Springer Optimization and Its Applications 113

Nicholas J. Daras
Themistocles M. Rassias *Editors*

Operations Research, Engineering, and Cyber Security

Trends in Applied Mathematics and Technology

Springer

Springer Optimization and Its Applications

VOLUME 113

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

Nicholas J. Daras • Themistocles M. Rassias
Editors

Operations Research, Engineering, and Cyber Security

Trends in Applied Mathematics and Technology

Springer

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

Themistocles M. Rassias
Department of Mathematics
National Technical University of Athens
Athens, Greece

ISSN 1931-6828 ISSN 1931-6836 (electronic)
Springer Optimization and Its Applications
ISBN 978-3-319-51498-7 ISBN 978-3-319-51500-7 (eBook)
DOI 10.1007/978-3-319-51500-7

Library of Congress Control Number: 2017932376

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

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Operations Research, Engineering, and Cyber Security: Trends in Applied Mathematics and Technology brings together a variety of mathematical methods and theories with several applications from a number of disciplines. It discusses new scientific perspectives of an interdisciplinary nature that pertain to several domains of research from pure and applied mathematical sciences including operations research, engineering, and cyber security.

The book presents 18 papers written by eminent scientists from the international mathematical community. Some representative papers in this book had been communicated during the International Conference held at the Hellenic Artillery School in May 2015.

These contributions focus on new developments of mathematical sciences with emphasis to the solvability of the direct electromagnetic scattering problem, geometric approaches to cyber security, ellipsoid targeting with overlap, nonequilibrium solutions of dynamic networks, measuring ballistic dispersion, elliptic regularity theory for the numerical solution of variational problems, approximation theory for polynomials on the real line and the unit circle, complementarity and variational inequalities in electronics, new two-slope parameterized achievement scalarizing functions for nonlinear multiobjective optimization, and strong and weak convexity of closed sets in a Hilbert space. Furthermore, two papers provide expositions on optimization problems related to security in network systems as well as an investigation of some recent inequalities for relative operator entropy. Some papers in this volume could be particularly useful for a broader readership, specifically in the optimal batch production with time-varying demand over finite planning horizon, electromagnetic compatibility in challenging environment, cybersecurity investments with budget constraints, region-based watermarking for images, optimal inventory policies for finite horizon inventory models with time-varying demand, metrical Pareto efficiency, and monotone Ekeland's variational principle.

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

Vari Attikis, Greece	Nicholas J. Daras
Athens, Greece	Themistocles M. Rassias

Contents

Contributors

Khalid Addi University of La Reunion, PIMENT, Sainte-Clotilde, Reunion, France

C.E. Athanasiadis Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece

Lakdere Benkherouf Department of Statistics and Operations Research, Kuwait University, Al-khaldiya, Kuwait

Jack Brimberg Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON, Canada

Kenier Castillo CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal

C. Christopoulos Emeritus Professor of Electrical Engineering, University of Nottingham, Nottingham, UK

Monica-Gabriela Cojocaru University of Guelph, Guelph, ON, Canada

Patrizia Daniele Department of Mathematics and Computer Science, University of Catania, Catania, Italy

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

Bhaskar DasGupta Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

C.T.J. Dodson School of Mathematics, University of Manchester, Manchester, UK

Silvestru Sever Dragomir Mathematics, College of Engineering & Science, Victoria University, Melbourne, Australia
School of Computer Science and Applied Mathematics, University of Witwatersrand, Johannesburg, South Africa

Axel Dreves Department of Aerospace Engineering, Universität der Bundeswehr München, München, Germany

Daniel Goeleven University of La Reunion, PIMENT, Sainte-Clotilde, Reunion, France

Vladimir V. Goncharov CIMA, Universidade de Évora, Évora, Portugal

Georgios Goudelis National Technical University of Athens, Zografou, Greece

Scott Greenhalgh Queen's University, Kingston, ON, Canada

Joachim Gwinner Department of Aerospace Engineering, Universität der Bundeswehr München, München, Germany

W.J. Hurley Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON, Canada

Grigorii E. Ivanov Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
Institute of Systems Dynamics and Control Theory of Siberian Branch of RAS, Irkutsk, Russia

Stefanos D. Kollias National Technical University of Athens, Zografou, Greece

Ioannis Konstantaras Department of Business Administration, School of Business Administration, University of Macedonia, Thessaloniki, Greece

Marko M. Mäkelä Department of Mathematics and Statistics, University of Turku, Turku, Finland

Francisco Marcellán Departamento de Matemáticas, Universidad Carlos III de Madrid, Leganés, Madrid, Spain

Antonino Maugeri Department of Mathematics and Computer Science, University of Catania, Catania, Italy

Anna Nagurney Isenberg School of Management, University of Massachusetts, Amherst, MA, USA

Yury Nikulin Department of Mathematics and Statistics, University of Turku, Turku, Finland

Klimis S. Ntalianis Technical Educational Institute of Athens, Egaleo, Greece

Nina Ovcharova Department of Aerospace Engineering, Universität der Bundeswehr München, München, Germany

Nikolaos Papadakis Hellenic Military Academy, Vari Attikis, Greece

Andrey Pavlov Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, ON, Canada

Konstantinos A. Raftopoulos National Technical University of Athens, Zografou, Greece
The American College of Greece, Agia Paraskevi, Greece

Jorge Rivero Departamento de Matemáticas, Universidad Carlos III de Madrid, Leganés, Madrid, Spain
Instituto de Ciencias Matemáticas (ICMAT) Campus de Cantoblanco, UAM, Madrid, Spain

V. Sevroglou Department of Statistics and Insurance Science, University of Piraeus, Piraeus, Greece

Konstantina Skouri Department of Mathematics, University of Ioannina, Ioannina, Greece

K.I. Skourogiannis Department of Mathematics, National and Kapodistrian University of Athens, Athens, Greece

Venkatkumar Srinivasan Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

Mihai Turinici "A. Myller" Mathematical Seminar, "A. I. Cuza" University, Iaşi, Romania

Paraskevi Tzouveli National Technical University of Athens, Zografou, Greece

Outi Wilppu Department of Mathematics and Statistics, University of Turku, Turku, Finland

Complementarity and Variational Inequalities in Electronics

Khalid Addi and Daniel Goeleven

Abstract The purpose of this chapter is to review and describe the main mathematical models applicable to the study of electrical networks involving devices like diodes whose Ampere–Volt characteristics are set-valued graphs. The mathematical models in question are related to complementarity problems, variational inequalities, and non-regular dynamical systems.

Introduction

In this expository work, we review and discuss some methodology that has been recently developed by several authors for the rigorous formulation and the mathematical analysis of circuits in electronics like slicers, amplitude selectors, sampling gates, operational amplifiers, four-diode bridge full-wave rectifiers, etc. All these circuits use semiconductors like diodes and transistors leading to some highly nonlinear phenomena like switching and clipping. The peculiarity of devices like diodes is that their Ampere-Volt characteristics are described by graphs including vertical branches. Such graphs are thus set-valued and their mathematical treatment requires the use of appropriate tools. The objective of this work is to explain to engineers and mathematicians how advanced tools from convex analysis can be used to build rigorous mathematical models for the qualitative study and numerical simulation of electrical networks involving devices like diodes and transistors. Our objective is also to show that mathematical models like complementarity problems, variational inequalities, and differential inclusions can be used to analyze diverse problems in electronics. These last models are indeed well known for their applications in mechanics and economics but we show here that electronics is also an important source applications. We will review the main mathematical models applicable to the study of electrical networks involving devices like diodes and transistors. It is, however, not our intention here to discuss theoretical mathematical results like

K. Addi (✉) • D. Goeleven
University of La Reunion, PIMENT, Sainte-Clotilde 97715, Reunion, France
e-mail: khalid.addi@univ-reunion.fr; daniel.goeleven@univ-reunion.fr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_1

existence and uniqueness of a solution or stability of a stationary solution, but in developing our subject, we will refer the reader to the appropriate articles. Mathematical models like complementarity problems, variational inequalities, and non-regular dynamical systems are indeed particularly useful to characterize the qualitative properties of the circuits (see [5–7, 18, 21–23, 25–27, 39–41, 50]) as well as to compute some defined output signal (see [1–4, 24, 30, 33]). Such mathematical models are also useful for the determination of the stationary points of dynamical circuits and to determine the corresponding Lyapunov stability and attractivity properties (see [8, 16, 17, 35]) a topic of major importance for further dynamical analysis and control applications (see [11, 14, 15, 19, 42]). Hemivariational inequalities are also important mathematical models that can be used to study electrical networks involving devices like thyristors (see [5]).

On the Use of Complementarity Problem in Electronics

In this section, we show how complementarity problems can be used to develop a suitable approach for the formulation and mathematical analysis of electrical networks involving devices like ideal diodes. For $U, V \in \mathbb{R}^n$, the notation $\langle U, V \rangle = \sum_{i=1}^{n} U_i V_i$ is used to denote the euclidean scalar product on $\mathbb{R}^n$ and $\|U\| = \sqrt{\langle U, U \rangle}$ to denote the corresponding norm. The identity mapping on $\mathbb{R}^n$ will be denoted by $id_{\mathbb{R}^n}$ while the identity matrix of order n is denoted by $I_{n\times n}$. We set $\mathbb{R}^n_+ = [0, +\infty[^n$ and we denote by "$\leq$" the ordering defined by $\mathbb{R}^n_+$, i.e., $U \leq V$ if and only if $V - U \in \mathbb{R}^n_+$. We will also use the notations:

$$\min\{U, V\} = \begin{pmatrix} \min\{U_1, V_1\} \\ \min\{U_2, V_2\} \\ \vdots \\ \min\{U_n, V_n\} \end{pmatrix}, \ \max\{U, V\} = \begin{pmatrix} \max\{U_1, V_1\} \\ \max\{U_2, V_2\} \\ \vdots \\ \max\{U_n, V_n\} \end{pmatrix}.$$

The Complementarity Relation

We say that two vectors $U, V \in \mathbb{R}^n$ satisfy the complementarity relation provided that

$$U \geq 0, V \geq 0 \text{ and } \langle U, V \rangle = 0.$$

The equation $\langle U, V \rangle = 0$ being an orthogonality condition, we also present the complementarity relation as

$$0 \leq U \perp V \geq 0$$

or also

$$\mathbb{R}^n_+ \ni U \perp V \in \mathbb{R}^n_+.$$

It is easy to check that the complementarity relation model is the following set of relations:

$$\begin{cases} (\forall i): & U_i \geq 0 \\ (\forall i): & V_i \geq 0 \\ (\forall i): & U_i > 0 \implies V_i = 0 \\ (\forall i): & V_i > 0 \implies U_i = 0 \end{cases}$$

which is equivalent to the equation

$$\min\{U, V\} = 0.$$

The Complementarity Relation in Electronics

The diode is a device that constitutes a rectifier which permits the easy flow of charges in one direction but restrains the flow in the opposite direction. Diodes are used in power electronics applications like rectifier circuits, switching, and inverter and converter circuits. Figure 1 illustrates the Ampere-Volt characteristic of an ideal diode. This kind of diode is a simple switch. If $V < 0$, then $i = 0$ and the diode is blocking while if $i > 0$, then $V = 0$ and the diode is conducting. We see that the ideal diode is described by the complementarity relation.

$$V \leq 0,\ i \geq 0,\ Vi = 0 \Leftrightarrow 0 \leq -V \perp i \geq 0.$$

The Complementarity Problem

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a given function. The complementarity problem consists to find $x \in \mathbb{R}^n$ such that x and $F(x)$ satisfy the complementarity relation

$$\begin{cases} x \geq 0 \\ F(x) \geq 0 \\ \langle x, F(x) \rangle = 0 \end{cases}$$

$$\Leftrightarrow 0 \ \leq \ x \perp F(x) \ \geq 0 \Leftrightarrow \mathbb{R}^n_+ \ni x \perp F(x) \in \mathbb{R}^n_+.$$

The complementarity problem is equivalent to the equation

$$\min\{x, F(x)\} = 0.$$

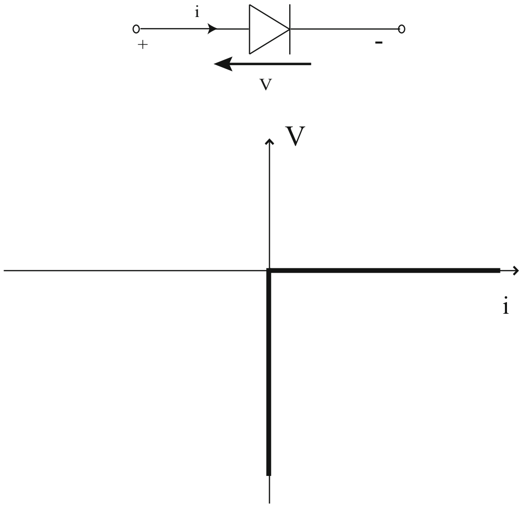


Fig. 1 Ideal diode model

Let $\alpha > 0$, it is also possible to give an equivalent fixed point formulation of the complementarity problem as follows:

$$0 \leq x \perp F(x) \geq 0 \Leftrightarrow \min\{x, F(x)\} = 0$$
$$\Leftrightarrow \max\{-x, -\alpha F(x)\} = 0 \Leftrightarrow x = \max\{0, x - \alpha F(x)\}.$$

Recall also that if $F = \nabla G$ for some $G \in C^1(\mathbb{R}^n; \mathbb{R})$, then any solution x^* of the optimization problem:

$$\min_{x \in \mathbb{R}^n_+} G(x)$$

satisfies the complementarity problem $0 \leq x^* \perp F(x^*) \geq 0$. The converse is also true provided that G is convex. The complementarity mathematical theory has known important developments. Both qualitative results and numerical methods have been developed by several authors in using tools from convex analysis, optimization, and fixed point theory. We refer the readers to the following books [28, 32, 44] and [53] where various results in the field are discussed.

The Complementarity Problem in Electronics

Theoretical tools from complementarity theory can be used to develop a rigorous mathematical study of electrical networks involving devices like ideal diodes. We present here only one example because the variational inequality model that we will discuss in the following section is more general and recovers the complementarity model. The use of complementarity problems in electronics originates in different papers devoted to the mathematical study of dynamical systems in which certain variables are coupled by means of a static piecewise linear characteristic (see, e.g., [21, 22, 24–26, 39–42, 48, 50]).

A Clipping Circuit with Ideal Diode

Let us consider the circuit of Fig. 2 involving a load resistance $R > 0$, an input-signal source u, corresponding instantaneous current i, an ideal diode as a shunt element, and a supply voltage E. Kirchhoff's voltage law gives

$$u = U_R + V + E$$

where $U_R = Ri$ denotes the difference of potential across resistor and V is the difference of potential across diode. Thus

$$\begin{aligned} &0 \leq i \perp E + Ri - u \geq 0 \Leftrightarrow \min\{i, E - u + Ri\} = 0 \\ &\Leftrightarrow \min\{i, \frac{E-u}{R} + i\} = 0 \Leftrightarrow i + \min\{0, \frac{E-u}{R}\} = 0 \\ &\Leftrightarrow i = -\min\{0, \frac{E-u}{R}\} = \frac{1}{R}\max\{0, u - E\}. \end{aligned}$$

If $u \leq E$, then the diode is blocking while if $u > E$, then the diode is conducting. Let us now consider a driven time depending input $t \mapsto u(t)$ and define the output-signal $t \mapsto V_o(t)$ as

$$V_o(t) = E + V(t).$$

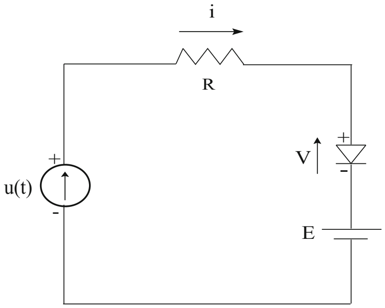


Fig. 2 Clipping circuit 1: diode as shunt element

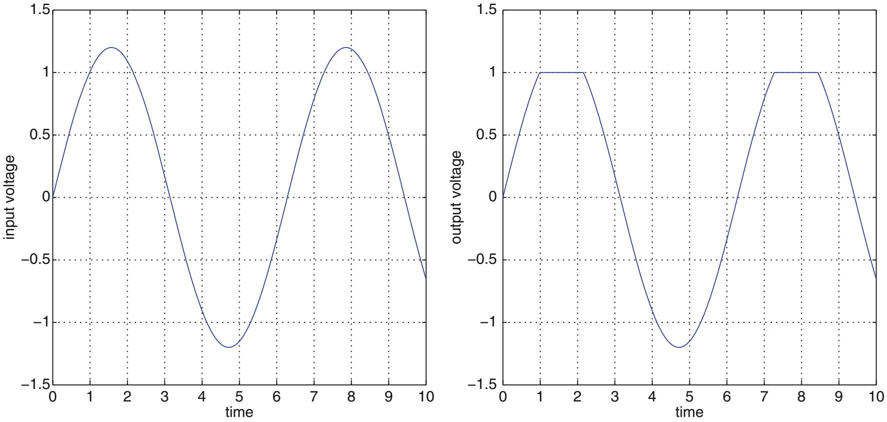


Fig. 3 Clipping circuit 1: ideal diode as shunt element, $E = 1$

The time depending current $t \mapsto i(t)$ is given by

$$i(t) = \frac{1}{R}\max\{0, u(t) - E\} \tag{1}$$

and thus

$$V_o(t) = V(t) + E = u(t) - Ri(t) = u(t) + \min\{0, E - u(t)\} = \min\{u(t), E\}. \tag{2}$$

This shows that the circuit in Fig. 2 can be used to transmit the part of a given input-signal u which lies below some given reference level E (Fig. 3).

On the Use of Variational Inequalities in Electronics

In this section, we show how variational inequalities can be used to develop a suitable method for the formulation and mathematical analysis of electrical networks involving devices like different types of diodes (not necessarily ideal) and transistors.

The Convex Subdifferential Relation

We denote by $\Gamma_0(\mathbb{R}^n; \mathbb{R} \cup \{+\infty\})$ the set of proper, convex, and lower semicontinuous functions from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$. The domain $D(\Phi)$ of Φ is defined by

$$D(\Phi) = \{x \in \mathbb{R}^n : \Phi(x) < +\infty\}.$$

Let $\Phi \in \Gamma_0(\mathbb{R}^n; \mathbb{R} \cup \{+\infty\})$ be given. The convex subdifferential $\partial\Phi(x)$ (see, e.g., [43, 59]) of Φ at x is defined by

$$\partial\Phi(x) = \{w \in \mathbb{R}^n : \Phi(v) - \Phi(x) \geq \langle w, v - x\rangle, \forall v \in \mathbb{R}^n\}.$$

The set $\partial\Phi(x)$ describes the differential properties of Φ by means of the supporting hyperplanes to the epigraph of Φ at $(x, \Phi(x))$. Let $\Phi \in \Gamma_0(\mathbb{R}^n; \mathbb{R} \cup \{+\infty\})$ be given. The Fenchel transform Φ^* of Φ is the function defined by

$$(\forall z \in \mathbb{R}^n) : \ \Phi^*(z) = \sup_{x \in D(\Phi)} \{\langle x, z\rangle - \Phi(x)\}.$$

The function $\Phi^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper convex and lower semicontinuous. A well-known result in convex analysis (see, e.g., [43, 59]) ensures that

$$z \in \partial\Phi(x) \Longleftrightarrow x \in \partial\Phi^*(z) \Longleftrightarrow \Phi(x) + \Phi^*(z) = \langle x, z\rangle.$$

We say that $U, V \in \mathbb{R}^n$ satisfy a convex subdifferential relation provided that

$$(\forall U \in \mathbb{R}^n) : \ V \in \partial\Phi(U),$$

for some $\Phi \in \Gamma_0(\mathbb{R}^n; \mathbb{R} \cup \{+\infty\})$.

The Convex Subdifferential Relation and the Complementarity Relation

Let $K \subset \mathbb{R}^n$ be a nonempty closed convex set. We denote by Ψ_K the indicator function of K, that is:

$$\Psi_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K. \end{cases} \tag{3}$$

Then

$$\partial\Psi_K(x) = \begin{cases} \{w \in \mathbb{R}^n : \langle w, h - x\rangle \leq 0, \forall h \in K\} & \text{if } x \in K \\ \emptyset & \text{if } x \notin K. \end{cases}$$

We may use this last result to prove that the complementarity relation can be written equivalently as a convex subdifferential relation. More precisely:

$$0 \leq U \perp V \geq 0 \Leftrightarrow -V \in \partial\Psi_{\mathbb{R}^n_+}(U).$$

Indeed, let $U, V \in \mathbb{R}^n$ satisfying the complementarity relation $0 \leq U \perp V \geq 0$. Then $(\forall h \geq 0) : \langle V, h\rangle \geq 0$ and since $\langle V, U\rangle = 0$, we see that $(\forall h \geq 0) : \langle V, h - U\rangle \geq 0$ meaning that $-V \in \partial\Psi_{\mathbb{R}^n_+}(U)$. Reciprocally, if $-V \in \partial\Psi_{\mathbb{R}^n_+}(U)$, then $U \geq 0$ and $(\forall h \geq 0) : \langle V, h - U\rangle \geq 0$. Setting $h = 2U$ we obtain $\langle V, U\rangle \geq 0$. Then setting $h = 0$, we get $\langle V, U\rangle \leq 0$. Thus $V \perp U$. Moreover, for any $H \geq 0$, we may set $h = H + U$ to see that $\langle V, H\rangle \geq 0$. It results that $V \geq 0$. Thus $U, V \in \mathbb{R}^n$ satisfy the complementarity relation.

The Convex Subdifferential Relation in Electronics

Electrical devices like diodes are described in terms of Ampere-Volt characteristics (i, V) that is a graph expressing the difference of potential V across the device as a function of current i through the device. The schematic symbol of a circuit element is given in Fig. 4. The conventional current flow i will be depicted on the conductor in the direction of the arrow and the potential $V = V_A - V_B$ where V_A (resp. V_B) denotes the potential of point A (resp. B) across the device will be denoted alongside the device. Experimental measures as well as empirical and physical models lead to a variety of monotone graphs that may present vertical branches. The reader can find general descriptions of devices and Ampere-Volt characteristics either in the appropriate electronics literature (see, e.g., [10, 49]) or in the various electronics society catalogs available on the web.

Let us suppose here that we may write

$$(\forall i \in \mathbb{R}) : V \in \mathscr{F}(i),$$

for some set-valued function $\mathscr{F} : \mathbb{R} \rightrightarrows \mathbb{R}$. The domain $D(\mathscr{F})$ of $\mathscr{F}$ is defined by

$$D(\mathscr{F}) = \{x \in \mathbb{R} : \mathscr{F}(x) \neq \emptyset\}.$$

We assume that $\mathscr{F}$ is maximal monotone. That means that $\mathscr{F}$ is monotone, i.e.,

$$\forall\, x_1, x_2 \in D(\mathscr{F}), z_1 \in \mathscr{F}(x_1), z_2 \in \mathscr{F}(x_2) : \ (z_1 - z_2)(x_1 - x_2) \geq 0$$

and the graph $G(\mathscr{F})$ of $\mathscr{F}$, i.e.,

$$G(\mathscr{F}) := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : x \in D(\mathscr{F}),\ y \in \mathscr{F}(x)\}$$

is not properly included in any other monotone subset of $\mathbb{R} \times \mathbb{R}$.

