatMe);
}

void repeatMe(){
   mk=analogRead(A0)*5/1023;
   if(mk-xn>0){
     me=HIGH;
     aux=5;
   }
   else{
     me=LOW;
     aux=-5;
   }
   xn=xn+h*aux;
   x=r*x*(1-x);
   if(x>0.5){
      digitalWrite(7,me);
      digitalWrite(2,HIGH);
   }
   else{
     digitalWrite(7,!me);
     digitalWrite(2,LOW);
  }
}

void loop(){
  timer.run();
}
```

Arduino Receiver Code

```
#include <SimpleTimer.h>
SimpleTimer timer;
int me; //Encrypted signal from the transmitter me(k).

void setup() {
pinMode(2,INPUT);
    pinMode(7,INPUT);
    pinMode(3,OUTPUT);
    timer.setInterval(0.1,repeatMe);
}

void repeatMe(){
```

```
me=digitalRead(7);
     if(digitalRead(2)==HIGH{
      digitalWrite(3,me);
     }
     else{
      digitalWrite(3,!me);
     }
}

void loop(){
timer.run();
}
```

References

1. Agiza, H.N., Yassen, M.T.: Synchronization of Rossler and Chen chaotic dynamical systems using active control. Phys. Lett. A. **278**, 191–197 (2001)
2. Alvarez, G., Li, S.: Some basic cryptographic requirements for chaos-based cryptosystems. Int. J. Bifurcation Chaos **16**(8), 2129–2151 (2006)
3. Arduino: (2015). http://www.store.arduino.cc/product/A000066
4. Chang, S.-M.: Chaotic generator in digital secure communication. In: Proceedings of the World Congress on Engineering 2009 (WCE2009), London, 1–3 July 2009
5. Chen, C.K., Lin, C.L.: Text encryption using ECG signals with chaotic logistic map. In: The 5th IEEE Conference on Industrial Electronics and Applications (ICIEA 2010), Taichung, 15–17 June 2010
6. Fallalih, K., Leung, H.: A chaos secure communication scheme based on multiplication modulation. Commun. Nonlinear Sci. Numer. Simul. **15**, 368–383 (2010)
7. He, L.-F., Zhang, G., Tian, Z.-S.: A chaotic secure communication scheme based on logistic map. In: 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 22–24 Oct 2010
8. Huang, J.: Adaptive synchronization between different hyper-chaotic systems with fully uncertain parameters. Phys. Lett. A. **372**, 4799–4804 (2008)
9. Kint, J., Constales, D., Vanderbauwhede, A.: Pierre-Franǫis Verhulst's final triumph. In: Ausloos, M., Dirickx, M. (eds.) The Logistic Map and the Route to Chaos, p. 13. Springer, Heidelberg (2006)
10. Kokarev, L., Jakimoski, G.: Logistic map as a block encryption algorithm. Phys. Lett. A **289**, 199–206 (2001)
11. Larger, L., Goedgebuer, J.-P.: Encryption using chaotic dynamics for otpimal telecommunications. C. R. Phys. **5**, 609–611 (2004)
12. Lawrence, A.J., Wolff, R.C.: Binary time series generated by chaotic logistic maps. Stochastics Dyn. **3**(4), 529–544 (2003)
13. Lee, P.-H., Pei, S.-C., Chen, Y.-Y.: Generating chaotic stream ciphers using chaotic systems. Chin. J. Phys. **41**(6), 559–581 (2003)
14. Liu, S.T., Sun, F.Y.: Spatial chaos-based image encryption design. Sci. China Ser. G Phys. Mech. Astron. **52**(2), 177–183 (2009)
15. Lorenz, E.N.: Does the flap pf a butterfly's wings in Brazil set off a tornado in Texas? In: 139th Meeting of the American Association for the Advancement of Science, 29 Dec 1972

16. May, R.M.: Simple mathematical models with very complicated dynamics. Nature **261**, 459–467 (1976)
17. Murillo-Escobar, M.A., Abundiz-Pérez, F., Cruz-Hernández, C., López-Gutiérrez, R.M.: A novel symmetric text encryption algorithm based on logistic map. In: Proceedings of the 2014 International Conference on Communications, Signal Processing and Computers (ICNC 2014), Honolulu, 3–6 Feb 2014