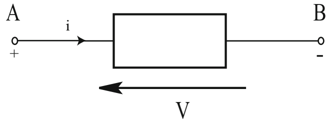


Fig. 4 Electrical device

A classical result (see, e.g., Proposition 1.3.15 in [38]) ensures that there exists a proper, convex, and lower semicontinuous function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ such that

$$(\forall i \in \mathbb{R}) : \mathscr{F}(i) = \partial\varphi(i).$$

Remark 1 Note that there exists $-\infty \leq a \leq b \leq +\infty$ such that $]a, b[\subset D(\mathscr{F}) \subset [a, b]$ and φ can be determined by the formula

$$\varphi(i) = \begin{cases} \int_{i_0}^{i} \beta^0(s)ds & \text{if } \ i \in [a, b] \\ +\infty & \text{if } \ i \in \mathbb{R}\backslash[a, b] \end{cases} \tag{4}$$

where $i_0 \in]a, b[$ and $\beta^0 : D(\mathscr{F}) \to \mathbb{R}$ denotes the minimal section of $\mathscr{F}$, i.e., $\beta^0(x) \in \mathscr{F}(x)$ and $|\beta^0(x)| = \inf\{|w| : w \in \mathscr{F}(x)\}$. Remark that the function φ in (4) is determined by $\mathscr{F}$ up to an additive constant.

Note also that

$$(\forall i \in]a, b[) : \partial\varphi(i) = \left[\beta^0(i^-), \beta^0(i^+)\right],$$

where

$$\beta^0(i^-) = \lim_{z \to i, z < i} \beta(z)$$

and

$$\beta^0(i^+) = \lim_{z \to i, z > i} \beta(z).$$

Any Ampere-Volt characteristic that can be described by a maximal monotone graph can thus also be formulated as a convex subdifferential relation

$$V \in \partial\varphi(i)$$

for some $\varphi \in \varGamma_0(\mathbb{R}; \mathbb{R} \cup \{+\infty\})$. Recall also that

$$V \in \partial\varphi(i) \iff i \in \partial\varphi^*(V) \iff \varphi(i) + \varphi^*(V) = iV.$$

The function φ will be called the electrical superpotential (determined up to an additive constant) of the device. Roughly speaking, the electrical superpotential φ appears as a "primitive" of $\mathscr{F}$ in the sense that the "derivative" (in the generalized sense determined by the convex subdifferential) of φ recovers the set-valued function $\mathscr{F}$.

The notion of superpotential has been introduced by Moreau [51] for convex but generally non-differentiable energy functionals so as to manage nonlinear phenomena like unilateral contact and Coulomb friction. This approach has led to a major generalization of the concept of superpotential by Panagiotopoulos [54] so as to recover the case of non-convex energy functionals. The approach of Moreau as well as the one of Panagiotopoulos is now well-established and often used for the treatment of various problems in elasticity, plasticity, fluid mechanics, and robotics (see, e.g., [37, 38, 52, 54] and [55]). More recently, the superpotential approach of Moreau and Panagiotopoulos has been used to develop a suitable method for the formulation and mathematical analysis of circuits involving devices like diodes, diacs, and thyristors in [5]. The case of circuits with transistors has been studied in [34] and a mathematical general theory applicable to a large class of electrical networks has been developed in [6].

Ideal Diode Model

Let us come back again in this section to the ideal diode model. Figure 1 illustrates the Ampere-Volt characteristic of this kind of diode. We have previously seen that the ideal diode is described by the complementarity relation

$$0 \leq -V \perp i \geq 0$$

which is equivalent to the convex subdifferential relation

$$V \in \partial \Psi_{R_+}(i).$$

The electrical superpotential of the ideal diode is

$$\varphi_D(x) = \Psi_{R_+}(x).$$

Then

$$\varphi_D^*(z) = \Psi_{R_-}(z).$$

We also have

$$\partial \varphi_D(x) := \begin{cases} \mathbb{R}_- & \text{if } x = 0 \\ 0 & \text{if } x > 0 \\ \emptyset & \text{if } x < 0 \end{cases}$$

and

$$\partial \varphi_D^*(z) := \begin{cases} \mathbb{R}_+ & \text{if } z = 0 \\ 0 & \text{if } z < 0 \\ \emptyset & \text{if } z > 0 \end{cases} .$$

The complementarity relation can thus be written as

$$V \in \partial\varphi_D(i) \Longleftrightarrow i \in \partial\varphi_D^*(V) \Longleftrightarrow \varphi_D(i) + \varphi_D^*(V) = iV.$$

Practical Diode Model

Figure 5 illustrates the Ampere-Volt characteristic of a practical diode model. Figure 5 illustrates the Ampere-Volt characteristic of a practical diode model. There is a voltage point, called the knee voltage V_1, at which the diode begins to conduct and a maximum reverse voltage, called the peak reverse voltage V_2, that will not force the diode to conduct. When this voltage is exceeded, the depletion may breakdown and allow the diode to conduct in the reverse direction. Note that usually $\mid V_2 \mid >> \mid V_1 \mid$ and the model is locally ideal. For general purpose diodes used in low frequency/speed applications, $\mid V_1 \mid \simeq 0.7$–2.5 V and $\mid V_2 \mid \simeq 5$ kV; for high voltage rectifier diodes, $\mid V_1 \mid \simeq 10$ V and $\mid V_2 \mid \simeq 30$ kV; for fast diodes used in switched mode power supply and inverter circuits, $\mid V_1 \mid \simeq 0.7$-1.5 V and $\mid V_2 \mid \simeq 3$ kV and for Schottky diodes used in high frequency applications, $\mid V_1 \mid \simeq 0.2$–0.9 V and $\mid V_2 \mid \simeq 100$ V.

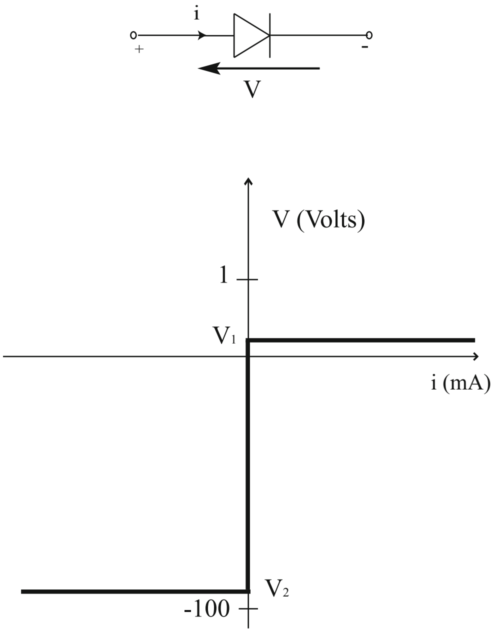


Fig. 5 Practical diode model

The electrical superpotential of the practical diode is

$$\varphi_{PD}(x) = \begin{cases} V_1 x & \text{if } x \geq 0 \\ V_2 x & \text{if } x < 0 \end{cases}.$$

Then

$$\varphi^*_{PD}(z) = \Psi_{[V_2,V_1]}(z).$$

We see that

$$\partial\varphi_{PD}(x) = \begin{cases} V_2 & \text{if } x < 0 \\ [V_2, V_1] & \text{if } x = 0 \\ V_1 & \text{if } x > 0 \end{cases}$$

recovers the Ampere-Volt characteristic (i, V) while

$$\partial\varphi^*_{PD}(z) = \begin{cases} \mathbb{R}_- & \text{if } z = V_2 \\ 0 & \text{if } z \in]V_2, V_1[\\ \mathbb{R}_+ & \text{if } z = V_1 \\ \emptyset & \text{if } z \in \mathbb{R}\setminus [V_2, V_1] \end{cases}$$

recovers the volt-ampere characteristic (V, i). The Ampere-Volt characteristic of the practical diode can thus be written as

$$V \in \partial\varphi_{PD}(i) \Longleftrightarrow i \in \partial\varphi^*_{PD}(V) \Longleftrightarrow \varphi_{PD}(i) + \varphi^*_{PD}(V) = iV.$$

Complete Diode Model

Figure 6 illustrates a complete diode model which includes the effect of the natural resistance of the diode, called the bulk resistance, the reverse current I_{R_1}, the diode capacitance, and the diffusion current. This last model is more accurate and represents the true operating characteristics of the diode.

Note that $\mid V_4 \mid << \mid V_1 \mid$. For example, the 10ETS.. rectifier (SAFEIR series) has been designed with $\mid V_1 \mid = 1.1\,\text{V}$, $\mid V_4 \mid = 800\text{–}1600\,\text{V}$, $I_{R1} = 0.05\,\text{mA}$ and with a bulk resistance equal to $20\,\text{m}\Omega$. Let us use the notation of Fig. 6. It is implicitly assumed that

$$I_{R2} < 0 < I_{R1}, \; V_4 < V_2 < 0 < V_1 < V_3.$$

Let us also set

$$\alpha := \frac{(V_3 - V_1)}{(I_{R3} - I_{R1})}, \; \beta := \frac{(I_{R1}V_3 - I_{R3}V_1)}{(I_{R3} - I_{R1})}, \; \gamma := \frac{I_{R1}(I_{R1}V_3 - I_{R3}V_1)}{2(I_{R3} - I_{R1})}.$$

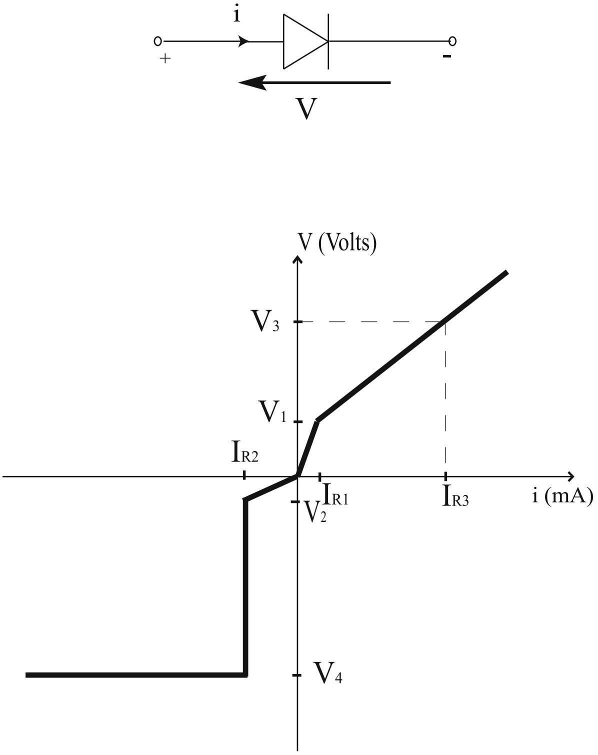


Fig. 6 Complete diode model

The electrical superpotential of the complete diode is

$$\varphi_{CD}(x) = \begin{cases} V_4 x + I_{R2}(\frac{V_2}{2} - V_4) & \text{if } x \leq I_{R2} \\ \frac{V_2}{2I_{R2}} x^2 & \text{if } I_{R2} < x \leq 0 \\ \frac{V_1}{2I_{R1}} x^2 & \text{if } 0 < x \leq I_{R1} \\ \frac{1}{2}\alpha x^2 - \beta x + \gamma & \text{if } I_{R1} < x \end{cases}$$

and simple calculations yield

$$\partial\varphi_{CD}(x) = \begin{cases} V_4 & \text{if } x < I_{R2} \\ [V_4, V_2] & \text{if } x = I_{R2} \\ \frac{V_2}{I_{R2}} x & \text{if } I_{R2} < x \leq 0 \\ \frac{V_1}{I_{R1}} x & \text{if } 0 < x \leq I_{R1} \\ \alpha x - \beta & \text{if } I_{R1} < x. \end{cases} .$$

On the other hand, we may compute the conjugate function

$$\varphi^*_{CD}(z) = \begin{cases} +\infty \text{ if } z \leq V_4 \\ I_{R2}(z - \dfrac{V_2}{2}) \text{ if } V_4 < z \leq V_2 \\ \dfrac{I_{R2}}{2V_2}z^2 \text{ if } V_2 < z \leq 0 \\ \dfrac{I_{R1}}{2V_1}z^2 \text{ if } 0 < z \leq V_1 \\ \dfrac{1}{2}\alpha z^2 + (I_{R_1} - \alpha V_1)z + \dfrac{1}{2}V_1(\alpha V_1 - I_1) \text{ if } V_1 < z \end{cases}$$

and

$$\partial\varphi^*_{CD}(z) = \begin{cases} \emptyset \text{ if } z < V_4 \\ ]-\infty, I_{R2}] \text{ if } z = V_4 \\ I_{R2} \text{ if } V_4 < z \leq V_2 \\ \dfrac{I_{R2}}{V_2}z \text{ if } V_2 < z \leq 0 \\ \dfrac{I_{R1}}{V_1}z \text{ if } 0 < z \leq V_1 \\ \alpha z + (I_{R1} - \alpha V_1) \text{ if } V_1 < z. \end{cases}$$

The Ampere-Volt characteristic of the complete diode can then be written as

$$V \in \partial\varphi_{CD}(i) \Longleftrightarrow i \in \partial\varphi^*_{CD}(V) \Longleftrightarrow \varphi_{CD}(i) + \varphi^*_{CD}(V) = iV.$$

Zener Diode Models

The Zener diodes are made to permit current to flow in the reverse direction if the voltage is larger than the rated breakdown or "Zener voltage" V_2. For example, for a common Zener diode, $V_1 \simeq 0.7\,\text{V}$ and $V_2 \simeq -7\,\text{V}$.

The Zener diode (see Fig. 7) is a good voltage regulator to maintain a constant voltage regardless of minor variations in load current or input voltage. There is a current point I_Z, called the Zener knee current, which is the minimum value of the Zener current required to maintain voltage regulation and a maximum allowable value of Zener current I_M. Currents above this value will damage or destroy the system. The graph corresponding to the Ampere-Volt characteristic (i, V) is maximal monotone and there exists a proper convex and continuous electrical superpotential $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$(\forall i \in \mathbb{R}) : \ V \in \partial\varphi(i).$$

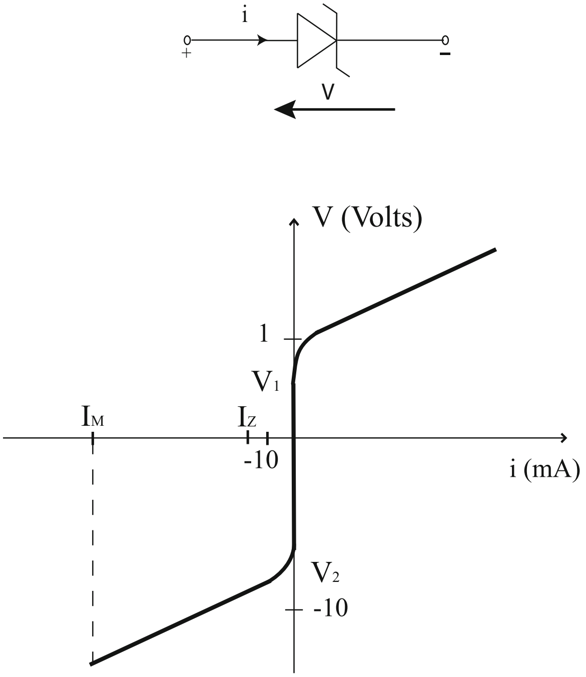


Fig. 7 Zener diode model

The ideal Zener diode model is given by the practical diode model (see Fig. 5) with the appropriate values for V_1 and V_2. This means that the voltage across the diode is constant over a wide range of device current values. The practical Zener diode model (see Fig. 8) is a piecewise linear model that includes the effects of the Zener impedance. Let us use the notation of Fig. 8. It is here implicitly assumed that

$$I_1 < 0 < I_2, \quad V_1 < V_3 < 0 < V_4 < V_2.$$

The electrical superpotential of the Zener diode is

$$\varphi_Z(x) = \begin{cases} \frac{(V_1 - V_3)}{2I_1}x^2 + V_3 x & \text{if } x < 0 \\ \frac{(V_2 - V_4)}{2I_2}x^2 + V_4 x & \text{if } x \geq 0. \end{cases}$$

Then

$$\varphi_Z^*(z) = \begin{cases} \frac{I_1}{2(V_1 - V_3)}(z^2 - 2V_3 z + V_3^2) & \text{if } z < V_3 \\ 0 \text{ if } V_3 \leq z \leq V_4 \\ \frac{I_2}{2(V_2 - V_4)}(z^2 - 2V_4 z + V_4^2) & \text{if } V_4 < z. \end{cases}$$

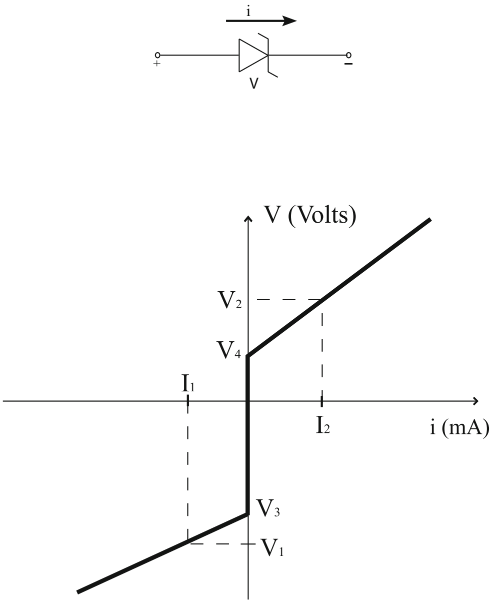


Fig. 8 Practical Zener diode model

Moreover

$$\partial\varphi_Z(x) = \begin{cases} \frac{(V_1 - V_3)}{I_1}x + V_3 \text{ if } x < 0 \\ [V_3, V_4] \text{ if } x = 0 \\ \frac{(V_2 - V_4)}{I_2}x + V_4 \text{ if } x > 0 \end{cases}$$

and

$$\partial\varphi_Z^*(z) = \begin{cases} \frac{I_1}{V_1 - V_3}(z - V_3) \text{ if } z < V_3 \\ 0 \text{ if } V_3 \le z \le V_4 \\ \frac{I_2}{V_2 - V_4}(z - V_4) \text{ if } V_4 < z. \end{cases}$$

The Ampere-Volt characteristic of the concrete Zener diode can thus be written as

$$V \in \partial\varphi_Z(i) \Longleftrightarrow i \in \partial\varphi_Z^*(V) \Longleftrightarrow \varphi_Z(i) + \varphi_Z^*(V) = iV.$$

Varistor Model

A varistor is a nonlinear device that has an electrical behavior similar to the Zener diode (with $| V_1 | = | V_2 |$). More precisely, the varistor (see Fig. 9) is a voltage-dependent resistor with a symmetrical monotone Ampere-Volt characteristic. It is

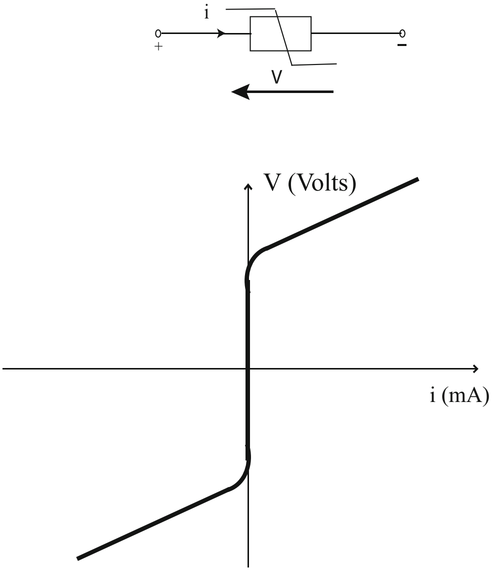


Fig. 9 Varistor

used connected in parallel with the electronic device or circuit that is to be guarded in order to form a low-resistance shunt when voltage increases and thus prevent any further rise in the overvoltage. The graph corresponding to the Ampere-Volt characteristic (i, V) is maximal monotone and there exists a proper convex and continuous electrical superpotential $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$(\forall i \in \mathbb{R}) : \ V \in \partial\varphi(i).$$

Transistor Models

A junction transistor is a semiconductor triode capable of producing amplification. A P-N-P (resp. N-P-N) transistor consists of a silicon (or germanium) crystal in which a layer of N-type silicon (resp. P-type) is sandwiched between two layers of P-type silicon (resp. N-type). The three portions of transistor are known as emitter, base, and collector.

The behavior of a transistor can be described by means of the Ebers-Moll model (see, e.g., [49]) involving two diodes placed back to back and two dependent current-controlled sources $\alpha_I I_C$ and $\alpha_N I_E$ shunting the diodes. Here $\alpha_N \in [0, 1[$ is known as the current gain in normal operation and $\alpha_I \in [0, 1[$ is known as the inverted common-base gain current. Throughout this paper, we will use the notations and conventions of Figs. 10 and 11.

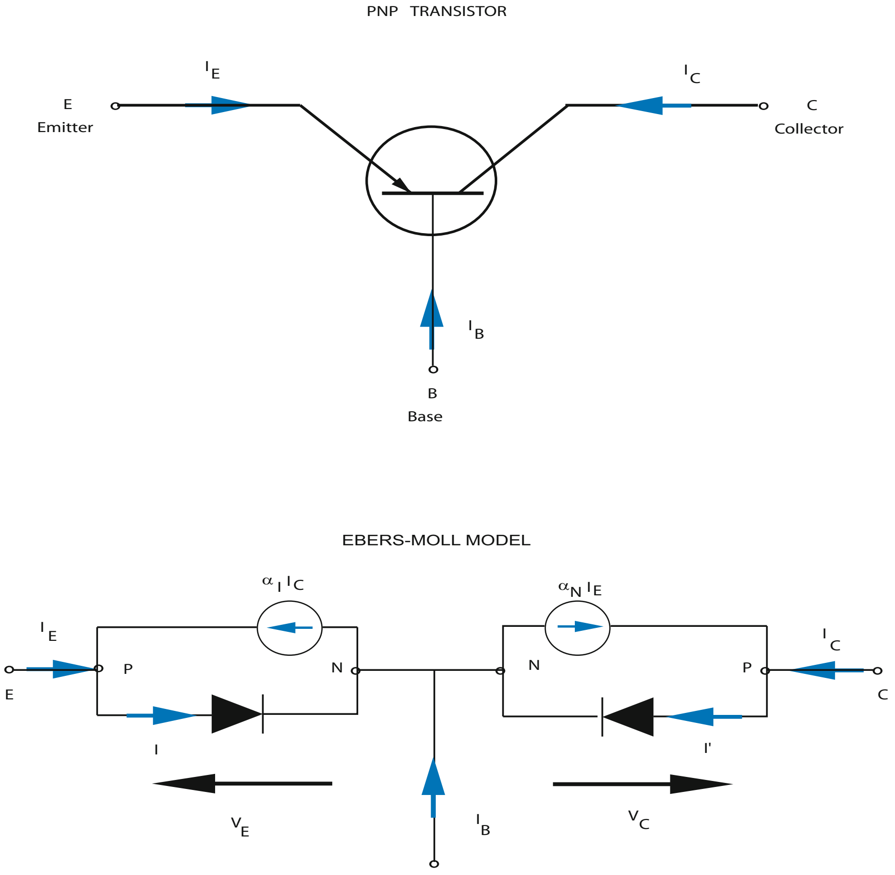


Fig. 10 Transistor P-N-P

Let us here assume that the two diodes of Ebers-Moll are ideal. That means that each diode acts as a simple switch. If $V_E < 0$ (resp. $V_C < 0$), then $I = 0$ (resp. $I' = 0$) and the diode is blocking. If $I > 0$ (resp. $I' > 0$), then $V_E = 0$ (resp. $V_C = 0$) and the diode is conducting. We may then write $V_E \leq 0,\ I \geq 0,\ V_E I = 0$ and $V_C \leq 0,\ I' \geq 0,\ V_C I' = 0$. That is also:

$$\begin{pmatrix} -V_E \\ -V_C \end{pmatrix} \geq 0, \begin{pmatrix} I \\ I' \end{pmatrix} \geq 0,\ \langle \begin{pmatrix} -V_E \\ -V_C \end{pmatrix}, \begin{pmatrix} I \\ I' \end{pmatrix} \rangle = 0 \tag{5}$$

or equivalently

$$\begin{pmatrix} V_E \\ V_C \end{pmatrix} \in \begin{pmatrix} \partial\psi_{\mathbb{R}_+}(I) \\ \partial\psi_{\mathbb{R}_+}(I') \end{pmatrix}. \tag{6}$$

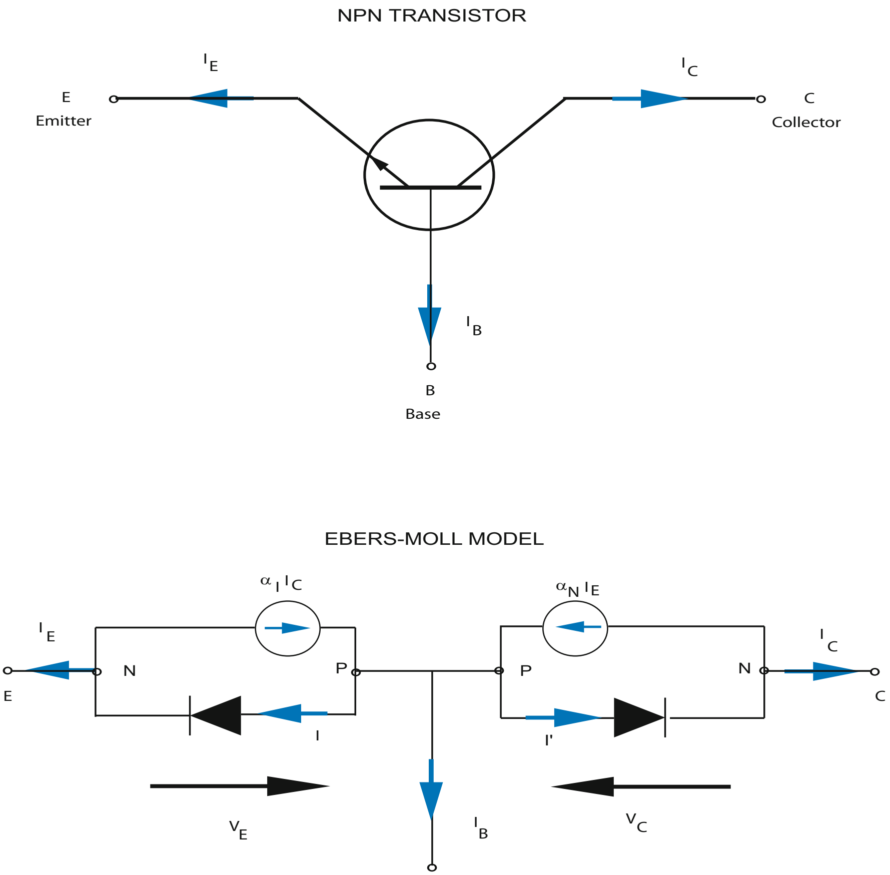


Fig. 11 Transistor N-P-N

Moreover,

$$\begin{pmatrix} I \\ I' \end{pmatrix} = \begin{pmatrix} 1 & \alpha_I \\ \alpha_N & 1 \end{pmatrix} \begin{pmatrix} I_E \\ I_C \end{pmatrix} \tag{7}$$

and

$$I_B = -(I_E + I_C). \tag{8}$$

The relations in (6)–(8) constitute a handy mathematical model for the transistor.

Let us now consider a more general mathematical model in assuming that there exist proper convex and lower semicontinuous functions φ_E, φ_C such that the Ampere-Volt characteristics of the two diodes of Ebers-Moll model can be formulated as

$$V_E \in \partial\varphi_E(I), \quad V_C \in \partial\varphi_C(I').$$

The function φ_E is called the emitter electrical superpotential of the transistor while the function φ_C is named the collector electrical superpotential. The mathematical model of the transistor reads

$$\begin{pmatrix} V_E \\ V_C \end{pmatrix} \in \begin{pmatrix} \partial\varphi_E(I) \\ \partial\varphi_C(I') \end{pmatrix}, \tag{9}$$

$$\begin{pmatrix} I \\ I' \end{pmatrix} = \begin{pmatrix} 1 & \alpha_I \\ \alpha_N & 1 \end{pmatrix} \begin{pmatrix} I_E \\ I_C \end{pmatrix} \tag{10}$$

and

$$I_B = -(I_E + I_C). \tag{11}$$

The different models of diodes that have been discussed in the previous section can be here used so as to define the corresponding models of transistors.

The Variational Inequality Model

Let $\Phi \in \Gamma_0(\mathbb{R}^n; \mathbb{R} \cup \{+\infty\})$ and let $F : \mathbb{R}^n \to \mathbb{R}$ be a given function. The variational inequality problem consists to find $u \in \mathbb{R}^n$ such that

$$\langle F(u), v - u \rangle + \Phi(v) - \Phi(u) \geq 0, \ \forall v \in \mathbb{R}^n. \tag{12}$$

It is easy to see that (12) is equivalent to the convex subdifferential relation

$$F(u) \in -\partial\Phi(u). \tag{13}$$

Problem (12) is called a "variational inequality of the second kind" or "mixed variational inequality" (see, e.g., [32, 37, 47] and [55]). This model recovers the one called "variational inequality of the first kind" which consists to find $u \in C$ such that

$$\langle F(u), v - u \rangle \geq 0, \ \forall v \in C, \tag{14}$$

with C a nonempty closed convex set. It suffices indeed to set $\Phi = \Psi_C$ to see that in this case, (12) is equivalent to (14). Let us also recall here that if $C = \mathbb{R}^n_+$, then (14) is equivalent to the complementarity problem

$$0 \leq u \perp F(u) \geq 0. \tag{15}$$

It is well known that for each $y \in \mathbb{R}^n$, there exists a unique $x \in \mathbb{R}^n$ such that

$$\langle x - y, v - x\rangle + \Phi(v) - \Phi(x) \geq 0, \ \forall v \in \mathbb{R}^n,$$

that is

$$y \in x + \partial\Phi(x).$$

The mapping $P_\Phi : \mathbb{R}^n \to \mathbb{R}^n; y \mapsto P_\Phi(y)$, called the proximal operator (see, e.g., [56]), and defined by

$$(\forall y \in \mathbb{R}^n) : \ P_\Phi(y) = (id_{\mathbb{R}^n} + \partial\Phi)^{-1}(y), \tag{16}$$

is thus a well-defined single-valued operator. Moreover, it is easy to check that

$$y \in x + \partial\Phi(x) \Longleftrightarrow x = (id_{\mathbb{R}^n} + \partial\Phi)^{-1}(y) \Longleftrightarrow x = \ \text{argmin}_{v\in\mathbb{R}^n}\{\frac{1}{2}||v - y||^2 + \Phi(v)\}.$$

If C is a nonempty closed convex set, then

$$P_{\Psi_C} \equiv P_C$$

where P_C denotes the projector from $\mathbb{R}^n$ onto C, i.e.,

$$P_C(x) = \ \text{argmin}_{v\in C}\{\frac{1}{2}||v - x||^2\}.$$

Let $\alpha > 0$ be given. Using the proximal operator, we see that (12) can be formulated as an equivalent fixed point problem

$$u = (id_{\mathbb{R}^n} + \partial\Phi)^{-1}(u - \alpha F(u)).$$

Finally, we recall that if $F = \nabla G$ for some $G \in C^1(\mathbb{R}^n; \mathbb{R})$, then any solution x^* of the optimization problem:

$$\min_{x\in\mathbb{R}^n} G(x) + \Phi(x)$$

satisfies the variational inequality

$$\langle F(x), v - x\rangle + \Phi(v) - \Phi(x) \geq 0, \ \forall v \in \mathbb{R}^n.$$

The converse is also true provided that G is convex.

The Variational Inequality Model in Electronics

A circuit in electronics is formed by the interconnection of electrical devices like generators, resistors, capacitors, inductors, transistors, diodes, and various others. The behavior of a circuit is usually described in terms of currents and voltages that can be specified through each involved electrical device. The approach to state a mathematical model that can be used to determine these currents and voltages consists to formulate the Ampere-Volt characteristic of each electrical device, to write the Kirchhoff's voltage law expressing that the algebraic sum of the voltages between successive nodes in all meshes in the circuit are zero and to write the Kirchhoff's current law stating that the algebraic sum of the currents in all branches which converge to a common node equals zero. We will see in this section that general electrical circuits with diodes and transistors can be studied in using the variational inequalities modelling approach.

The approach using variational inequalities of the second kind so as to study electrical networks involving devices like diodes and transistors has been developed in [6] and [34]. The mathematical approach studied in [6] uses recession tools so as to define a new class of problems that is called "semi-complementarity problems" (see also [36]). It is first shown that the study of semi-complementarity problems can be used to prove qualitative results applicable to the study of linear variational inequalities of the second kind. In using variational inequalities of the second kind, the authors study diode circuits like amplitude selectors that are used to transmit the part of a given waveform which lies above or below some given reference level, double-diode clippers that are used to limit the input amplitude at two independent levels, sampling gates which are transmission circuits in which the output is a reproduction of an input waveform during a selected time interval and is zero otherwise and other circuits involving both diodes, transistors, and operational amplifiers. Further theoretical results, applications in electronics and numerical simulations can be find in the following papers: [7, 23] and [33].

A General Clipping Circuit

Let us again consider the circuit of Fig. 2. We discuss here the case of a diode with electrical superpotential φ. Kirchhoff's voltage law gives

$$u = U_R + V + E$$

where $U_R = Ri$ denotes the difference of potential across the resistor and $V \in \partial\varphi(i)$ is the difference of potential across diode (Fig. 12). Thus

$$E + Ri - u \in -\partial\varphi(i) \tag{17}$$

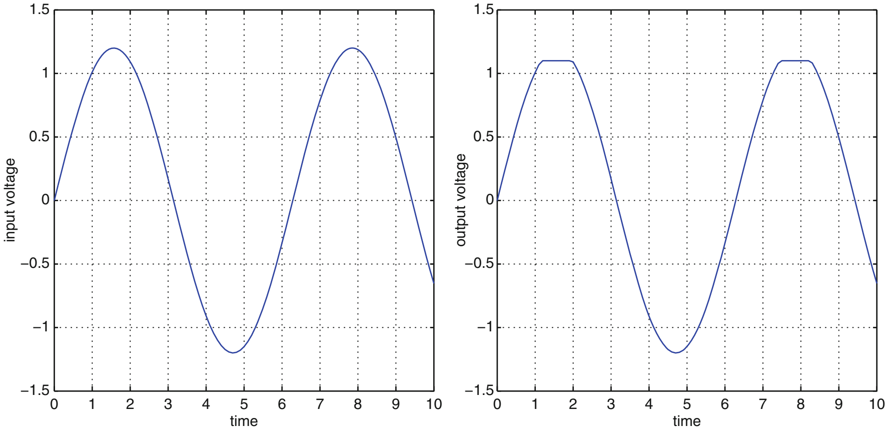


Fig. 12 Clipping circuit 1: practical diode as shunt element using, $V_1 = 0.1, V_2 = -90,\ E = 1$

which is equivalent to the variational inequality

$$(Ri + E - u)(v - i) + \varphi(v) - \varphi(i) \geq 0, \forall v \in \mathbb{R}.$$

Moreover,

$$\frac{E}{R} + i - \frac{u}{R} \in -\frac{1}{R}\partial\varphi(i) \Longleftrightarrow -\frac{E}{R} + \frac{u}{R} \in i + \frac{1}{R}\partial\varphi(i)$$
$$\Longleftrightarrow i = (id_{\mathbb{R}} + \frac{1}{R}\partial\varphi)^{-1}(\frac{u - E}{R}).$$

Let us now consider a driven time depending input $t \mapsto u(t)$ and define the output-signal $t \mapsto V_o(t)$ as

$$V_o(t) = E + V(t) = u(t) - Ri(t).$$

The time depending current $t \mapsto i(t)$ is given by

$$\begin{aligned} i(t) &= (id_{\mathbb{R}} + \tfrac{1}{R}\partial\varphi)^{-1}(\tfrac{u(t)-E}{R}) \\ &= \text{argmin}_{x\in\mathbb{R}}\{\tfrac{1}{2}|x - (\tfrac{u(t)-E}{R})|^2 + \tfrac{1}{R}\varphi(x)\}. \end{aligned} \tag{18}$$

A Rectifier-Stabilizer Circuit

In this section, we illustrate our mathematical modelling approach with a rectifier-stabilizer circuit (Fig. 13). The rectifier-stabilizer circuit involves four diodes D_1, D_2, D_3, and D_4, a Zener diode D_z, an N-P-N transistor T, two resistors R_1 and R_2, and two capacitors C_1 and C_2. This circuit is supplied by the signal input u. We first follow a classical compartmental approach to split it into two blocks: the "rectifier" circuit depicted in Fig. 14 and the "stabilizer" one presented in Fig. 15. We suppose that all diodes of the rectifier block are ideal. We denote by V_i the voltage of diode D_i $(1 \leq i \leq 4)$, V the voltage of the capacitor and use the other notation indicated in Fig. 14. Kirchhoff's laws yield the system

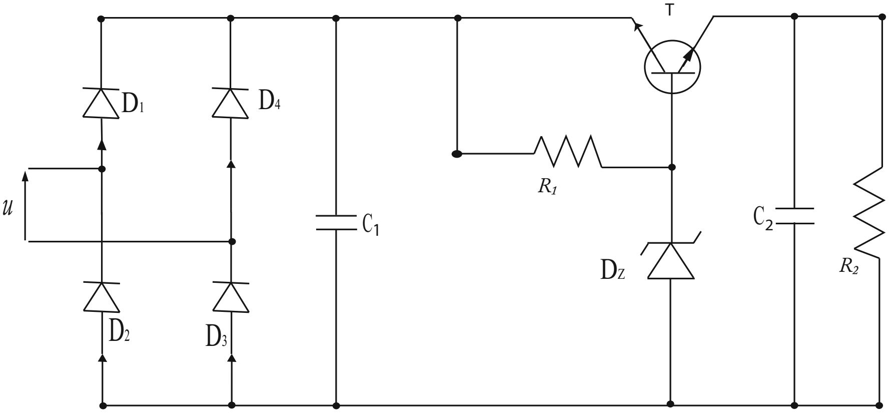


Fig. 13 Rectifier-stabilizer circuit

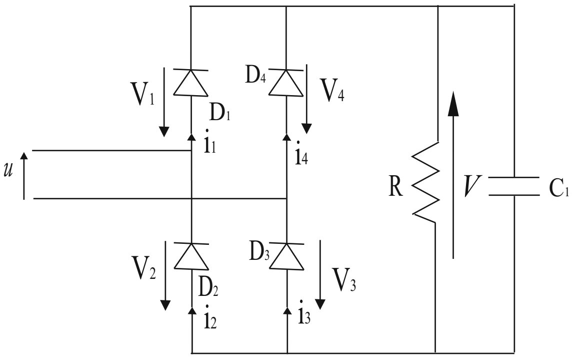


Fig. 14 Rectifier circuit

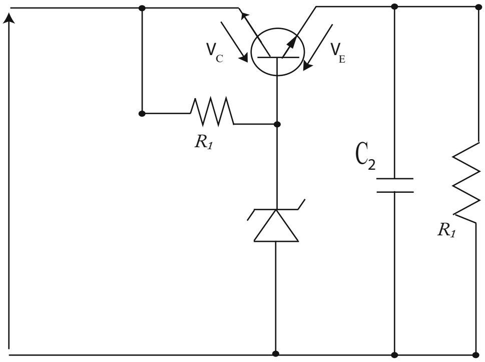


Fig. 15 Stabilizer circuit

$$\begin{cases} i_1 + i_4 = \dfrac{V}{R} + C_1 \dfrac{dV}{dt}, \\ -V_4 \quad = V + V_3, \\ i_3 \quad = i_4 + i_1 - i_2, \\ -V_1 \quad = V + V_3 - u, \\ -V_2 \quad = -V_3 + u. \end{cases}$$

We have

$$\begin{cases} -V_4 \in -\partial \varphi_D(i_4), \\ -V_1 \in -\partial \varphi_D(i_1), \\ -V_2 \in -\partial \varphi_D(i_2), \\ -V_3 \in -\partial \varphi_D(i_3). \end{cases}$$

Moreover $V_3 \in \partial \varphi_D(i_3)$ if and only if $i_3 \in \partial \varphi_D^*(V_3)$. Setting

$$(\forall\ V \in \mathbb{R}) : \ \theta_D(V) = \varphi_D^*(-V)$$

we get

$$(\forall\ V \in \mathbb{R}) : \ \partial \theta_D(V) = -\partial \varphi_D^*(-V)$$

Note that here $\theta_D(V) = \Psi_{\mathbb{R}_-}(-V) = \Psi_{\mathbb{R}_+}(V) = \varphi_D(V)$.
Therefore

$$V_3 \in \partial \varphi_D(i_3) \Leftrightarrow i_3 \in -\partial \varphi_D(-V_3).$$

We also set

$$\Phi(x) = \Psi_{\mathbb{R}_+}(x_1) + \Psi_{\mathbb{R}_+}(x_2) + \Psi_{\mathbb{R}_+}(x_3) + \Psi_{\mathbb{R}_+}(x_4) = \Psi_{(\mathbb{R}_+)^4}$$

It results that the dynamical behavior of the circuit in Fig. 14 is described by

$$\frac{dV}{dt} = \frac{-1}{RC_1}V + \overbrace{\begin{pmatrix} \frac{1}{C_1} & 0 & \frac{1}{C_1} & 0 \end{pmatrix}}^{B} \begin{pmatrix} i_4 \\ -V_3 \\ i_1 \\ i_2 \end{pmatrix}, \tag{19}$$

$$\overbrace{\begin{pmatrix} -V_4 \\ i_3 \\ -V_1 \\ -V_2 \end{pmatrix}}^{y} = \overbrace{\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}}^{C} V + \overbrace{\begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}}^{N} \overbrace{\begin{pmatrix} i_4 \\ -V_3 \\ i_1 \\ i_2 \end{pmatrix}}^{y_L} + \overbrace{\begin{pmatrix} 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}}^{F} u \tag{20}$$

and

$$y \in -\partial \Phi(y_L). \tag{21}$$

At equilibrium, the dynamical circuit in Fig. 14 reduces to the circuit in Fig. 16 and the stationary solutions of (19)–(21) satisfy the problem

$$\begin{cases} -\dfrac{1}{RC_1}V + By_L = 0 \\ \\ \langle Ny_L + CV + Fu, v - y_L \rangle + \Phi(v) - \Phi(y_L) \geq 0, \quad \forall\ v \in \mathbb{R}^4. \end{cases} \tag{22}$$

From the first equation of (22) one deduces that $V = RC_1By_L$, so that $y = (N + \frac{1}{a}CB)y_L + Fu$ and our problem reduces to the variational inequality $VI(M, \Phi, Fu)$

$$y_L \in \mathbb{R}^4 : \ \langle My_L + Fu, v - y_L \rangle + \Phi(v) - \Phi(y_L) \geq 0, \quad \forall\ v \in \mathbb{R}^4. \tag{23}$$

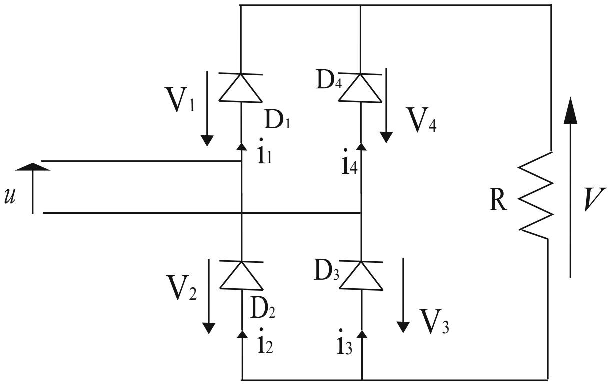


Fig. 16 Rectifier circuit

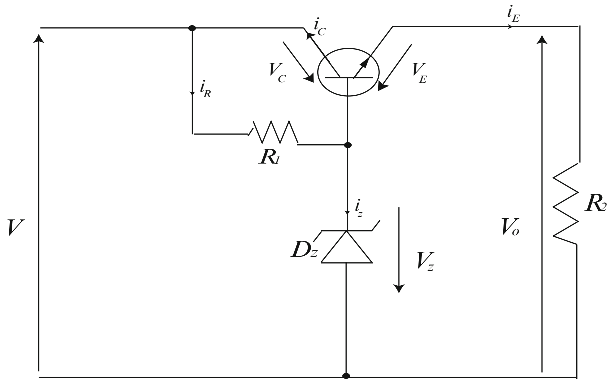


Fig. 17 Stabilizer circuit at equilibrium

with

$$M := N + RC_1CB = \begin{pmatrix} R & -1 & R & 0 \\ 1 & 0 & 1 & -1 \\ R & -1 & R & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Let us now consider the stabilizer block as in Fig. 17.