18. Pande, A., Zambreno, J.: A chaotic encryption scheme for real-time embedded systems: design and implementation. Telecommun. Syst. (2011). doi:10.1007/s11235-011-9460-1
19. Pareek, N.K., Patidar, V., Sud, K.K.: Image encryption using chaotic logistic map. Image Vision Comput. **24**, 926–934 (2006)
20. Park, J.H.: Chaos synchronization between two different chaotic dynamical systems. Chaos, Solitons Fractals **27**, 549–554 (2006)
21. Pastijn, H.: Chaotic growth with the logistic model of P.-F. Verhulstin. In: Ausloos, M., Dirickx, M. (eds.) The Logistic Map and the Route to Chaos, p. 3. Springer, Heidelberg (2006)
22. Patidar, V., Sud, K.K.: A pseudo random bit generator based on chaotic logistic map and its statistical testing. Informatica **33**, 441–452 (2009)
23. Pearl, R., Reed, L.J.: On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proc. Natl. Acad. Sci. **6**, 275–288 (1920)
24. Pecora, L.M., Caroll, T.L.: Synchronization in chaotic systems. Phys. Rev. Lett. **64**, 821–825 (1990)
25. Romani, M.: SimpleTimer library for Arduino (2010). http://www.playground.arduino.cc/Code/SimpleTimer
26. Singh, N., Sinha, A.: Chaos-based secure communication system using logistic map. Opt. Lasers Eng. **48**, 398–404 (2010)
27. Taylor, D.S.: Design of continuously variable slope delta modulation communication systems. Motorola Technical Document AN1544 (1996)
28. Ursulean, R.: Reconsidering the generalized logistic map as a pseudo random bit generator. Elektronika ir Elektrotechnila **7**(56), 100–113 (2004)
29. Verhulst, P.F.: Recherches mathématiques sur la loi d'acroissement de la population. Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique **18**, 1–38 (1845)
30. Verhulst, P.F.: Deuxième mémoire sur la loi d'acroissement de la population. Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique **20**, 1–32 (1847)
31. Volos, C.K.: Chaotic random bit generator realized with a mocrocontroller. J. Comput. Model. **3**(4), 115–136 (2013)
32. Volos, C.K., Doukas, N., Kyprianidis, I.M., Stouboulos, I.N., Kostis, T.G.: Chaotic autonomous mobile robot for military missions. In: The 17th International Conference on Communications, Rhodes Island, July 2013
33. Wolf, A., Swift, J.B., Swinney, H.L., Vastano, J.A.: Determining Lyapunov exponents form a time series. Phys. D **16**, 285–317 (1985)
34. Yang, T., Chua, L.O.: Impulsive stabilization for control and synchronization of chaotic systems: theory and application to secure communication. IEEE Trans. Circuits Syst. I Fundam. Theory Appl. **44**(10), 976–988 (1997)
35. Zapateiro, M., Vidal, Y., Acho, L.: A secure communication scheme based on chaotic duffing oscillators and frequency estimation for the transmission of binary-coded messages. Commun. Nonlinear Sci. Numer. Simul. **19**(4), 991–1003 (2014)
36. Zapateiro De la Hoz, M., Acho, L., Vidal, Y.: A modified Chua chaotic oscillator and its application to secure communications. Appl. Math. Comput. **247**, 712–722 (2014)
37. Zhang, X., Cao, Y.: A novel chaotic map and an improved chaos-based image encryption scheme. Sci. World J. **2014** (2014). http://www.dx.doi.org/10.1155/2014/713541
38. Zhang, L., Liao, X., Wang, X.: An image encryption approach based on chaotic maps. Chaos, Solitons Fractals **24**, 759–765 (2005)