We denote by V_E, V_C, and V_z the voltages of the transistor and the Zener diode, respectively, as indicated in Fig. 17. Note that we omit the capacitor C_2, thanks to the equilibrium, and use the other notation indicated in Fig. 17. Kirchhoff's laws yield the system

$$\begin{cases} -V_z & = V + -R_1(i_z + i_e + i_c) \\ -V_z - V_C & = V \\ V_E - V_C & = V - R_2 i_e. \end{cases}$$

The N-P-N transistor behavior is described by means of the Ebers-Moll model as given in the previous section while the ideal Zener diode behavior is depicted in Fig. 18.

Setting

$$V_{ze} = V_z - V_s,$$

we see that the ideal Zener diode is then described by the complementarity relation

$$V_{ze} \geq 0, \quad i_z \geq 0, \quad V_{ze} i_z = 0.$$

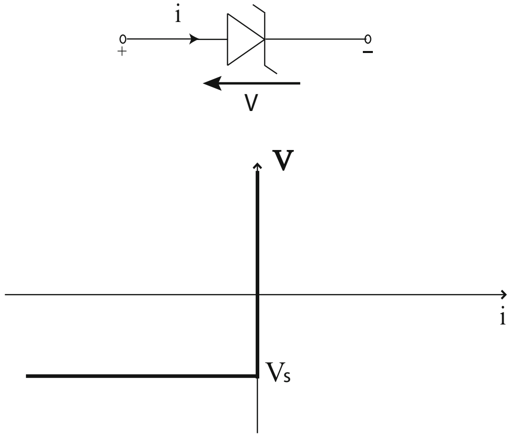


Fig. 18 An ideal Zener Diode

We have

$$\begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} V_{ze} \\ -V_E \\ -V_C \end{pmatrix} = \begin{pmatrix} -R_1 & -R_1 & -R_1 \\ 0 & 0 & 0 \\ 0 & -R_2 & 0 \end{pmatrix}\begin{pmatrix} i_z \\ i_e \\ i_c \end{pmatrix} + \begin{pmatrix} V+V_s \\ V+V_s \\ V \end{pmatrix}.$$

and

$$\begin{pmatrix} i_E \\ i_C \end{pmatrix} = \frac{1}{1-\alpha_I\alpha_N}\begin{pmatrix} 1 & -\alpha_I \\ -\alpha_N & 1 \end{pmatrix}\begin{pmatrix} I \\ I' \end{pmatrix}$$

and thus

$$\begin{pmatrix} -1 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} V_{ze} \\ -V_E \\ -V_C \end{pmatrix} = \frac{1}{K}\begin{pmatrix} -R_1K & R_1(\alpha_N-1) & R_1(\alpha_I-1) \\ 0 & 0 & 0 \\ 0 & -R_2 & R_2\alpha_I \end{pmatrix}\begin{pmatrix} i_z \\ I \\ I' \end{pmatrix}$$
$$+\begin{pmatrix} V+V_s \\ V+V_s \\ V \end{pmatrix},$$

where $K = 1-\alpha_I\alpha_N$. Then

$$\overbrace{\begin{pmatrix} V_{ze} \\ -V_E \\ -V_C \end{pmatrix}}^{w} = \overbrace{\frac{1}{K}\begin{pmatrix} R_1K & R_1(1-\alpha_N) & R_1(1-\alpha_I) \\ R_1K & R_1(1-\alpha_N)+R_2 & R_1(1-\alpha_I)-\alpha_IR_2 \\ R_1K & R_1(1-\alpha_N) & R_1(1-\alpha_I) \end{pmatrix}}^{\Theta}\overbrace{\begin{pmatrix} i_z \\ I \\ I' \end{pmatrix}}^{z}$$

$$+\overbrace{\begin{pmatrix} V+V_s \\ V+V_s \\ V \end{pmatrix}}^{q}.$$

We also have

$$\begin{cases} V_{ze} & \in -\partial \Psi_{\mathbb{R}+}(i_z) \\ -V_E & \in -\partial \Psi_{\mathbb{R}+}(I) \\ -V_C & \in -\partial \Psi_{\mathbb{R}+}(I'). \end{cases}$$

Setting

$$(\forall x \in \mathbb{R}_+):\ \Xi(x) = \Psi_{\mathbb{R}^3_+}(x) = \Psi_{\mathbb{R}_+}(x_1) + \Psi_{\mathbb{R}_+}(x_2) + \Psi_{\mathbb{R}_+}(x_3),$$

we obtain the variational inequality model $VI(\Theta, \Xi, q)$:

$$z \in \mathbb{R}^3:\ \langle \Theta z + q, v - z\rangle + \Xi(v) - \Xi(z) \geq 0,\ \forall v \in \mathbb{R}^3. \tag{24}$$

Let $u : \mathbb{R}_+ \to \mathbb{R}$ be a given supplied voltage (Figs. 19, 20). Using the results proved in [6], we may assert that for each $t \in \mathbb{R}_+$, the rectifier output-signal $V(t)$ is uniquely defined by

$$(\forall t \in \mathbb{R}_+):\ V(t) = R(i_1(t) + i_4(t)) \tag{25}$$

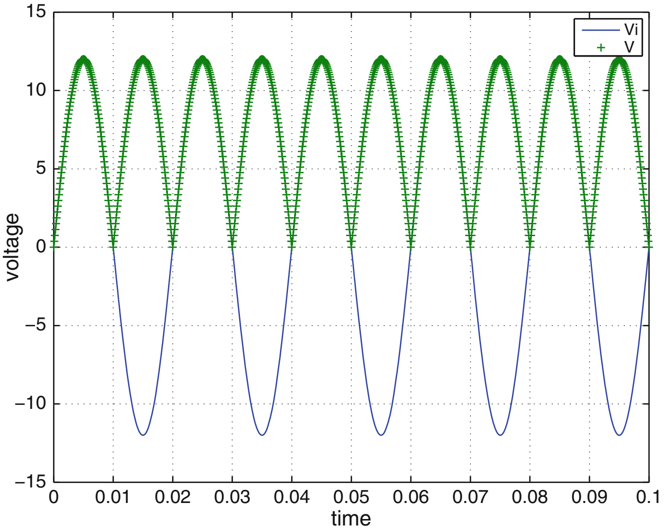


Fig. 19 Rectifier circuit

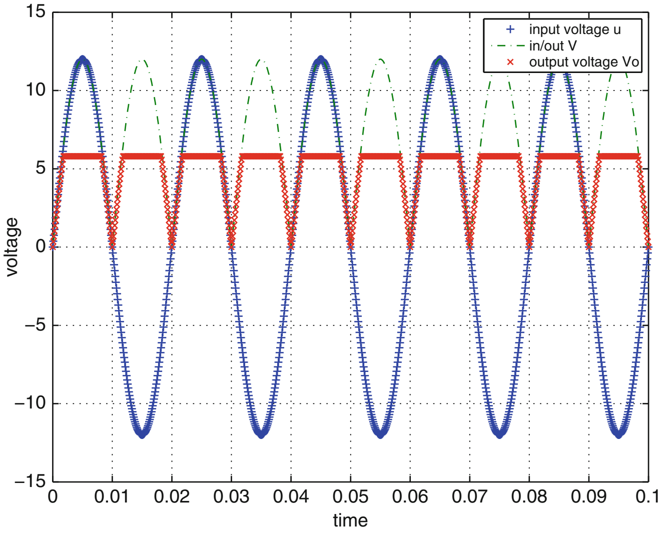


Fig. 20 Stabilizer circuit

where for each $t \in \mathbb{R}_+$, $i_1(t)$ and $i_4(t)$ are computed as solutions of the variational inequality $VI(M, \Phi, Fu(t))$ in (23). Setting

$$(\forall t \in \mathbb{R}_+) : \; q(t) = \begin{pmatrix} V(t) + V_s \\ V(t) + V_s \\ V(t) \end{pmatrix},$$

with $V(t)$ defined in (25), we may also use the results proved in [6], to assert that for each $t \in \mathbb{R}_+$, the stabilizer output-signal $V_o(t)$ is uniquely defined by

$$(\forall t \in \mathbb{R}_+) : \; V_o(t) = \frac{R_2}{K}(I(t) - \alpha_I I'(t)), \tag{26}$$

where $I(t)$ and $I'(t)$ are determined in solving the variational inequality $VI(\Theta, \Xi, q(t))$ in (24).

A General Framework

The practice (see [5] and [15]) shows that a large class of circuits can be studied via the following general mathematical formalism.

Let $\Xi \in \Gamma_0(\mathbb{R}^m; \mathbb{R} \cup \{+\infty\})$ be a given. Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times n}$, and $D \in \mathbb{R}^{n\times p}$ be given matrices. Let $u \in \mathbb{R}^p$ be given, we consider the problem

$NRM(A, B, C, D, u, \Xi)$: Find $(x, y_L) \in \mathbb{R}^n \times \mathbb{R}^m$ such that

$$Ax - By_L + Du = 0, \tag{27}$$

$$y = Cx, \tag{28}$$

and

$$y_L \in \partial\Xi(y). \tag{29}$$

The matrices A, B, C, and D in (27) are structural matrices used to state Kirchhoff's voltage laws and Kirchhoff's current laws in matrix form. The matrix A depends of electrical parameters like resistances, capacitances, and inductances. Usually u is a control vector that drives the system, x denotes a current vector, and y_L is a voltage vector corresponding to electrical devices like diodes whose (possibly set-valued) Ampere-Volt characteristics can be described as in (29).

It is noteworthy that (27)–(29) may represent not only the equations of a static circuit, but also the generalized equation that is to be satisfied by the equilibrium points of a dynamical circuit, or more generally of a class of differential inclusions (see [15] for applications in the absolute stability problem).

Let us now make the following two assumptions:

Assumption $(H1)$*: There exists* $\bar{x}_0 \in \mathbb{R}^n$ *such that* Ξ *is finite and continuous at* $\bar{y}_0 = C\bar{x}_0$.

Assumption $(H2)$*: There exists an invertible matrix* $P \in \mathbb{R}^{n \times n}$ *such that* $PB = C^T$.

We set

$$(\forall x \in \mathbb{R}^n) : \Phi(x) = \Xi(Cx). \tag{30}$$

Then

$$D(\Phi) = \{x \in \mathbb{R}^n : Cx \in D(\Xi)\}. \tag{31}$$

Assumption $(H1)$ entails that $D(\Phi) \neq \emptyset$ and it is clear that $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper convex and lower semicontinuous function. The following result shows that our general framework can be reduced to a variational inequality of the second kind.

Proposition 1 *Suppose that assumptions* $(H1) - (H2)$ *are satisfied and let* Φ *be defined as in (30). i) If* (x, y_L) *is a solution of Problem* $NRM(A, B, C, D, u, \Xi)$ *then* $x \in \mathbb{R}^n$ *is a solution of the variational inequality*

$$\langle -PAx - PDu, v - x\rangle + \Phi(v) - \Phi(x) \geq 0, \forall v \in \mathbb{R}^n. \tag{32}$$

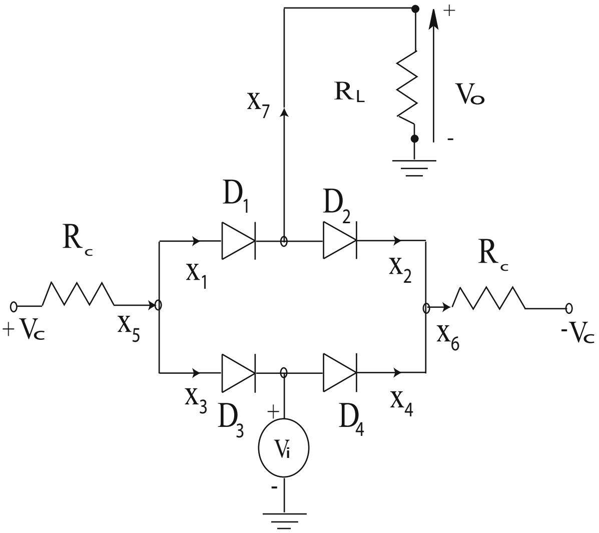


Fig. 21 Four-diode-bridge sampling gate

ii) If $x \in \mathbb{R}^n$ is a solution of the variational inequality (32) then there exists $y_L \in \mathbb{R}^m$ such that (x, y_L) is a solution of Problem $NRM(A, B, C, D, u, \Xi)$.

Indeed, let (x, y_L) be a solution of Problem (27)–(29). Then $0 \in Ax - B\partial\Xi(Cx) + Du$ which is equivalent to $0 \in PAx - PB\partial\Xi(Cx) + PDu$ since P is invertible. Thus $0 \in PAx - C^T\partial\Xi(Cx) + PDu$. The existence of a vector $\bar{y}_0 = C\bar{x}_0$ at which Ξ is finite and continuous ensures that (see, e.g., [55]) $C^T\partial\Xi(Cz) = \partial\Phi(z)$. Thus $0 \in PAx + PDu - \partial\Phi(x)$ and (32) holds. Suppose now that x is solution of Problem ˜(32). We see as above that $0 \in Ax - B\partial\Xi(Cx) + Du$. It results that there exists $y_L \in \partial\Xi(Cx)$ such that $0 = Ax - By_L + Du$. Then we obtain the relations in (27)–(29) by setting $y = Cx$.

A Sampling Gate

The circuit in Fig. 21 is a sampling gate, i.e., a circuit in which the output is a reproduction of the input waveform during a selected time interval and is zero otherwise. The time interval is selected by the gate signal V_c. The circuit involves a bridge of four diodes D_1, D_2, D_3, D_4 and symmetrically controlled by gate voltages $+V_c$ and $-V_c$ through the control resistors $R_c > 0$. The input-signal is given by V_i and the output signal is defined by the voltage V_o through the load resistor $R_L > 0$. Usually, V_i is sinusoidal while V_c is rectangular shaped. We denote by V_j the voltage of the diode D_j and by x_j the current across the diode D_j $(1 \leq j \leq 4)$. Moreover, x_5 denotes the current through the left resistor R_c, x_6 is the current through the right resistor R_c and x_7 denotes the current trough resistor R_L.

Kirchhoff's laws yield the system

$$\overbrace{\begin{pmatrix} -R_L & 0 & 0 \\ 0 & -2R_c & 0 \\ 0 & 0 & 0 \end{pmatrix}}^{A} \overbrace{\begin{pmatrix} x_7 \\ x_6 \\ x_1 \end{pmatrix}}^{x} - \overbrace{\begin{pmatrix} 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}}^{B} \overbrace{\begin{pmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{pmatrix}}^{V} +$$

$$+ \overbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}}^{D} \overbrace{\begin{pmatrix} V_i \\ 2V_c \end{pmatrix}}^{u} = 0$$

and we suppose that the electrical superpotentials of the four diodes D_1, D_2, D_3, D_4 are respectively given by $\varphi_1, \varphi_2, \varphi_3, \varphi_4 \in \Gamma_0(\mathbb{R}; \mathbb{R} \cup \{\infty\})$

$$\begin{aligned} V_1 &\in \partial\varphi_1(x_1), \\ V_2 &\in \partial\varphi_2(x_2) = \partial\varphi_2(x_1 - x_7), \\ V_3 &\in \partial\varphi_3(x_3) = \partial\varphi_3(x_6 - x_1), \\ V_4 &\in \partial\varphi_4(x_4) = \partial\varphi_4(x_7 + x_6 - x_1). \end{aligned}$$

Setting

$$y = \overbrace{\begin{pmatrix} 0 & 0 & 1 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & 1 & -1 \end{pmatrix}}^{C} \begin{pmatrix} x_7 \\ x_6 \\ x_1 \end{pmatrix}$$

and defining the function $\Xi(x) = \varphi_1(x_1) + \varphi_2(x_2) + \varphi_3(x_3) + \varphi_4(x_4)$, we may write $V \in \partial\Xi(y)$ and then consider problem $NRM(A, B, C, D, u, \Xi)$. It is easy to see that in practice, $\begin{pmatrix} 2 & 1 & 1 & 2 \end{pmatrix}^T = C \begin{pmatrix} 1 & 3 & 2 \end{pmatrix}^T \in]0, +\infty[^4$ is a point at which Ξ is finite and continuous since electrical superpotentials φ_i $(1 \leq i \leq 4)$ of any type of diode are finite and continuous on $]0, +\infty[$. It results that Assumption $(H1)$ holds (Fig. 22). We remark also that $C^T = B$ and thus Assumption $(H2)$ holds with $P = I$. So, for a driven time depending input $t \mapsto V_i(t)$ and control gate signals $t \mapsto V_c(t)$ and $t \mapsto -V_c(t)$, the output time depending voltage $t \mapsto V_o(t)$ through the resistor R_L is determined by $V_o(t) = R_L x_7(t)$ where the current function $t \mapsto x_7(t)$ is determined in solving the variational inequality (32).

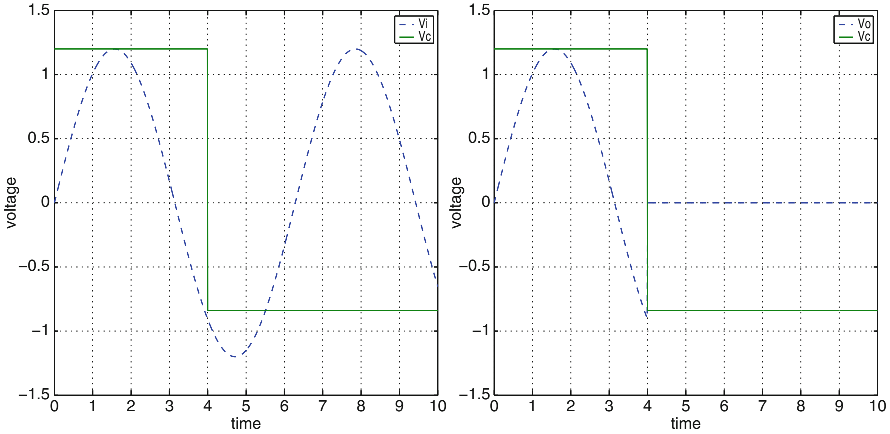


Fig. 22 Four-diode-bridge sampling gate with ideal diodes

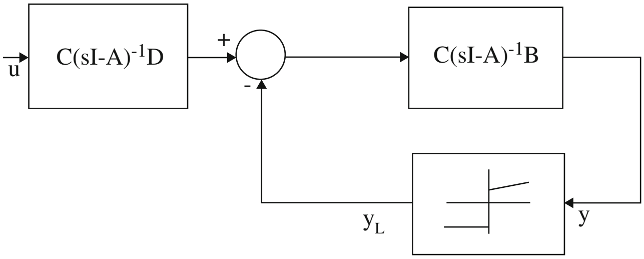


Fig. 23 Illustration of the circuit with a feedback branch

Non-regular Dynamical Systems

In this section, we introduce a general formalism whose study has been initiated in [14]. We refer also the reader to [5, 16] and [35] for some related recent works. Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{m \times n}$, and $D \in \mathbb{R}^{n \times p}$ be given matrices. Let $\Xi : \mathbb{R}^m \to \mathbb{R}$ be a given mapping. It is assumed that $\Xi \in \Gamma_0(\mathbb{R}^m; \mathbb{R} \cup \{+\infty\})$. Our aim is to introduce a system described by a transfer function

$$H(s) = C(sI - A)^{-1}B$$

and a feedback branch containing a sector static nonlinearity as depicted in Fig. 23. The feedback nonlinearity that describes the graph (y, y_L) is here defined by the model

$$y_L \in \partial \Xi(y).$$

Moreover, the system is driven by inputs Du for some given function

$$u : [0, +\infty[\to \mathbb{R}^p; t \mapsto u(t).$$

The state-space equations of such a system are given by

$$\frac{dx}{dt}(t) = Ax(t) - By_L(t) + Du(t), \tag{33}$$

$$y(t) = Cx(t), \tag{34}$$

and

$$y_L(t) \in \partial \Xi(y(t)). \tag{35}$$

Note that if $(\forall t \geq 0) : u(t) = u$ for some given $u \in \mathbb{R}^p$, then the stationary solutions of (33)–(35) are given by the solutions of the problem $NRM(A, B, C, D, u, \Xi)$ discussed in the previous section.

We suppose that $u \in L^1_{loc}(0, +\infty; \mathbb{R}^p)$ and for $x_0 \in \mathbb{R}^n$, we consider the problem $P(x_0)$: Find a function $x : [0, +\infty[\to \mathbb{R}^n;\ t \mapsto x(t)$ and a function $y_L : [0, +\infty[\to \mathbb{R}^m;\ t \mapsto y_L(t)$ such that

$$x \in C^0([0, +\infty[; \mathbb{R}^n), \tag{36}$$

$$By_L \in L^1_{loc}(0, +\infty; \mathbb{R}^n), \tag{37}$$

$$\frac{dx}{dt} \in L^1_{loc}(0, +\infty; \mathbb{R}^n), \tag{38}$$

$$x(0) = x_0, \tag{39}$$

$$\frac{dx}{dt}(t) = Ax(t) - By_L(t) + Du(t),\ \text{a.e. } t \geq 0, \tag{40}$$

$$y(t) = Cx(t),\ \forall\ t \geq 0, \tag{41}$$

and

$$y_L(t) \in \partial \Xi(y(t)),\ \text{a.e. } t \geq 0. \tag{42}$$

Let us now make the following two assumptions:

Assumption $(G1)$*: There exists a symmetric and invertible matrix $R \in \mathbb{R}^{n \times n}$ such that*

$$R^{-2}C^T = B.$$

Assumption $(G2)$*: There exists* $z_0 \in \mathbb{R}^n$ *such that* Ξ *is finite and continuous at* $y_0 = CR^{-1}z_0$.

Note that $R^{-2} = (R^{-1})^2$. Using (40)–(42), we may consider the differential inclusion

$$\frac{dx}{dt} \in Ax - B\partial\Xi(Cx) + Du.$$

Setting $z = Rx$, we remark that

$$\begin{aligned} &\frac{dx}{dt} \in Ax - B\partial\Xi(Cx) + Du \\ \Leftrightarrow\ & R\frac{dx}{dt} \in RAR^{-1}Rx - RB\partial\Xi(CR^{-1}Rx) + RDu \\ \Leftrightarrow\ & \frac{dz}{dt} \in RAR^{-1}z - R^{-1}R^2B\partial\Xi(CR^{-1}z) + RDu \\ \Leftrightarrow\ & \frac{dz}{dt} \in RAR^{-1}z - R^{-1}C^T\partial\Xi(CR^{-1}z) + RDu. \end{aligned}$$

We set

$$(\forall z \in \mathbb{R}^n) : \Phi(x) = \Xi(CR^{-1}z).$$

Then

$$D(\Phi) = \{z \in \mathbb{R}^n : CR^{-1}z \in D(\Xi)\}.$$

and with Assumption $(G1)$, we have

$$(\forall z \in \mathbb{R}^n) : \ \partial\Phi(z) = R^{-1}C^T\partial\Xi(CR^{-1}z).$$

This allows us to consider, for $x_0 \in \mathbb{R}^n$, the problem $Q(x_0)$: Find a function $z : [0, +\infty[\to \mathbb{R}^n;\ t \mapsto z(t)$ such that

$$z \in C^0([0, +\infty[; \mathbb{R}^n), \tag{43}$$

$$\frac{dz}{dt} \in L^1_{loc}(0, +\infty; \mathbb{R}^n), \tag{44}$$

$$z(0) = Rx_0, \tag{45}$$

$$\frac{dz}{dt}(t) \in RAR^{-1}z(t) + RDu(t) - \partial\Phi(z(t)),\ \text{a.e. } t \geq 0. \tag{46}$$

Note that this last differential inclusion is equivalent to the variational inequality

$$\langle \frac{dz}{dt}(t) - RAR^{-1}z(t) - RDu(t), v - z(t)\rangle + \Phi(v) - \Phi(z(t)) \geq 0, \forall v \in \mathbb{R}^n, \quad \text{a.e. } t \geq 0.$$

Proposition 2 *Suppose that assumptions* $(G1) - (G2)$ *are satisfied. If* (x, y_L) *is solution of Problem* $P(x_0)$*, then* $z = Rx$ *is solution of Problem* $Q(x_0)$*. Reciprocally, if* z *is solution of Problem* $Q(x_0)$*, then there exists a function* y_L *such that* $(R^{-1}z, y_L)$ *is solution of Problem* $P(x_0)$*.*

Indeed, we have seen above that if (x, y_L) is solution of Problem $P(x_0)$, then $z = Rx$ is solution of Problem $Q(x_0)$. Suppose now that z is solution of Problem $Q(x_0)$. Then setting $x = R^{-1}z$, we see as above that

$$\frac{dx}{dt} \in Ax - B\partial\Xi(Cx) + Du.$$

It results that there exists a function $y_L \in \partial\Xi(Cx)$ such that

$$\frac{dx}{dt} = Ax - By_L + Du.$$

Note that

$$By_L = -\frac{dx}{dt} + Ax + Du \in L^1_{loc}(0, +\infty; \mathbb{R}^n).$$

Then we obtain the relations in (36)–(42) by setting

$$y = Cx.$$

So, using assumptions $(G1) - (G2)$, we may reduce the study of problem $P(x_0)$ to the one of problem $Q(x_0)$ which can be investigated by means of mathematical tools from set-valued analysis, theory of maximal monotone operators, and variational inequality theory (see, e.g., [9, 12, 13, 19, 29, 31, 37, 38, 45, 52, 55]). The equivalence between complementarity systems, projected systems, and unilateral differential inclusions are recapitulated in [20]. General results allowing a stability analysis of the stationary solutions of non-regular dynamical systems can be found in [18] and [35]. A generalization of Krakovskii-LaSalle invariance theory for non-regular systems can be found in [16] and related results in [17]. The stability analysis applicable to the study of a DC-DC Buck converter is detailed in [8]. Piecewise affine dynamical systems and linear complementarity systems with applications in electronics are given in [11, 14, 21–27]. Numerical methods have been proposed in [4] and [3] so as to study switched circuits. The nonsmooth approach applied to simulating integrated circuits and power electronics is detailed in [30]. We refer also the readers to [1] for a book on numerical methods for nonsmooth dynamical systems with applications in electronics. Let us here also mention a study including mathematical formulation and numerical simulations of higher order Moreau's sweeping process in electronics [2].

Assumption $(G1)$ *and Kalman–Yakubovich–Popov Lemma*

Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{m\times n}$. One says that the representation (A, B, C) is minimal provided that (A, B) is controllable and (A, C) is observable, i.e., the matrices $(B\ AB\ A^2B\ \ldots\ A^{n-1}B)$ and $(C\ CA\ CA^2\ \ldots\ CA^{n-1})^T$ have full rank. Let us now consider the real, rational matrix-valued transfer function $H : \mathbb{C} \to \mathbb{C}^{m\times m}$ given by

$$H(s) = C(sI_n - A)^{-1}B. \tag{47}$$

Definition 1 One says that H is positive real if

- H is analytic in $\mathbb{C}^+ := \{s \in \mathbb{R} : Re[s] > 0\}$,
- $H(s) + H^T(\bar{s})$ is positive semi-definite for all $s \in \mathbb{C}^+$,

where $\bar{s}$ is the conjugate of s.

The following result is called Kalman–Yakubovich–Popov lemma [46, 57] and [61] (see also, e.g., [19, 58]).

Lemma 1 *Let (A, B, C) be a minimal realization and let H be defined in (47). The transfer function matrix H is positive real if and only if there exist a symmetric and positive definite matrix $P \in \mathbb{R}^{n\times n}$ and a matrix $L \in \mathbb{R}^{n\times m}$, such that*

$$\begin{aligned} PA + A^TP &= -LL^T \\ PB &= C^T. \end{aligned} \tag{48}$$

So, if the realization (A, B, C) is minimal and the transfer function H is positive real then there exists a symmetric and positive definite matrix $P \in \mathbb{R}^{n\times n}$ and a matrix $L \in \mathbb{R}^{n\times m}$ such that $PA + A^TP = -LL^T$ and $PB = C^T$. Choosing R as the symmetric square root of P, i.e., $R = R^T$, R positive definite and $R^2 = P$, we see that $B^TR^2 = C$ and thus

$$R^{-2}C^T = B.$$

It results that assumption $(G1)$ holds.

A Non-regular Circuit

Let us consider the following dynamics that corresponds to the circuit depicted in Fig. 24:

$$\begin{pmatrix} \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \\ \frac{dx_3}{dt} \end{pmatrix} = \overbrace{\begin{pmatrix} 0 & 1 & 0 \\ -\frac{1}{L_3 C_4} & -\frac{(R_1+R_3)}{L_3} & \frac{R_1}{L_3} \\ 0 & \frac{R_1}{L_2} & -\frac{(R_1+R_2)}{L_2} \end{pmatrix}}^{A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$-\overbrace{\begin{pmatrix} 0 & 0 \\ \frac{1}{L_3} & \frac{1}{L_3} \\ -\frac{1}{L_2} & 0 \end{pmatrix}}^{B} \begin{pmatrix} y_{L,1} \\ y_{L,2} \end{pmatrix} + \overbrace{\begin{pmatrix} 0 \\ 0 \\ \frac{1}{L_2} \end{pmatrix}}^{D} u,$$

and

$$\begin{cases} y_{L,1} \in \partial\varphi_D(-x_3 + x_2) \\ y_{L,2} \in \partial\varphi_Z(x_2) \end{cases} \tag{49}$$

where $R_1 > 0, R_2 > 0, R_3 > 0$ are resistors, $L_2 > 0$, $L_3 > 0$ are inductors, $C_4 > 0$ is a capacitor, x_1 is the time integral of the current across the capacitor, x_2 is the current across the capacitor, x_3 is the current across the inductor L_2 and resistor R_2, $y_{L,1}$ is the voltage of the Zener diode, $y_{L,2}$ is the voltage of the diode, φ_Z is the electrical superpotential of the Zener diode, and φ_D is the electrical superpotential of the diode. Setting

$$y = \overbrace{\begin{pmatrix} 0 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix}}^{C} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

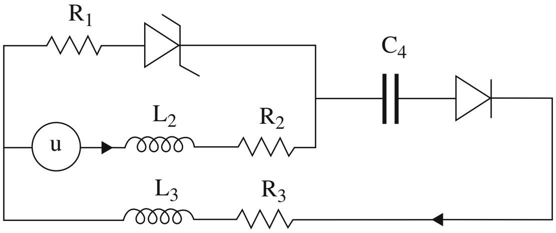


Fig. 24 Non-regular circuit

and defining the function $\Xi : \mathbb{R}^2 \to \mathbb{R}; X \mapsto \Xi(X)$ by the formula

$$\Xi(X) = \varphi_D(X_1) + \varphi_Z(X_2),$$

we may write the relations in (49) equivalently as

$$y_L \in \partial \Xi(Cx).$$

It is easy to see that

$$\text{rank}\{(B\ AB\ A^2B)\} \ = \ \text{rank}\{(C\ CA\ CA^2)^T\} = 3$$

and a simple computation shows that the transfer function

$$H(s) = C(sI - A)^{-1}B =$$

$$\frac{1}{D(s)} \begin{pmatrix} s^2C_4L_3 + s^2C_4L_2 + sC_4R_2 + sC_4R_3 + 1 & \frac{C_4L_3}{L_2}s(sL_2 + R_2) \\ C_4s(sL_2 + R_2) & \frac{C_4L_3}{L_2}s(sL_2 + R_1 + R_2) \end{pmatrix},$$

where

$$D(s) = s^3C_4L_3L_2 + s^2C_4L_3R1 + s^2C_4L_3R_2 + s^2C_4R_1L_2 + sC_4R_1R_2 + s^2C_4R_3L_2 + \\ sC_4R_3R_1 + sC_4R_3R_2 + sL_2 + R_1 + R_2,$$

is positive real. The existence of a matrix R that satisfies condition $(G1)$ is thus here also a consequence of the Kalman–Yakubovich–Popov lemma. A simple computation shows that the matrix

$$R = \begin{pmatrix} \frac{1}{\sqrt{C_4}} & 0 & 0 \\ 0 & \sqrt{L_3} & 0 \\ 0 & 0 & \sqrt{L_2} \end{pmatrix}$$

is convenient.

References

1. V. Acary, B. Brogliato, Numerical methods for nonsmooth dynamical systems, in *Applications in Mechanics and Electronics*. Lecture Notes in Applied and Computational Mechanics, vol. 35 (Springer, Berlin, 2008)
2. V. Acary, B. Brogliato, D. Goeleven, Higher order Moreau's sweeping process. Mathematical formulation and numerical simulation. Math. Program. A **113**(1), 133–217 (2008)

3. V. Acary, O. Bonnefon, B. Brogliato, Time-stepping numerical simulation of switched circuits with the nonsmooth dynamical systems approach. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **29**(7), 1042–1055 (2010)
4. V. Acary, O. Bonnefon, B. Brogliato, *Nonsmooth Modeling and Simulation for Switched Circuits*. Lecture Notes in Electrical Engineering, vol. 69 (Springer, Dordrecht, 2011)
5. K. Addi, S. Adly, B. Brogliato, D. Goeleven, A method using the approach of Moreau and Panagiotopoulos for the mathematical formulation of non-regular circuits in electronics. Nonlinear Anal. Ser. C Hybrid Syst. Appl. **1**, 30–43 (2007)
6. K. Addi, B. Brogliato, D. Goeleven, A qualitative mathematical analysis of a class of linear variational inequalities via semi-complementarity problems. Appl. Electron. Math. Program. A **126**(1), 31–67 (2011)
7. K. Addi, Z. Despotovic, D. Goeleven, A. Rodic, Modelling and analysis of a non-regular electronic circuit via a variational inequality formulation. Appl. Math. Model. **35**, 2172–2184 (2011)
8. S. Adly, D. Goeleven, B.K. Le, Stability analysis and attractivity results of a DC-DC Buck converter. Set Valued Var. Anal. **20**(3), 331–353 (2012)
9. J.P. Aubin, A. Cellina, *Differential Inclusions* (Springer, Berlin, 1984)
10. U.A. Bakshi, A.P. Godse, *Electronic Devices and Circuits* (Technical Publications, Pune, 2008)
11. A. Bemporad, M.K. Camlibel, W.P.M.H. Heemels, A.J. van der Schaft, J.M. Schumacher, B. de Schutter, Further switched systems, in *Handbook of Hybrid Systems Control*, ed. by F. Lamnabhi-Lagarrigue, J. Lunze, Theory, Tools, Applications. Cambridge University Press, Cambridge (2009)
12. H. Brézis, *Opérateurs Maximaux Monotones et Semigroupes de Contractions dans les Espaces de Hilbert* (North-Holland/American Elsevier, Amsterdam/New York, 1972)
13. H. Brézis, Problèmes unilatéraux. Journal de Mathématiques Pures et Appliquées **51**, 1–168 (1972)
14. B. Brogliato, Some perspectives on the analysis and control of complementarity systems. IEEE Trans. Autom. Control **48**, 918–935 (2003)
15. B. Brogliato, Absolute stability and the Lagrange-Dirichlet theorem with monotone multivalued mappings. Syst. Control Lett. **51**, 343–353 (2004)
16. B. Brogliato, D. Goeleven, The Krakovskii-LaSalle invariance principle for a class of unilateral dynamical systems. Math. Control Signals Syst. **17**, 57–76 (2005)
17. B. Brogliato, D. Goeleven, Well-posedness, stability and invariance results for a class of multivalued Lur'e dynamical systems. Nonlinear Anal. Theory Methods Appl. **74**, 195–212 (2011)
18. B. Brogliato, D. Goeleven, Existence, uniqueness of solutions and stability of nonsmooth multivalued Lur'e dynamical systems. J. Convex Anal. **20**, 881–900 (2013)
19. B. Brogliato, R. Lozano, B.M. Maschke, O. Egeland, *Dissipative Systems Analysis and Control*. Springer CCE Series (Springer, London, 2006)
20. B. Brogliato, A. Daniilidis, C. Lemarchal, C. Acary, On the equivalence between complementarity systems, projected systems and unilateral differential inclusions. Syst. Control Lett. **55**, 45–51 (2006)
21. M.K. Camlibel, J.M. Schumacher, Existence and uniqueness of solutions for a class of piecewise linear systems. Linear Algebra Appl. 147–184 (2002)
22. M.K. Camlibel, J.M. Schumacher, Existence and uniqueness of solutions for a class of piecewise linear dynamical systems. Linear Algebra Appl. **351–352**, 147–184 (2002)
23. M.K. Camlibel, J.M. Schumacher, Linear passive systems and maximal monotone mappings. Math. Program. (2015, accepted for publication)
24. M.K. Camlibel, W.P.M.H. Heemels, J.M. Schumacher, Consistency of a time-stepping method for a class of piecewise-linear networks. IEEE Trans. Circuits Syst. I. **49**, 349–357 (2002)
25. M.K. Camlibel, W.P.M.H. Heemels, H. Schumacher, On linear passive complementarity systems. Eur. J. Control. **8**, 220–237 (2002)
26. M.K. Camlibel, W.P.M.H. Heemels, J.M. Schumacher, On linear passive complementarity systems. Eur. J. Control **8**(3), 220–237 (2002)

27. M.K. Camlibel, L. Iannelli, F. Vasca, Passivity and complementarity. Math. Program. **145**(1–2), 531–563 (2014)
28. R.W. Cottle, J.S. Pang, R.E. Stone, *The Linear Complementarity Problem* (Academic, New York, 1992)
29. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
30. P. Denoyelle, V. Acary, The non-smooth approach applied to simulating integrated circuits and power electronics. Evolution of electronic circuit simulators towards fast-SPICE performance. INRIA Technical Report RT-0321 (2006)
31. G. Duvaut, J.L. Lions, *Les Inéquations en Mécanique et en Physique* (Dunod, Paris, 1972)
32. F. Facchinei, J.-S. Pang, *Finite Dimensional Variational Inequalities and Complementarity Problems* (Springer, Berlin, 2003)
33. C. Georgescu, B. Brogliato, V. Acary, Switching, relay and complementarity systems: a tutorial on their well-posedness and relationships. Phys. D. Dyn. Bifurcations Nonsmooth Syst. **241**(22), 1985–2002 (2012)
34. G. Goeleven, An existence and uniqueness result for a linear mixed variational inequality arising in electrical circuits with transistors. J. Optim. Theory Appl. **138**, 397–406 (2008)
35. D. Goeleven, B. Brogliato, Stability and instability matrices for linear evolution variational inequalities. IEEE Trans. Autom. Control **49**, 521–534 (2004)
36. D. Goeleven, D. Motreanu, On the solvability of variational inequalities via relaxed complementarity problems. Commun. Appl. Anal. **4**, 533–546 (2000)
37. D. Goeleven, D. Motreanu, *Variational and Hemivariational Inequalities – Theory, Methods and Applications*. Vol. 2: Unilateral Problems and Unilateral Mechanics (Kluwer Academic, Boston, 2003)
38. D. Goeleven, D. Motreanu, Y. Dumont, M. Rochdi, *Variational and Hemivariational Inequalities – Theory, Methods and Applications*. Vol. 1: Unilateral Analysis and Unilateral Mechanics (Kluwer Academic, Boston, 2003)
39. W.P.M.H. Heemels, J.M. Schumacher, S. Weiland, Linear complementarity systems. SIAM J. Appl. Math. **60**, 1234–1269 (2000)
40. W.P.M.H. Heemels, M.K. Camlibel, J.M. Schumacher, On the dynamic analysis of piecewise-linear networks. IEEE Trans. Circuits Syst. I **49**, 315–327 (2002)
41. W.P.M.H. Heemels, M.K. Camlibel, A.J. Van der Schaft, J.M. Schumacher, Well-posedness of hybrid systems, in *Control Systems, Robotics and Automation*, ed. by H. Unbehauen. Theme 6.43 of Encyclopedia of Life Support Systems (Eolss, Oxford, 2004)
42. W.P.M.H. Heemels, M.K. Camlibel, J.M. Schumacher, B. Brogliato, Observer-based control of linear complementarity systems. Int. J. Robust Nonlinear Control **21**(10), 1193–1218 (2011)
43. J.B. Hiriart-Urruty, C. Lemaréchal, *Fundamentals of Convex Analysis*. Springer Grundlehren Text Editions (Springer, Heidelberg, 2001)
44. G. Isac, *Complementarity Problems*. Lecture Notes in Mathematics, vol. 1528 (Springer, Berlin, 1992)
45. A. Jofré, R.T. Rockafellar, R.J.-B. Wets, Variational inequalities and economic equilibrium. Math. Oper. Res. **32**, 32–50 (2007)
46. R.E. Kalman, Lyapunov functions for the problem of Lur'e in automatic control. Proc. Natl. Acad. Sci. **49**, 201–205 (1963)
47. I.V. Konnov, E.O. Volotskaya, Mixed variational inequalities and economic equilibrium problems. J. Appl. Math. **6**, 289–314 (2002)
48. D. Leenaerts, W.M.G. van Bokhoven, *Piecewise Linear Modeling and Analysis* (Kluwer Academic, Norwell, MA, 1998)
49. J. Millman, C.C. Halkias, *Integrated Electronics* (McGraw-Hill Kogakusha, Sydney, 1985)
50. Y. Murakami, A Method for the Formulation and solution of circuits composed of switches and linear RLC networks. IEEE Trans. Circuits Syst. I **49**, 315–327 (2002)
51. J.J. Moreau, La Notion du Surpotentiel et les Liaisons Unilatérales on Elastostatique. C.R. Acad. Sci. Paris **167A**, 954–957 (1968)
52. J.J. Moreau, Nonsmooth mechanics and applications, in *CISM*, ed. by P.D. Panagiotopoulos, vol. 302 (Springer, Wien, 1988)

53. K.G. Murty, *Linear Complementarity*. Linear and Nonlinear Programming (Hederman, Berlin, 1998)
54. P.D. Panagiotopoulos, Non-convex superpotentials in the sense of F.H. Clarke and applications. Mech. Res. Comm. **8**, 335–340 (1981)
55. P.D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*. Convex and Nonconvex Energy Functions (Birkhaüser, Basel, 1985)
56. D. Pascali, S. Sburlan, *Nonlinear Mappings of Monotone Type* (Sijthoff and Noordhoff International Publishers, Amsterdam, 1978)
57. V.M. Popov, Absolute stability of nonlinear systems of automatic. Control Autom. Remote Control **22**, 857–875 (1962)
58. A. Rantzer, On the Kalman-Yakubovich-Popov Lemma. Syst. Control Lett. **28**, 7–10 (1996)
59. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
60. M. Vidyasagar, *Nonlinear Systems Analysis*. Prentice Hall International Editions (Prentice Hall, Princeton, NJ, 1993)
61. V.A. Yakubovich, Solution of certain matrix inequalities in the stability theory of nonlinear control systems. Dokl. Akad. Nauk. SSSR. **143**, 1304–1307 (1962)

Electromagnetic Scattering by a Chiral Impedance Screen

C.E Athanasiadis, V. Sevroglou, and K.I. Skourogiannis

Abstract In this paper the solvability of the direct electromagnetic scattering problem by an impedance screen in a chiral environment is presented. Time-harmonic electromagnetic plane waves in a chiral medium are considered as incident fields. These propagating fields are scattered by an obstacle which is a partially coated open surface Γ, well known as the “screen". Uniqueness results are proved using appropriate relations for Beltrami fields, and in addition, existence results are established by using a variational method in suitable functional space setting.

Keywords Chiral media • Beltrami fields • Impedance boundary conditions

A. M. S. Mathematics Subject Classifications: 35P25, 35Q60, 78A40

Introduction

In this work the scattering problem of plane time-harmonic electromagnetic waves by a partially coated chiral obstacle embedded in an infinite homogeneous isotropic chiral medium is studied. From the mathematical point of view, chiral media satisfy a set of constitutive relations in which the magnetic and electric fields are coupled. Different expressions exist for the constitutive relations [14]; in this work the well-known Drude-Born-Fedorov (DBF) constitutive relations are used. These constitutive relations are chosen because they are symmetric under time

C.E. Athanasiadis (✉) • K.I. Skourogiannis
Department of Mathematics, National and Kapodistrian University of Athens, Panepistimiopolis, GR 15784 Zographou, Athens, Greece
e-mail: cathan@math.uoa.gr; skouroco@otenet.gr

V. Sevroglou
Department of Statistics and Insurance Science, University of Piraeus, 80 Karaoli and dimitriou Str., Piraeus 18534, Greece
e-mail: bsevro@unipi.gr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_2

reversality and duality transformations. Chiral obstacles are characterized by the so-called chirality (or preferential handedness) and the related electromagnetic fields are composed of left circularly polarized (LCP) and right circularly polarized (RCP) components. These fields have independent directions of propagation and different wave numbers. Chirality is common in a variety of naturally occurring and man-made objects (e.g. DNA in molecular scale, helices) and has also played an important role to the study of optical activity. Properties and scattering problems involving chiral media have been studied by many scientists; for an excellent source we refer to [15, 16] and [17] (and therein references). Solvability results concerning direct scattering problems where the obstacle is a perfect conductor or a dielectric (penetrable scatterer) in chiral media can be found in [5, 6]. In these cases, Bohren decomposition is used and an equivalent boundary integral formulation to the scattering problems is considered. Furthermore, boundary integral equations for electromagnetic scattering by a homogeneous chiral obstacle were studied in [4], by using a generalization of Müller's equations for scattering by a non- chiral obstacle. In [1], existence and uniqueness of the solution to the diffraction problem of a plane electromagnetic field by a chiral curved layer covering a perfectly conducting object have been studied. In particular, approximative impedance conditions are given for thin chiral curved layers and optimal error estimates are obtained (the reader can also see [2]). We end up with the work studied in [3], where the LCP and the RCP Beltrami Herglotz functions were defined by an integral representation over the unit sphere where the corresponding kernels are exactly the Beltrami far-field patterns. These functions will play an important role for the investigation of the inverse electromagnetic problem for a mixed-impedance screen in chiral media. For non-chiral media, mixed boundary value problems which describe model of scattering by obstacles that are covered by a thin layer of material on part of their boundaries are studied in [10]. The direct and inverse scattering problem of a time-harmonic electromagnetic plane wave by a mixed perfectly conducting-impedance screen is studied in [8, 11] and [9]. Further, we mention that problems with mixed-impedance boundary conditions in elasticity have been considered in [7].

Setting Up the Problem

We consider a plane time-harmonic electromagnetic wave $\mathbf{E}^{\mathrm{inc}}$ which is propagated in an infinite homogeneous isotropic chiral medium. This field is disturbed by a very thin partially coated chiral obstacle (the scatterer), known as *screen*, which is an open, bounded, smooth surface $\Gamma \in \mathbb{R}^3$ with two sides coated by impedance material. This surface is also a part of a piecewise smooth surface $\partial \mathrm{D}$ of a bounded domain $\mathrm{D} \subset \mathbb{R}^3$. The domain D as well as the infinite medium is filled up with a homogeneous and isotropic chiral medium of *chirality measure* β. For our case we assume that β is a positive constant. We denote $\hat{n}$ the unit normal vector to Γ which coincides with the outward normal vector defined almost everywhere on $\partial \mathrm{D}$. The boundary condition on each side of this surface obstacle is described by an impedance boundary condition.

For a vector $\mathbf{u}$ we use the notation $\hat{\mathbf{n}} \times \mathbf{u}^{+}|_{\Gamma}, \hat{\mathbf{n}} \cdot \mathbf{u}^{+}|_{\Gamma}$, $\gamma_T^{+}\mathbf{u}|_{\Gamma}$ for the restriction to Γ of the traces $\hat{\mathbf{n}} \times \mathbf{u}^{+}|_{\partial D}, \hat{\mathbf{n}} \cdot \mathbf{u}^{+}|_{\partial D}$ and $\gamma_T^{+}\mathbf{u}|_{\partial D}$, respectively, from the outside of the ∂D, where $\gamma_T^{+}\mathbf{u} := \hat{\mathbf{n}} \times (\mathbf{u}^{+} \times \hat{\mathbf{n}})$ is the tangential component of $\mathbf{u}^{+}$. Similar considerations for the traces from the inside of the ∂D which are notated by $\hat{\mathbf{n}} \times \mathbf{u}^{-}|_{\Gamma}, \hat{\mathbf{n}} \cdot \mathbf{u}^{-}|_{\Gamma}$, $\gamma_T^{-}\mathbf{u}|_{\Gamma}$ also hold. We also use the notation $\mathbf{u}^{\pm}|_{\Gamma}$ when a relation is hold for both the restrictions of the vector $\mathbf{u}$ on Γ.

The total electric field $\mathbf{E}$ is the superposition of the incident electric field $\mathbf{E}^{\mathrm{inc}}$ and the scattered electric field $\mathbf{E}^{\mathrm{sc}}$, i.e.,

$$\mathbf{E} = \mathbf{E}^{\mathrm{inc}} + \mathbf{E}^{\mathrm{sc}}. \tag{1}$$

The scattering electromagnetic problem by a double impedance screen in chiral media is to determine the total electric field $\mathbf{E}$ that satisfies

$$\nabla \times \nabla \times \mathbf{E} = 2\gamma^2 \beta \,\nabla \times \mathbf{E} + \gamma^2 \mathbf{E} \quad \text{in } \mathbb{R}^3 \setminus \overline{\Gamma}, \tag{2}$$

$$\hat{\mathbf{n}} \times \nabla \times \mathbf{E}^{-} = \frac{i\lambda^{-}\gamma^2}{k^2}\, \hat{\mathbf{n}} \times \mathbf{E}^{-} \times \hat{\mathbf{n}} + \gamma^2 \beta\, \hat{\mathbf{n}} \times \mathbf{E}^{-} \quad \text{on } \Gamma, \tag{3}$$

$$\hat{\mathbf{n}} \times \nabla \times \mathbf{E}^{+} = \frac{i\lambda^{+}\gamma^2}{k^2}\, \hat{\mathbf{n}} \times \mathbf{E}^{+} \times \hat{\mathbf{n}} + \gamma^2 \beta\, \hat{\mathbf{n}} \times \mathbf{E}^{+} \quad \text{on } \Gamma, \tag{4}$$

$$\hat{\mathbf{r}} \times \nabla \times \mathbf{E}^{sc} - \beta\gamma^2\, \hat{\mathbf{r}} \times \mathbf{E}^{sc} + \frac{i\gamma^2}{k} \mathbf{E}^{sc} = o(\frac{1}{r}) \quad r \to \infty, \tag{5}$$

where $\gamma^2 = k^2/(1-k^2\beta^2)$, $k = \omega\sqrt{\varepsilon\mu}$, with ω the angular frequency, ε, μ been the electric permittivity and magnetic permeability, respectively, and λ^{-}, $\lambda^{+} \in L_{\infty}(\Gamma)$ with $\lambda^{-}, \lambda^{+} \geq \lambda_0 > 0$. The Silver-Müller radiation condition (5) holds uniformly in all directions $\hat{\mathbf{r}} = \mathbf{r}/r$ where $r := |\mathbf{r}|$. We note that the electric field $\mathbf{E}$ is divergence-free, that is $\nabla \cdot \mathbf{E} = 0$. In addition, k is not a wave number and its notation has not any particular physical significance.

In what follows we deal with the uniqueness and existence of the solution of the scattering problem (2)–(5) in an appropriate space setting . Hence, we define the following Sobolev spaces:

$$H(\mathrm{curl}, B_\rho \setminus \overline{\Gamma}) := \{\mathbf{u} \in \left[L^2(B_\rho \setminus \overline{\Gamma})\right]^3 : \mathrm{curl}\mathbf{u} \in \left[L^2(B_\rho \setminus \overline{\Gamma})\right]^3\}, \tag{6}$$

$$L_t^2(\Gamma) := \{\mathbf{u} \in [L^2(\Gamma)]^3 : \mathbf{u} \cdot \hat{\mathbf{n}} = 0, \text{ on } \Gamma\}, \tag{7}$$

$$H_{loc}(\mathrm{curl}, \mathbb{R}^3 \setminus \overline{\Gamma}) := \left\{\mathbf{u} \in H(\mathrm{curl}, B_\rho \setminus \overline{\Gamma}) \text{ for every } B_\rho \text{ such that } D \subset B_\rho\right\}, \tag{8}$$

and

$$X(\mathrm{curl},\, \mathbb{R}^3\backslash\overline{\Gamma}) := \left\{\mathbf{u} \in H_{loc}(\mathrm{curl},\, \mathbb{R}^3 \setminus \overline{\Gamma}) : \hat{\mathbf{n}} \times \mathbf{u}^-|_\Gamma \,, \hat{\mathbf{n}} \times \mathbf{u}^+|_\Gamma \in L^2_t(\Gamma)\right\}, \tag{9}$$

where B_ρ is a sphere with radius ρ large enough, containing the bounded domain D. The last space is equipped with the graph norm

$$\begin{aligned} \| \mathbf{u} \|^2_{X(\mathrm{curl},\mathbb{R}^3\backslash\overline{\Gamma})} &:= \| \nabla \times \mathbf{u} \|^2_{\left(L^2(B_\rho\backslash\overline{\Gamma})\right)^3} + \| \mathbf{u} \|^2_{\left(L^2(B_\rho\backslash\overline{\Gamma})\right)^3} \\ &\quad + \| \hat{\mathbf{n}} \times \mathbf{u}^- \|^2_{L^2_t(\Gamma)} + \| \hat{\mathbf{n}} \times \mathbf{u}^+ \|^2_{L^2_t(\Gamma)} \,. \end{aligned} \tag{10}$$

Uniqueness Results

In order to prove uniqueness for the scattering problem (2)–(5) we will be based on the Bohren decomposition of the electric field $\mathbf{E}$ and magnetic field $\mathbf{H}$ into the $\mathbf{Q}_L$ (LCP) and $\mathbf{Q}_R$ (RCP) Beltrami fields

$$\mathbf{E} = \mathbf{Q}_L - i\eta\mathbf{Q}_R, \quad \mathbf{H} = \frac{1}{i\eta}\mathbf{Q}_L + \mathbf{Q}_R, \tag{11}$$

where $\eta = \sqrt{\frac{\mu}{\varepsilon}}$ is the intrinsic impedance of the chiral medium. In view of (11) the Beltrami fields are expressed as

$$\mathbf{Q}_L = \frac{\mathbf{E} + i\eta\,\mathbf{H}}{2}, \quad \mathbf{Q}_R = \frac{i\eta^{-1}\mathbf{E} + \mathbf{H}}{2}\,. \tag{12}$$

In addition the Beltrami fields satisfy the following equations:

$$\nabla \times \mathbf{Q}_L = \gamma_L\,\mathbf{Q}_L, \quad \nabla \times \mathbf{Q}_R = -\gamma_R\,\mathbf{Q}_R, \tag{13}$$

where $\gamma_L = k(1-k\beta)^{-1}$, $\gamma_R = k(1+k\beta)^{-1}$ are the wave numbers for the Beltrami fields, $\mathbf{Q}_L$, $\mathbf{Q}_R$, respectively.

The scattered Beltrami fields $\mathbf{Q}^{sc}_L, \mathbf{Q}^{sc}_R$, satisfy the Silver-Müller type radiation conditions [3, 5]

$$\hat{\mathbf{r}} \times \mathbf{Q}^{sc}_L + i\,\mathbf{Q}^{sc}_L = o(\frac{1}{r}), \quad \hat{\mathbf{r}} \times \mathbf{Q}^{sc}_R - i\,\mathbf{Q}^{sc}_R = o(\frac{1}{r}), \quad \text{as} \quad r \to \infty, \tag{14}$$

as well as the asymptotic relations

$$\mathbf{Q}^{sc}_L = O(\frac{1}{r}), \quad \mathbf{Q}^{sc}_R = O(\frac{1}{r}), \quad \text{as} \quad r \to \infty. \tag{15}$$

Relations (14) and (15) are obtained via (12) with the aid of the asymptotic behaviour of $\mathbf{E}$, $\mathbf{H}$ as $r \to \infty$, [12]. In what follows, with the notation $\overline{\mathbf{Q}}_A$, $A = L, R$, the bar "−" will denote the conjugate vector of $\mathbf{Q}_A$ and with $\mathbf{Q}_A^-$, $\mathbf{Q}_A^+$ we denote the limit from inside and outside of the boundary ∂D, respectively. In addition, the notation $\mathbf{Q}_A^\pm$ is for both the previous limits. We are now ready to proceed with the following proposition:

Theorem 2.1 *The Beltrami fields* $\mathbf{Q}_A$, $A = L, R$, *with* $\mathbf{Q}_A \in X(curl, \mathbb{R}^3 \setminus \overline{\Gamma})$, *satisfy the following relation:*

$$\int_{S_\rho} \hat{\mathbf{x}} \cdot (\mathbf{Q}_A \times \overline{\mathbf{Q}}_A)\, ds = \int_\Gamma \hat{\mathbf{n}} \cdot (\mathbf{Q}_A^- \times \overline{\mathbf{Q}}_A^-)\, ds - \int_\Gamma \hat{\mathbf{n}} \cdot (\mathbf{Q}_A^+ \times \overline{\mathbf{Q}}_A^+)\, ds, \tag{16}$$

where $S_\rho = \{\mathbf{x} \in \mathbb{R}^3 : |\mathbf{x}| = \rho\}$ *and* $\hat{\mathbf{x}}$ *is the unit normal vector to the spherical surface* S_ρ.

Proof The reader can be found an analogous proposition in [8], and hence the proof is omitted for brevity. ■

Further we have the following result:

Theorem 2.2 *The Beltrami fields* $\mathbf{Q}_L$, $\mathbf{Q}_R \in X(curl, \mathbb{R}^3 \setminus \overline{\Gamma})$ *satisfy the relation*

$$\Im\left(\frac{1}{\eta}\int_\Gamma \hat{\mathbf{n}} \cdot (\mathbf{Q}_L^\pm \times \overline{\mathbf{Q}}_L^\pm)\, ds - \eta \int_\Gamma \hat{\mathbf{n}} \cdot (\mathbf{Q}_R^\pm \times \overline{\mathbf{Q}}_R^\pm)\, ds\right)$$
$$= \frac{k}{\eta}\int_\Gamma \frac{1}{\lambda^\pm}(|\mathbf{U}^\pm|^2 - |\hat{\mathbf{n}} \cdot \mathbf{U}^\pm|^2)\, ds, \tag{17}$$

where $\mathbf{U}^\pm := \mathbf{Q}_L^\pm + i\,\eta\,\overline{\mathbf{Q}}_R^\pm$.

Proof The boundary conditions (3) and (4) via the relations (11) and (13) and the vector identity $\mathbf{u} = (\hat{\mathbf{n}} \cdot \mathbf{u})\hat{\mathbf{n}} - \hat{\mathbf{n}} \times (\mathbf{u} \times \hat{\mathbf{n}})$ lead to

$$\mathbf{Q}_L^\pm = i\eta\mathbf{Q}_R^\pm + [\hat{\mathbf{n}} \cdot (\mathbf{Q}_L^\pm - i\eta\mathbf{Q}_R^\pm)]\hat{\mathbf{n}}$$
$$+ \frac{ik^2}{\lambda^\pm}(\beta - \frac{\gamma_L}{\gamma^2})\, \hat{\mathbf{n}} \times \mathbf{Q}_L^\pm + \frac{\eta k^2}{\lambda^\pm}(\beta + \frac{\gamma_R}{\gamma^2})\, \hat{\mathbf{n}} \times \mathbf{Q}_R^\pm \quad \text{on } \Gamma, \tag{18}$$

but via

$$\beta - \frac{1}{\gamma_R} = -\frac{1}{k} \ \text{ and } \ \beta + \frac{1}{\gamma_L} = \frac{1}{k} \tag{19}$$

the relation (18) takes the form

$$\mathbf{Q}_L^\pm = i\eta\mathbf{Q}_R^\pm + [\hat{\mathbf{n}} \cdot (\mathbf{Q}_L^\pm - i\eta\mathbf{Q}_R^\pm)]\hat{\mathbf{n}}$$
$$- \frac{ik}{\lambda^\pm}\, \hat{\mathbf{n}} \times \mathbf{Q}_L^\pm + \frac{\eta}{\lambda^\pm}\, \hat{\mathbf{n}} \times \mathbf{Q}_R^\pm \quad \text{on } \Gamma. \tag{20}$$

Multiplying (20) by $\hat{\mathbf{n}}$ and then by $\overline{\mathbf{Q}}_L^{\pm}$ we arrive at

$$\begin{aligned}\hat{\mathbf{n}} \cdot (\mathbf{Q}_L^{\pm} \times \overline{\mathbf{Q}}_L^{\pm}) = i\eta\, \hat{\mathbf{n}} \cdot (\mathbf{Q}_R^{\pm} \times \overline{\mathbf{Q}}_L^{\pm}) \\ - \frac{ik}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_L^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_L^{\pm}) + \frac{\eta k}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_R^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_L^{\pm}) \\ + \frac{ik}{\lambda^{\pm}}\mathbf{Q}_L^{\pm} \cdot \overline{\mathbf{Q}}_L^{\pm} - \frac{\eta k}{\lambda^{\pm}}\mathbf{Q}_R^{\pm} \cdot \overline{\mathbf{Q}}_L^{\pm}\end{aligned} \tag{21}$$

as well as

$$\begin{aligned}\hat{\mathbf{n}} \cdot (\mathbf{Q}_R^{\pm} \times \overline{\mathbf{Q}}_R^{\pm}) = -\frac{i}{\eta}\, \hat{\mathbf{n}} \cdot (\mathbf{Q}_L^{\pm} \times \overline{\mathbf{Q}}_R^{\pm}) \\ + \frac{ik}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_R^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_R^{\pm}) + \frac{k}{\eta\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_L^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_R^{\pm}) \\ - \frac{ik}{\lambda^{\pm}}\mathbf{Q}_R^{\pm} \cdot \overline{\mathbf{Q}}_R^{\pm} - \frac{k}{\eta\lambda^{\pm}}\mathbf{Q}_L^{\pm} \cdot \overline{\mathbf{Q}}_R^{\pm}.\end{aligned} \tag{22}$$

For the remaining of the proof of (17), we use (21) and (22) in order to evaluate the quantity

$$\frac{1}{\eta}\hat{\mathbf{n}} \cdot (\mathbf{Q}_L^{\pm} \times \overline{\mathbf{Q}}_L^{\pm}) - \eta\hat{\mathbf{n}} \cdot (\mathbf{Q}_R^{\pm} \times \overline{\mathbf{Q}}_R^{\pm}). \tag{23}$$

Taking into account that

$$i\hat{\mathbf{n}} \cdot (\mathbf{Q}_R^{\pm} \times \overline{\mathbf{Q}}_L^{\pm}) + i\hat{\mathbf{n}} \cdot (\mathbf{Q}_L^{\pm} \times \overline{\mathbf{Q}}_R^{\pm}) \tag{24}$$

is a real number, after some calculations we can arrive at the relations

$$\frac{ik}{\eta\lambda^{\pm}}\mathbf{Q}_L^{\pm} \cdot \overline{\mathbf{Q}}_L^{\pm} - \frac{k}{\lambda^{\pm}}\mathbf{Q}_R^{\pm} \cdot \overline{\mathbf{Q}}_L^{\pm} + \frac{k}{\lambda^{\pm}}\mathbf{Q}_L^{\pm} \cdot \overline{\mathbf{Q}}_R^{\pm} + \frac{i\eta k}{\lambda^{\pm}}\mathbf{Q}_R^{\pm} \cdot \overline{\mathbf{Q}}_R^{\pm} = \frac{ik}{\eta\lambda^{\pm}}|\mathbf{U}^{\pm}|^2, \tag{25}$$

and

$$\begin{aligned}-\frac{ik}{\eta\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_L^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_L^{\pm}) + \frac{k}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_R^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_L^{\pm}) - \frac{k}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_L^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_R^{\pm}) \\ -\frac{i\eta k}{\lambda^{\pm}}(\hat{\mathbf{n}} \cdot \mathbf{Q}_R^{\pm})(\hat{\mathbf{n}} \cdot \overline{\mathbf{Q}}_R^{\pm}) = -\frac{ik}{\eta\lambda^{\pm}}|\hat{\mathbf{n}} \cdot \mathbf{U}^{\pm}|^2,\end{aligned} \tag{26}$$

and hence, the assertion of the proposition is proved. ■

In the sequel, uniqueness for the boundary value problem (2)–(5) will be established. We will consider the corresponding homogeneous scattering problem of (2)–(5), i.e., incident electric field $\mathbf{E}^{\text{inc}} = \mathbf{0}$. Relations (12), (16) and (23) will be used in order to prove the following uniqueness theorem:

Theorem 2.3 *The electromagnetic scattering problem (2)–(5) in chiral media, for* $\mathbf{E}^{inc} = \mathbf{0}$*, has the trivial solution.*

Proof By radiation conditions (14) we have

$$\lim_{\rho\to\infty}\left(\frac{1}{\eta}\int_{S_\rho}|\hat{\boldsymbol{\rho}}\times\mathbf{Q}_L + i\mathbf{Q}_L|^2 ds + \eta\int_{S_\rho}|\hat{\boldsymbol{\rho}}\times\mathbf{Q}_R - i\mathbf{Q}_R|^2 ds\right) = 0. \tag{27}$$

If $D_{ex} = \mathbb{R}^3\backslash D$, relation (27) with the aid of the divergence theorem in D and $D_{ex}\cap B_\rho$ for the vectors $\mathbf{Q}_A\times\overline{\mathbf{Q}}_A,\ A = L, R$ with $\mathbf{Q}_A \in X(\text{curl}, \mathbb{R}^3\setminus\overline{\Gamma})$, and due to (16), yields to

$$\begin{aligned}
&\lim_{\rho\to\infty}\left(\frac{1}{\eta}\int_{S_\rho}|\hat{\boldsymbol{\rho}}\times\mathbf{Q}_L|^2 ds + \frac{1}{\eta}\int_{S_\rho}|\mathbf{Q}_L|^2 ds + \eta\int_{S_\rho}|\hat{\boldsymbol{\rho}}\times\mathbf{Q}_R|^2 ds + \eta\int_{S_\rho}|\mathbf{Q}_R|^2 ds\right)\\
&+2\,\Im\left(\frac{1}{\eta}\int_\Gamma(\mathbf{Q}_L^+\times\overline{\mathbf{Q}}_L^+)\cdot\hat{\mathbf{n}}\,ds - \eta\int_\Gamma(\mathbf{Q}_R^+\times\overline{\mathbf{Q}}_R^+)\cdot\hat{\mathbf{n}}\,ds\right)\\
&+2\,\Im\left(\frac{1}{\eta}\int_\Gamma(\mathbf{Q}_L^-\times\overline{\mathbf{Q}}_L^-)\cdot\hat{\mathbf{n}}\,ds - \eta\int_\Gamma(\mathbf{Q}_R^-\times\overline{\mathbf{Q}}_R^-)\cdot\hat{\mathbf{n}}\,ds\right) = 0
\end{aligned} \tag{28}$$

Taking into account (17) and (28), via Rellich's lemma in chiral media [6], we arrive at $\mathbf{Q}_L = \mathbf{Q}_R = \mathbf{0}$, and from (11) the theorem now easily follows. ■

Existence of the Solution

In this section we will prove the existence of the solution of the scattering problem (2)–(5) using a variational method. Having in mind the Sobolev spaces defined in (6)–(9), we multiply equation (2) by a test function $\mathbf{w} \in X(\text{curl}, \mathbb{R}^3\setminus\overline{\Gamma})$ and we integrate by parts in D and $D_{ex}\cap B_\rho$. If we apply the divergence and the first vector Green's theorem in D and $D_{ex}\cap B_\rho$, in view of the continuity of $\hat{\mathbf{n}}\times\mathbf{E}$ and $\hat{\mathbf{n}}\times\nabla\times\mathbf{E}$ across $\partial D\setminus\overline{\Gamma}$ we can obtain the variational form of the scattering problem

$$\begin{aligned}
&\int_D(\nabla\times\mathbf{E})\cdot(\nabla\times\overline{\mathbf{w}})\,du + \int_{D_{ex}\cap B_\rho}(\nabla\times\mathbf{E})\cdot(\nabla\times\overline{\mathbf{w}})\,du\\
&-2\gamma^2\beta\int_D\mathbf{E}\cdot(\nabla\times\overline{\mathbf{w}})\,du - 2\gamma^2\beta\int_{D_{ex}\cap B_\rho}\mathbf{E}\cdot(\nabla\times\overline{\mathbf{w}})\,du\\
&-\gamma^2\int_D\mathbf{E}\cdot\overline{\mathbf{w}}\,du - \gamma^2\int_{D_{ex}\cap B_\rho}\mathbf{E}\cdot\overline{\mathbf{w}}\,du
\end{aligned}$$

$$\begin{aligned}
&+\frac{i\gamma^2}{k^2}\int_\Gamma \lambda_+\,\gamma_T^+\mathbf{E}\cdot\gamma_T^+\,\overline{\mathbf{w}}\,ds-\frac{i\gamma^2}{k^2}\int_\Gamma \lambda^-\,\gamma_T^-\mathbf{E}\cdot\gamma_T^-\,\overline{\mathbf{w}}\,ds\\
&-\gamma^2\beta\int_\Gamma(\hat{\mathbf{n}}\times\mathbf{E})\cdot\gamma_T^+\,\overline{\mathbf{w}}\,ds+\gamma^2\beta\int_\Gamma(\hat{\mathbf{n}}\times\mathbf{E})\cdot\gamma_T^-\,\overline{\mathbf{w}}\,ds\\
&+\int_{S_\rho} G_{kce}(\hat{\mathbf{x}}\times\mathbf{E})\cdot\gamma_T\overline{\mathbf{w}}ds\\
=\;&-\frac{i\gamma^{\,2}}{k^2}\int_\Gamma \lambda^+\,\gamma_T^+\,\mathbf{E}^{\text{inc}}\cdot\gamma_T^+\overline{\mathbf{w}}ds+\frac{i\gamma^{\,2}}{k^2}\int_\Gamma \lambda^-\,\gamma_T^-\mathbf{E}^{\text{inc}}\cdot\gamma_T^-\overline{\mathbf{w}}ds\\
&+\gamma^2\beta\int_\Gamma(\hat{\mathbf{n}}\times\mathbf{E}^{\text{inc}})\cdot\gamma_T^+\overline{\mathbf{w}}\,ds-\gamma^2\beta\int_\Gamma(\hat{\mathbf{n}}\times\mathbf{E}^{\text{inc}})\cdot\gamma_T^-\overline{\mathbf{w}}\,ds\\
&-\int_{S_\varrho} G_{kce}(\hat{\mathbf{x}}\times\mathbf{E}^{\text{inc}})\cdot\gamma_T\overline{\mathbf{w}}\,ds\;, \qquad (29)
\end{aligned}$$

where $\mathbf{E}^{\text{inc}}$ is a given field and G_{kce} is a Calderon type operator in chiral media which maps a tangential vector field $\hat{\mathbf{x}}\times\mathbf{E}$ on S_ρ to an also tangential vector field $\hat{\mathbf{x}}\times(\nabla\times\mathbf{E}-2\gamma^2\beta\,\mathbf{E})$ on the same surface space. These operators for non-chiral media have been studied in [13] and [18]. The authors of this article will present Calderon type operators in chiral media, as well as their identities, in a future work.

We are going to prove gradually the existence theorem:

Theorem 3.1 *For any given field $\mathbf{E}^{\text{inc}}\in X(curl,\mathbb{R}^3\setminus\overline{\Gamma})$ the electromagnetic scattering problem in chiral media (29) has a unique solution $\mathbf{E}\in X(curl,\mathbb{R}^3\setminus\overline{\Gamma})$.*

The scattered field $\mathbf{E}$ in (29) also satisfies

$$\nabla\times\nabla\times\mathbf{E}=2\,\gamma^2\beta\,\nabla\times\mathbf{E}+\gamma^2\mathbf{E},\quad \text{in } \mathbb{R}^3\setminus\overline{B}_\rho \qquad (30)$$

$$\hat{\mathbf{x}}\times\mathbf{E}=\xi,\quad \text{on } S_\rho \qquad (31)$$

$$\hat{\mathbf{r}}\times\nabla\times\mathbf{E}-\beta\,\gamma^2\,\hat{\mathbf{r}}\times\mathbf{E}+\frac{i\gamma^2}{k}\mathbf{E}=o(\frac{1}{r}),\qquad r\to\infty \qquad (32)$$

where $\xi\in L_t^2(S_\rho)$. We note that in (29), we have taken into account that for the incident electric field $\mathbf{E}^{\text{inc}}$ the equation

$$\nabla\times\nabla\times\mathbf{E}^{\text{inc}}-2\,\gamma^2\,\beta\,\nabla\times\mathbf{E}^{\text{inc}}-\gamma^2\mathbf{E}^{\text{inc}}=\mathbf{0},\quad \text{in } \mathbb{R}^3 \qquad (33)$$

holds. We now define the space

$$S:=\left\{p\in H^1(B_\rho\setminus\overline{\Gamma}):\,p^-|_\Gamma=c^-\text{ and }p^+|_\Gamma=c^+\right\}, \qquad (34)$$

where c^+ and c^- are constant numbers, as well as the space

$$X^0 := \left\{ \mathbf{u} \in X(\text{curl}, \mathbb{R}^3 \setminus \overline{\Gamma}) : < G_{kce}(\hat{\mathbf{x}} \times \mathbf{u}) , \nabla_{S_\rho} q > - \gamma^2 (\mathbf{u} , \nabla q)_{B_\rho} = 0, \quad \text{for } q \in S \right\}. \tag{35}$$

Then we write (29) in a more compact form

$$A(\mathbf{u}, \mathbf{w}) = B(\mathbf{w}), \tag{36}$$

where

$$\begin{aligned} A(\mathbf{u}, \mathbf{w}) &= (\nabla \times \mathbf{u}, \nabla \times \mathbf{w})_D + (\nabla \times \mathbf{u}, \nabla \times \mathbf{w})_{D_{ex} \cap B_\rho} \\ & -2\gamma^2 \beta \left((\mathbf{u}, \nabla \times \mathbf{w})_D + (\mathbf{u}, \nabla \times \mathbf{w})_{D_{ex} \cap B_\rho} \right) \\ & -\gamma^2 ((\mathbf{u}, \mathbf{w})_D + (\mathbf{u}, \mathbf{w})_{D_{ex} \cap B_\rho}) + < G_{kce}(\hat{\mathbf{x}} \times \mathbf{u}), \gamma_T \mathbf{w} >_{S_\rho} \\ & + \frac{i\gamma^2}{k^2} < \lambda^+ \gamma_T^+ \mathbf{u}, \gamma_T^+ \mathbf{w} >_\Gamma - \frac{i\gamma^2}{k^2} < \lambda^- \gamma_T^- \mathbf{u}, \gamma_T^- \mathbf{w} >_\Gamma \\ & -\gamma^2 \beta < \hat{\mathbf{n}} \times \mathbf{u}, \gamma_T^+ \mathbf{w} >_\Gamma + \gamma^2 \beta < \hat{\mathbf{n}} \times \mathbf{u}, \gamma_T^- \mathbf{w} >_\Gamma, \end{aligned} \tag{37}$$

and the right part of equation (36), due to (29), consists of boundary data

$$\begin{aligned} B(\mathbf{w}) = \frac{i\gamma^2}{k^2} &< \lambda^- \gamma_T^- \mathbf{E}^{\text{inc}}, \gamma_T^- \mathbf{w} >_\Gamma - \frac{i\gamma^2}{k^2} < \lambda^+ \gamma_T^+ \mathbf{E}^{\text{inc}}, \mathbf{w} >_\Gamma \\ & + \gamma^2 \beta < \hat{\mathbf{n}} \times, \mathbf{E}^{\text{inc}} \gamma_T^+ \mathbf{w} >_\Gamma - \gamma^2 \beta < \hat{\mathbf{n}} \times \mathbf{E}^{\text{inc}}, \gamma_T^- \mathbf{w} >_\Gamma \\ & - < G_{kce}(\hat{\mathbf{x}} \times \mathbf{E}^{\text{inc}}), \gamma_T \mathbf{w} >_{S_\rho} . \end{aligned} \tag{38}$$

The first step is to prove the following:

Lemma 3.2 *The equation $A(\nabla p , \nabla q) = B(\nabla q)$ has a unique solution for any $q \in S$.*

Proof We put $\mathbf{u} = \nabla p$ and $\mathbf{w} = \nabla q$ so the equation (36) takes the form

$$-\gamma^2 (\nabla p, \nabla q)_{L^2(B_\rho)} + < G_{kce}(\hat{\mathbf{x}} \times \nabla p), \nabla_{S_\rho} q >_{S_\rho} = B(\nabla q), \tag{39}$$

since $(\nabla p)_T = \hat{\mathbf{n}} \times \nabla p \times \hat{\mathbf{n}} = \gamma_T \nabla p = \mathbf{0}$ for $p \in S$. Then compactness properties of the Calderon type operators allow us to apply the usual procedure of the Fredholm alternative theory to (39) as in [18] in order to complete the proof. ■

We move on with the next step which is the lemma below:

Lemma 3.3 $\nabla\, S$ *is a closed subspace of* $X(curl,\, \mathbb{R}^3 \setminus \overline{\Gamma})$ *and*

$$X(curl,\, \mathbb{R}^3 \setminus \overline{\Gamma}) = X^0 \oplus \nabla S. \tag{40}$$

Proof The space ∇S is closed in $X(\text{curl},\, \mathbb{R}^3 \setminus \overline{\Gamma})$ since S is closed in $H^1(B_\rho \setminus \overline{\Gamma})$ [11].

Then if $\mathbf{u} \in X(\text{curl},\, \mathbb{R}^3 \setminus \overline{\Gamma})$ is a solution of (36), we have $A(\mathbf{u}, \nabla\, q) = B(\nabla\, q)$ for any $q \in S$. We consider that $\mathbf{u} = \mathbf{v} + \nabla\, p_0$, where $\nabla\, p_0$ is the unique solution of (39), so it holds $A(\mathbf{v}\,, \nabla\, q) = 0$ and from the definition (35) we take $\mathbf{v} \in X^0$. Then it is easy to prove that this expression of $\mathbf{u}$ as a sum of elements of $\nabla\, S$ and X^0 is the unique one. ■

Then we are going to deal with the equation $A(\mathbf{u}, \mathbf{v}) = B(\mathbf{v})$, for any $\mathbf{v} \in X^0$, which finally takes the form

$$A(\mathbf{w}, \mathbf{v}) = B(\mathbf{v}) - A(\nabla\, p_0, \mathbf{v}), \quad \text{for any } \mathbf{v} \in X^0. \tag{41}$$

We continue our proof with the following result, due to [8].

Lemma 3.4 *The space* X^0 *is compactly imbedded in* $L^2(B_\rho)$.

Proof We consider a sequence $\left\{\mathbf{u}_n^{ex}\right\}_{n=1}^{\infty}$ of solutions of the scattering problem

$$\nabla \times \nabla \times \mathbf{u}_n^{ex} = 2\,\gamma^2 \beta\, \nabla \times \mathbf{u}_n^{ex} + \gamma^2 \mathbf{u}_n^{ex} \quad \text{in } \mathbb{R}^3 \setminus \overline{B}_\rho, \tag{42}$$

$$\hat{\mathbf{x}} \times \mathbf{u}_n^{ex} = \hat{\mathbf{x}} \times \mathbf{u}_n \quad \text{on } S_\rho, \tag{43}$$

$$\hat{\mathbf{r}} \times \nabla \times \mathbf{u}_n^{ex} - \beta\, \gamma^2\, \hat{\mathbf{r}} \times \mathbf{u}_n^{ex} + \frac{i\,\gamma^2}{k}\, \mathbf{u}_n^{ex} = o(\frac{1}{r}) \quad r \to \infty, \tag{44}$$

where $\{\mathbf{u}_n\}_{n=1}^{\infty}$ is a given bounded sequence in X^0. For the solutions $\mathbf{u}_n^{ex}$ we can give series expansions using proper vector wave functions in chiral media analogous to [18]. The boundary condition (43) and the definition (35) lead to the conclusion that the vectors $\mathbf{u}_n^{ex}$ and $\mathbf{u}_n$ have equal normal and tangential components on S_ρ so each element of the sequence $\mathbf{u}_n$ can be extended to a function $\mathbf{u}_n^0 \in H_{loc}(\text{curl}, B_\rho \setminus \overline{\Gamma})$ to all $\mathbb{R}^3$, defined as,

$$\mathbf{u}_n^0 = \begin{cases} \mathbf{u}_n & \text{in } B_\rho, \\ \\ \mathbf{u}_n^{ex} & \text{in } \mathbb{R}^3 \setminus \overline{B}_\rho. \end{cases} \tag{45}$$

Following analogous ideas for chiral media as those in [11], the proof is completed. ■

The above result allows us to define proper compact operators in order to apply again the Fredholm alternative theory to (41), and with the aid of Lemma 3.4 to prove that Eq. (41) has a unique solution. This conclusion completes the proof of Theorem 3.1 for the existence of the solution of (29), and therefore establishes the existence result for the scattering problem (2)–(5).

Conclusions

This paper was concerned with the solvability of the direct electromagnetic scattering problem by a chiral impedance screen in a chiral environment. In particular, the terms $2\gamma^2\,\beta\nabla\times\mathbf{E}^{\mathrm{inc}}$, $\gamma^2\beta\,\hat{\mathbf{n}}\times\mathbf{E}$ in (2)–(4) were the main reason for using the Beltrami fields in order to prove uniqueness for the electromagnetic problem in chiral media. We also make the following remarks:

1. If $\beta = 0$, i.e., non-chiral environment, the approach for existence and uniqueness is similar to the case for the mixed scattering problem in [11], which holds for scattering by a screen in non-chiral media. In addition if $\lambda^+ = 0$ and $\lambda^- = 0$, we can analogous prove that the scattering problem (2)–(5) has a unique solution.
2. In the case where the chirality measure β is not a constant, our method can also be applied, since the modifications that occurred can be handled.

References

1. H. Ammari, J.C. Nedelec, Time-harmonic electromagnetic fields in thin chiral curved layers. SIAM J. Math. Anal. **29**(2), 395–423 (1998)
2. H. Ammari, K. Hamdache, J.C. Nedelec, Chirality in the Maxwell equations by the dipole approximation method. SIAM J. Appl. Math. **59**, 2045–2059 (1999)
3. C.E. Athanasiadis, E. Kardasi, Beltrami Herglotz functions for electromagnetic scattering. Appl. Anal. **84**(2), 145–163 (2005)
4. C.E. Athanasiadis, P.A. Martin, I.G. Stratis, Electromagnetic scattering by a homogeneous chiral obstacle: boundary integral equations and low–chirality approximations. SIAM J. Appl. Math. **59**(5), 1745–1762 (1999)
5. C.E. Athanasiadis, G. Costakis, I.G. Stratis, Electromagnetic scattering by a homogeneous chiral obstacle in a chiral environment. IMA J. Appl. Math. **64**, 245–258 (2000)
6. C.E. Athanasiadis, G. Costakis, I.G. Stratis, Electromagnetic scattering by a perfectly conducting obstacle in a homogeneous chiral environment: solvability and low-frequency theory. Math. Methods Appl. Sci. **25**, 927–944 (2002)
7. C.E. Athanasiadis, D. Natrosvili, V. Sevroglou, I.G. Stratis, A boundary integral equations approach for direct mixed impedance problems in elasticity. Integr. Equ. Appl. **23**(2), 183–222 (2011)
8. C.E. Athanasiadis, V. Sevroglou, K.I. Skourogiannis, The direct electromagnetic scattering problem by a mixed impedance screen in chiral media. Appl. Anal. **91**(11), 1–11 (2012)
9. C.E. Athanasiadis, V. Sevroglou, K.I. Skourogiannis, The inverse electromagnetic scattering problem by a mixed impedance screen in chiral media. Inv. Prob. Imag. **9**(4), 951–970 (2015)
10. F. Cakoni, D. Colton, *Qualitative Methods in Inverse Electromagnetic Scattering Theory* (Springer, Berlin, 2005)
11. F. Cakoni, E. Darringrand, The inverse electromagnetic scattering problem for a mixed boundary value problem for screens. Comput. Appl. Math. **174**, 251–269 (2005)
12. D. Colton, R. Kress, *Integral Equation Methods in Scattering Theory* (Wiley, New York, 1983)
13. D. Colton, R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory* (Springer, Berlin, 1998)
14. A. Lakhtakia, *Beltrami Fields in Chiral Media* (World Scientific, Singapore, 1994)

15. A. Lakhtakia, V.K. Varadan, V.V. Varadan, *Time-harmonic Electromagnetic Fields in Chiral Media*. Lecture Notes in Physics (Springer, Berlin, 1989)
16. A. Lakhtakia, V.K. Varadan, V.V. Varadan, Surface integral equations for scattering by PEC scatterers in isotropic chiral media. Int. J. Eng. Sci. **29**, 79–185 (1991)
17. I.V. Lindell, A.H. Sihvola, S.A. Tretyakov, A.J. Viitanen, *Electromagnetic Waves in Chiral and Bi-isotropic Media* (Artech House, Boston, 1994)
18. P. Monk, *Finite Element Methods for Maxwell's Equations* (Clarendon, Oxford, 2003)

Optimal Batch Production with Rework Process for Products with Time-Varying Demand Over Finite Planning Horizon

Lakdere Benkherouf, Konstantina Skouri, and Ioannis Konstantaras

Abstract In this paper a finite planning horizon, production–inventory model with rework, is considered. During the production process defective items are produced. These items, after the end of the production process, are repaired and converted into items of perfect quality. The demand for the item is assumed to be time varying. The objective is the determination of the production–reworking schedule that minimizes the total cost over the planning horizon. A procedure is proposed for the determination of such schedule.

Keywords Production • Imperfect • Rework • Finite horizon • Time-varying demand

2010 Mathematics Subject Classification: 90B05; 90B30

Introduction

Economic Production Quantity (EPQ) model determines the quantity that should be produced to minimize the total inventory costs by balancing the holding cost and fixed setup cost. The EPQ model uses the same assumptions as the basic EOQ model

L. Benkherouf
Faculty of Science, Department of Statistics and Operations Research,
Kuwait University, Safat, Kuwait
e-mail: lakdere.benkherouf@ku.edu.kw

K. Skouri (✉)
Department of Mathematics, University of Ioannina, Ioannina, Greece
e-mail: kskouri@uoi.gr

I. Konstantaras
Department of Business Administration, School of Business Administration, University
of Macedonia, Thessaloniki, Greece
e-mail: ikonst@uom.gr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_3

in Harris [1], except that it uses a finite replenishment rate. However, the possibility of producing defective items is not taken into account in EPQ model. Indeed, product quality is not always perfect and defective items may be produced during the regular production cycle. The reasons that defective items may be produced are the deterioration of the production process and the imperfect quality of the components and subassemblies that are procured from the suppliers. These defective items can be rejected or reworked and sold as new to secondary markets with a low price. However, there are products, like printed circuit boards (pcbs) with high production requirements (because of their use in smart phones and tablets), and at the same time, thousands of boards are damaged and need rework. Since the cost of these pcbs can be significant, good planning of the production rework process is important.

The complexity of a production/inventory model depends heavily on the assumptions that someone makes for demand pattern and treatment of defective items. Thus, several models have been proposed in the inventory control literature, which examine the effect of imperfect quality production on the economic production quantity. Rosenblatt and Lee [2] were the first who proposed a modified version of the EPQ model assuming that the production process shifts from an in-control to an out-of-control state. They derived the optimal production cycle and showed that it is shorter than that of the EPQ model. Hariga and Ben-Daya [3] extended the model proposed by Rosenblatt and Lee [2] by considering the general time shift distribution. Moon et al. [4] considered the Economic Lot Scheduling (ELS) problem with imperfect production process with sequence-independent setups and developed mathematical models under common cycle and time-varying lot sizes approaches. Wang [5] proposed a mathematical model to determine the joint optimal production run length and inspection policy under the assumption that product inspections are performed at the end of the production run. Sana et al. [6] extended the EPQ model in an imperfect production situation where the defective items are sold at a reduced price while Cardenas-Barron [7] examined the effects of shortages on the production quantity for the EPQ model with imperfect production.

Hayek and Salameh [8] studied an EPQ model with rework of imperfect quality items and derived an optimal solution for the production quantity. Chiu [9] studied the effects of the reworking of defective items on the EPQ model with shortages. Jamal et al. [10] developed a mathematical model for the optimal production quantity assuming rework of defective items. They considered two different policies according to the time that the defective items are reworked. Cardenas-Barron [11] proposed a simple algebraic procedure to determine the optimal solution for the two inventory models which were proposed in by Jamal et al. [10]. Biswas and Sarker [12] studied an inventory model assuming inspection of the new manufactured products, rework and scrap of the defective items. They assumed that the defective items are reworked within the same cycle and obtained closed-form expression for the optimal batch quantity with various cases of scrap detection. Chiu et al. [13] developed a mathematical model to determine the optimal run time for an imperfect finite production rate model with scrap and examined the joint effects of rework and machine breakdown. Chung et al. [14] proposed some EPQ-type inventory models for deterioration items with machine unavailability and shortages while Wee et al. [15] derived closed-form expressions of the optimal production and backordering lots for an EPQ model with imperfect quality items, shortages and screening constraints. Tai [16] studied a similar model to that of Wee et al. [15] assuming inspection errors and deterioration of good quality items. Pal et al. [17] considered an imperfect production system which undergoes in out-of-control state from in-control one after a time that follows a probability density function. Sarkar et al. [18] extended the model proposed in [7] assuming that defective rate is random and studied three different distribution density functions for this rate. Recently, Sivashankari and Panayappan [19] studied a single stage manufacturing system that generates imperfect quality products, where a proportion of defective products are reworked in the same cycle and for this system they developed a production–inventory model with planned backorders to determine the optimum production lot size.

Most of the above surveyed work adopted an infinite planning horizon with constant demand rate. When the demand rate varies with time the use of the demand information over a finite planning period is required (Silver et al. [20]). Production–inventory models with time-varying demand and finite planning horizon have been developed for deteriorating items, which can be considered as some kind of imperfect items. Specifically, Balkhi [21] studied a production–inventory system assuming that production, demand and deteriorating rates are continuous function of time. Sana et al. [22] presented a model with constant production and deterioration rates and linear time dependent demand rate. They used the box complex algorithm to find the optimal inventory policy. Roy et al. [23] developed a model for a randomly deteriorating item with linearly time-varying demand rate and where the production rate is considered as a decision variable. A genetic algorithm is used to minimize the total inventory costs. Yang [24] extended the model of [21] assuming partial backlogging. Benkherouf and Boushehri [25] considered a model with constant production and deterioration rates together with time varying demand rate. They provided technical conditions which ensure existence and uniqueness of an optimal inventory policy. Das et al. [26] developed a two warehouse production–inventory model for a deteriorating item with time-varying demand. However, in all these models no repair or rework process was considered.

Finally, we should note that a different approach, for catering for certain inventory production models with defective items, may be found in Benkherouf et al. [27] and the references therein. There, used products are returned by customers and after inspection they can be classified either as “remanufacturable” or as “refurbishable” items. There are two facilities for storing the return and new goods. The remanufacturing process brings “remanufacturable” items up to quality standards that are as rigorous as those of new items. The refurbished items are sold to a secondary market at a reduced price. The optimal inventory policy for a finite planning horizon reduces (in some cases) to a similar class of optimization problems treated in the present paper.

In the present paper, we study an imperfect production system that produces imperfect quality items and where a proportion of the defective items are reworked in the same cycle along the same lines of [10] and Benkherouf and Omar [28].

Table 1 Comparison between the basic characteristics of proposed model and other existing ones

Model	Demand rate	Production rate	Rate of imperfect	Reworked quantity	Horizon
Balkhi [21]	Time varying	Time varying	Time varying	None	Finite
Sana et al. [22]	Time varying	Constant	Constant	None	Finite
Biswas and Sarker [12]	Constant	Constant	Constant	Proportion	Infinite
Roy et al. [23]	Time varying	Decision variable	Constant	None	Finite
Yang [24]	Time varying	Time varying	Constant	None	Finite
Benkherouf and Boushehri [25]	Time varying	Constant	Constant	None	Finite
Tai et al. [16]	Constant	Constant	Constant	Proportion	Infinite
Sarkar et al. [18]	Constant	Constant	Random	All	Infinite
Das et al. [26]	Time varying	Constant	Constant	None	Finite
Sivashankari and Panayappan [19]	Constant	Constant	Constant	Proportion	Infinite
The proposed model	Time varying	Constant	Constant	All or proportion	Finite

After the reworking process, the reworked items are considered as good as new and used to satisfy the demand. We assume that the demand rate for the product is not constant but varying with time and also the planning horizon is not infinite but finite. These two modifications lead to a change in the corresponding optimization problem. In Table 1, the differences between the proposed model and the more relevant existing ones are summarized.

The remainder of this paper is organized as follows. The assumptions and notations of the inventory problem considered are presented in section “Assumptions and Notations”. Section “The mathematical model” contains the mathematical model while the solution procedure is presented in section “Optimization procedure”. Section “Numerical example” presents and discusses numerical results. Finally, this paper summarizes, concludes and proposes future research extension in section “Conclusions”.

Assumptions and Notations

The production–inventory model of the present paper is developed under the following assumptions:

(1) The planning horizon of the system is finite and is taken as H time units, $H > 0$. The initial and the final inventory levels are both zero.
(2) The demand rate at time t is given by a continuous function $D, D(t) : [0, H] \rightarrow (0, \infty)$.
(3) Items are produced at a fixed production rate $p, p > 0$.
(4) During the production process defective items are produced at fixed rate $a > 0$.
(5) The production rate is such that $p > D(t) + a$, for $t \in [0, H]$.
(6) The reworking process starts after the end of the regular production process.
(7) All defective items are reworked at a fixed rework rate $\gamma > 0$, where $D(t) \geq \gamma$.
(8) No defective items are produced during rework.
(9) The demand is satisfied from product which are produced during the production process and the reworking process.
(10) The lead time is zero.
(11) Shortages are not allowed during the planning horizon.

Notations

H	the planning horizon.
n	number of production–reworking cycles (cycles).
t_i	starting time of the production process in cycle i, $i = 1, \ldots, n$.
t_i^p	production stopping time in cycle i, $i = 1, \ldots, n$.
τ_i	stopping time of the reworking process in cycle i, $i = 1, \ldots, n$.
c_0	setup cost (sum of production and reworking setup cost).
c_h	holding cost per unit per unit time.
c_p	unit production cost.
c_r	cost of reworking per unit.

The Mathematical Model

In this section the mathematical formulation of the model is developed. The planning horizon is made up of multiple cycles. A typical cycle i (see Fig. 1), say, begins at time t_{i-1} and ends at time t_i, where $t_i > t_{i-1} \geq 0$, for $i = 1, \cdots, n$ with $t_0 = 0$ and $t_n = H$. During the i-th cycle, the regular production process lasts from time t_{i-1} to t_i^p and some of the produced items are defective. These items are immediately reworked on the interval $[t_i^p, \tau_i)$ where $\tau_i < t_i$ at a rate $\gamma > 0$. Then due to demand the inventory falls to zero at time t_i.

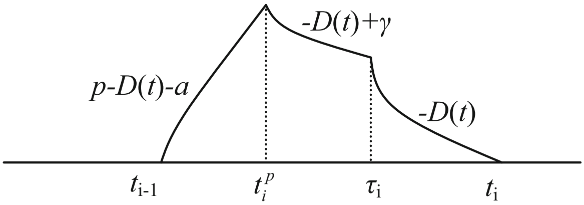


Fig. 1 The variation of the inventory for the *i*-th cycle

On the time interval $[t_{i-1}, t_i^p]$ the inventory level is affected by production, demand and defective items, so the following differential equation describes its variation:

$$\frac{dI(t)}{dt} = p - a - D(t),\ t_{i-1} \le t \le t_i^p, \tag{1}$$

with $I(t_{i-1}) = 0$. By setting $\eta = p - a$, then the solution of (1) is:

$$I(t) = \int_{t_{i-1}}^{t} \{\eta - D(u)\}du, \tag{2}$$

The inventory level during $[t_i^p, \tau_i]$ is affected by reworking process and demand and it is described by differential equation:

$$\frac{dI(t)}{dt} = -D(t) + \gamma,\ t_i^p \le t \le \tau_i, \tag{3}$$

with

$$I(t_i^p) = \int_{t_{i-1}}^{t_i^p} \{\eta - D(u)\}du. \tag{4}$$

The solution of the above differential equation is:

$$I(t) = -\int_{t_i^p}^{t} \{D(u) - \gamma\}du + I(t_i^p). \tag{5}$$

The inventory level during $[\tau_i, t_i]$ decreases due to demand and it is described by the differential equation:

$$\frac{dI(t)}{dt} = -D(t),\ \tau_i \le t \le t_i, \tag{6}$$

with $I(t_i) = 0$. The solution of (6) is:

$$I(t) = \int_t^{t_i} D(u)du. \tag{7}$$

Since $I(t)$ is continuous on $[t_i^p, t_i]$ the following relation prevails:

$$\int_{\tau_i}^{t_i} D(u)du = -\int_{t_i^p}^{\tau_i} \{D(u) - \gamma\}du + I(t_i^p), \tag{8}$$

or

$$\int_{t_i^p}^{t_i} D(u)du = \gamma(\tau_i - t_i^p) + I(t_i^p), \tag{9}$$

or using (4)

$$\int_{t_{i-1}}^{t_i} D(u)du = \gamma(\tau_i - t_i^p) + \eta(t_i^p - t_{i-1}) \tag{10}$$

Taking into account assumption (7) which is expressed as:

$$a(t_i^p - t_{i-1}) = \gamma(\tau_i - t_i^p), \tag{11}$$

relation (10) gives:

$$\int_{t_{i-1}}^{t_i} D(u)du = a(t_i^p - t_{i-1}) + \eta(t_i^p - t_{i-1}) = p(t_i^p - t_{i-1}), \tag{12}$$

or equivalently

$$t_i^p = t_{i-1} + \frac{1}{p}\int_{t_{i-1}}^{t_i} D(u)du. \tag{13}$$

In addition, by (11), the following relation holds:

$$\tau_i - t_i^p = \frac{a}{\gamma}(t_i^p - t_{i-1}) = \frac{a}{\gamma p}\int_{t_{i-1}}^{t_i} D(u)du, \tag{14}$$

and consequently τ_i is expressed as a function of t_{i-1} and t_i:

$$\tau_i = t_i^p + \frac{a}{\gamma p}\int_{t_{i-1}}^{t_i} D(u)du = t_{i-1} + \frac{a+\gamma}{\gamma p}\int_{t_{i-1}}^{t_i} D(u)du. \tag{15}$$

The total inventory cost over the finite horizon is the sum of: setup cost: c_0n, production cost:

$$c_p\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i^p} pdt = c_p p\sum_{i=1}^{n}(t_i^p - t_{i-1}) = c_p\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i} D(u)du = c_p\int_0^H D(u)du, \quad (16)$$

reworking cost:

$$c_r\sum_{i=1}^{n}\int_{t_i^p}^{\tau_i} \gamma dt = c_r\gamma\sum_{i=1}^{n}(\tau_i - t_i^p) = c_r\gamma\sum_{i=1}^{n}\frac{a}{\gamma p}\int_{t_{i-1}}^{t_i} D(u)du = c_r\frac{a}{p}\int_0^H D(t)dt, \quad (17)$$

holding cost. This is given by the following theorem:

Theorem 1 *The total holding costs is given by:*

$$\begin{aligned} &HC(t_1, t_2, \ldots, t_n, n) \\ &= c_h\sum_{i=1}^{n}\left[\int_{t_{i-1}}^{t_i}(t - t_{i-1})D(t)dt - \frac{a^2 + \gamma(a+p)}{2\gamma p^2}\left\{\int_{t_{i-1}}^{t_i} D(u)du\right\}^2\right]. \end{aligned} \quad (18)$$

Proof The proof follows from similar arguments used in [28].

The aim is to find the production–inventory policy, which minimizes the total cost over the finite planning horizon. This requires to find the optimal number of production–reworking cycles, n, the optimal production starting, t_i, and stopping, t_i^p, times and reworking stopping times, τ_i. From (16) and (17) it is obvious that the production and reworking costs are constants over the planning horizon. This means that they do not influence the inventory policy. From (13) and (15) when t_i are determined then t_i^p and τ_i are also determined. So, the optimal policy is obtained from the solution of the following optimization problem:

$$P:$$

$$\min_{(t_1,t_2,\ldots,t_n,n)} nc_0 + c_h\sum_{i=1}^{n}\left[\int_{t_{i-1}}^{t_i}(t - t_{i-1})D(t)dt - \frac{a^2 + \gamma(a+p)}{2\gamma p^2}\left\{\int_{t_{i-1}}^{t_i} D(t)dt\right\}^2\right]$$

subject to: $0 = t_0 \le t_1 \le \ldots \le t_n = H$, and n integer.

Remark 1 Assumption (7) can be easily modified in order for a proportion, say β, of defective items to be reworked. In this case (11) is modified as:

$$a\beta(t_i^p - t_{i-1}) = \gamma(\tau_i - t_i^p), \quad (19)$$

In the next section results are presented that ensure the existence of the optimal values for the decision variables $t_1, t_2, \ldots, t_n, n$.

Optimization Procedure

The optimization problem P can be handled using results found in Al-Khamis et al. [29]. These results are adapted to problem P and are stated with no proof. Let

$$p' = \frac{\gamma p^2}{a^2 + \gamma(a+p)}. \tag{20}$$

Theorem 2 *For fixed n and* $t \in [0, H]$*, if (i) the demand rate is log-concave and differentiable in t such that* $p' > D(t)$ *and (ii)* $D'(t)/\{p' - D(t)\}$ *is non-decreasing in t, Problem P has a unique optimal solution. Moreover, if* $Z(n)$ *denotes the value of the objective function at this minimum, then* $Z(n)$ *is convex in n.*

Note that the requirement for the demand rate to be log-concave is to ensure uniqueness of the optimal inventory policy. Log-concave functions include the linear and the exponential function and many other functions. For example, many probability distributions are log-concave such as the normal distribution, the logistic distribution and others: see Dharmadhikari and Kumar [30].

Theorem 3 below is a direct consequence of the convexity of $Z(n)$. This contains a procedure (similar to the one described in Rau and Ou Yang [31] or Benkherouf and Gilding [32]) which may be applied to find the optimal value of production–reworking cycles, n. Let $S(n) := Z(n) - nc_0$. Then, we have

Theorem 3 *The optimal number of replenishment schedule is such that:*

(i) If $c_0 > S(1) - S(2)$*, then the optimal number of replenishment schedule is* $n = 1$.

(ii) If there exists an $N \geq 0$ *such that* $S(N-1) - S(N) > c_0 > S(N) - S(N+1)$*, then the optimal n is N.*

(iii) If there exists an $N \geq 1$ *such that* $c_0 = S(N) - S(N+1)$*, then there are two optimal values for n: N and* $N+1$.

Numerical Example

In order to obtain some managerial insight from the presented model the following data are used $D(t) = 15\exp(0.2t)$, $P = 7000$, $a = 70$, $\gamma = 300$, $c_h = 5$, $c_o = 900$, $c_p = 50$, $c_r = 25$, $H = 5$. For these data the optimal number of production–rework cycles is $n = 6$ and the optimal cost is 263426.33. In Table 2 the optimal production–rework schedule is presented, i.e., the optimal production starting times, t_i, $i = 1, \cdots, n$, the optimal production stopping times, t_i^p, $i = 1, \cdots, n$ and the optimal reworking stopping times, $\tau_i, i = 1, \ldots, n$.

Table 3 shows the impact to optimal cost in consequence of changes to input parameters.

Table 2 The optimal t_i, t_i^p, and τ_i, $i = 1, \ldots, 6$ ($D(t) > \gamma$)

t_1	t_2	t_3	t_4	t_5	t_6
1.935	2.944	3.621	4.139	4.578	5
t_1^p	t_2^p	t_3^p	t_4^p	t_5^p	t_6^p
0.016	1.978	3.021	3.739	4.317	4.864
τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
0.020	1.988	3.038	3.767	4.358	4.931

Table 3 Sensitivity analysis

Parameter	n	Optimal cost
$P = 7500$	7	263487.74
$P = 8000$	6	263533.48
$a = 140$	6	264643.57
$a = 210$	6	265849.46
$\gamma = 400$	6	263427.68
$\gamma = 6500$	6	263431.61
$c_0 = 1000$	6	264026.38
$c_0 = 1100$	6	264626.38
$c_h = 6$	7	264304.68
$c_h = 7$	7	265176.18

From the results given in tables it seems that the model is not sensitive to the parameter changes both in terms of the optimal cost and in terms of decision variables. Even in the case of the significant change in the reworking rate, γ, the production–rework schedule remains virtually unchanged, probably because of the low defective rate.

Conclusions

In this study a production system is considered that produces defective items. All (or part of) defective items are reworked in the same production cycle. After the reworking process, the items are considered good as new and used to satisfy the demand. The demand rate for the product is time varying and the planning horizon is finite. The aim is the determination of production–reworking policy that minimizes the total cost of the system. By updating existing results of a similar optimization problem, conditions are found under which a unique production–reworking policy exists. This model can be extended by assuming that shortages could be allowed and also that reworking process takes place after a predetermined number of cycles.

References

1. F.W. Harris, How many parts to make at once. Fact. Mag. Manage. **10**, 135–136, 152 (1913)
2. M.J. Rosenblatt, H.L. Lee, Economic production cycles with imperfect production processes. IIE Trans. **18**, 48–55 (1986)
3. M. Hariga, M. Ben-Daya, Note: the economic manufacturing lot-sizing problem with imperfect production processes: bounds and optimal solutions. Nav. Res. Log. **45**, 423–433 (1998)
4. I. Moon, B.C. Giri, K. Choi, Economic lot scheduling problem with imperfect production processes and setup times. J. Oper. Res. Soc. **53**, 620–629 (2002)
5. C.H. Wang, Integrated production and product inspection policy for a deteriorating production system. Int. J. Prod. Econ. **95**, 123–134 (2005)
6. S.S. Sana, S.K. Goyal, K. Chaudhuri, An imperfect production process in a volume flexible inventory model. Int. J. Prod. Econ. **105**, 548–559 (2007)
7. L.E. Cardenas-Barron, Economic production quantity with rework process at a single-stage manufacturing system with planned backorders. Comput. Ind. Eng. **57**, 1105–1113 (2009)
8. P.A. Hayek, M.K. Salameh, Production lot sizing with the reworking of imperfect quality item produced. Prod. Plan. Control. **12**, 584–590 (2001)
9. Y.P. Chiu, Determining the optimal lot size for the finite production model with random defective rate, the rework process, and backlogging, Eng. Optim. **35**, 427–437 (2003)
10. A.M.M. Jamal, B.R. Sarker, S. Mondal, Optimal manufacturing batch size with rework process at a single stage production system. Comput. Ind. Eng. **47**, 77–89 (2004)
11. L.E. Cardenas–Barron, Optimal manufacturing batch size with rework in a single–stage production system–a simple derivation. Comput. Ind. Eng. **55**, 758–765 (2008)
12. P. Biswas, B. Sarker, Optimal batch quantity models for a lean production system with in-cycle rework and scrap. Int. J. Prod. Res. **46**, 6585–6610 (2008)
13. Y.S.P. Chiu, K.K. Chen, F.T. Cheng, M.F. Wu, Optimization of the finite production rate model with scrap, rework and stochastic machine breakdown. Comput. Math. Appl. **59**, 919–932 (2010)
14. C.J. Chung, G.A. Widyadana, H.M. Wee, Economic production quantity model for deteriorating inventory with random machine unavailability and shortage. Int. J. Prod. Res. **49**, 883–902 (2011)
15. H.M. Wee, W.T. Wang, P.C. Yang, A production quantity model for imperfect quality items with shortage and screening constraint. Int. J. Prod. Res. **51**, 1869–1884 (2013)
16. A.H. Tai, Economic production quantity models for deteriorating/imperfect products and service with rework. Comput. Ind. Eng. **66**, 879–888 (2013)
17. B. Pal, S.S. Sana, K. Chaudhuri, Maximising profits for an EPQ model with unreliable machine and rework of random defective items. Int. J. Syst. Sci. **44**, 582–594 (2013)
18. B. Sarkar, L.E. Cardenas–Barron, M. Sarkar, M.L. Singgih, An economic production quantity model with random defective rate, rework process and backorders for a single stage production system. J. Manuf. Syst. **33**, 423–435 (2014)
19. C.K. Sivashankari, S. Panayappan, Production inventory model with reworking of imperfect production, scrap and shortages. Int. J. Manag. Sci. Eng. Manag. **9**, 9–20 (2014)
20. E.A. Silver, D.F. Pyke, R. Peterson, *Inventory Management and Production Planning and Scheduling* (Wiley, New York, 1998
21. Z.T. Balkhi, On a finite production lot size inventory model for deteriorating items: an optimal solution. Eur. J. Oper. Res. **132**, 210–223 (2001)
22. S. Sana, S.K. Goyal, K.S. Chaudhuri, Production–inventory model for a deteriorating item with trended demand and shortages. Eur. J. Oper. Res. **157**, 357–371 (2004)
23. A. Roy, S. Kar, M. Maiti, Volume flexible production-policy for randomly deteriorating item with trended demand and shortages. Int. J. Prod. Econ. **128**, 188–199 (2010)
24. H.-L. Yang, A partial backlogging production-inventory lot-size model for deteriorating items with time-varying production and demand rate over a finite time horizon. Int. J. Syst. Sci. **42**, 1397–1407 (2011)

25. L. Benkherouf, D. Boushehri, Optimal policies for a finite-horizon production inventory model. Adv. Oper. Res. Article ID 768929, 16 p. (2012). doi:10.1155/2012/768929
26. D. Das, M.B. Kar, A. Roy, S. Kar, Two-warehouse production inventory model for a deteriorating item with time-varying demand and shortages: a genetic algorithm with varying population size approach. Optim. Eng. **15**, 889–907 (2014)
27. L. Benkherouf, K. Skouri, I. Konstantaras, Optimal control of production remanufacturing and refurbishing activities in a finite planning horizon inventory system. J. Optim. Theory Appl. **168**, 677–698 (2016)
28. L. Benkherouf, M.A. Omar, Optimal manufacturing batch size with rework for a finite-horizon and time-varying demand rates inventory model. RAIRO Oper. Res. **51**(1), 173–187 (2017)
29. T.M. Al-Khamis, L. Benkherouf, M.A. Omar, Optimal policies for a finite-horizon batching inventory model. Int. J. Syst. Sci **45**, 2196–2202 (2014)
30. S. Dharmadhikari, J.D. Kumar, Unimodality, convexity, and applications, in *Probability and Mathematical Statistics* (Academic, Boston, 1988)
31. H. Rau, B.C. Ouyang, A general and optimal approach for three inventory models with a linear trend in demand. Comput. Ind. Eng. **52**, 521–532 (2007)
32. L. Benkherouf, B.H. Gilding, On a class of optimization problems for finite time horizon models. SIAM J. Control Optim. **48**, 993–1030 (2009)

On Co-polynomials on the Real Line and the Unit Circle

Kenier Castillo, Francisco Marcellán, and Jorge Rivero

Abstract In this paper, we present an overview about algebraic and analytic aspects of orthogonal polynomials on the real line when finite modifications of the coefficients of the three-term recurrence relation they satisfy, the so-called co-polynomials on the real line, are considered. We investigate the behavior of their zeros, mainly interlacing and monotonicity properties. Furthermore, using a transfer matrix approach we obtain new structural relations, combining theoretical and computational advantages. In the case of orthogonal polynomials on the unit circle, we analyze the effects of finite modifications of Verblunsky coefficients on Szegő recurrences. More precisely, we study the structural relations and the corresponding $\mathcal{C}$-functions of the orthogonal polynomials with respect to these modifications from the initial ones. By using the Szegő's transformation we deduce new relations between the recurrence coefficients for orthogonal polynomials on the real line and the Verblunsky parameters of orthogonal polynomials on the unit circle as well as the relation between the corresponding $\mathcal{S}$-functions and $\mathcal{C}$-functions is studied.

Keywords Orthogonal polynomials on the real line • Orthogonal polynomials on the unit circle • Zeros • Spectral transformations • Co-polynomials • Szegő transformation

Mathematics Subject Classification (2010): 42C05

K. Castillo
CMUC, Department of Mathematics, University of Coimbra, 3001-501 Coimbra, Portugal
e-mail: kenier@mat.uc.pt

F. Marcellán
Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain
e-mail: pacomarc@ing.uc3m.es

J. Rivero (✉)
Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain

Instituto de Ciencias Matemáticas (ICMAT) Campus de Cantoblanco, UAM, 28049 Madrid, Spain
e-mail: riverodones@gmail.com

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_4

Introduction

Orthogonal Polynomials on the Real Line and Spectral Transformations

Let $d\mu$ be a nontrivial probability measure supported on $I \subseteq \mathbb{R}$. The sequence of polynomials $\{p_n(x)\}_{n\geqslant 0}$ where

$$p_n(x) = \gamma_n x^n + \delta_n x^{n-1} + \text{(lower degree terms)}, \quad \gamma_n > 0,$$

is said to be an orthonormal polynomial sequence with respect to $d\mu$ if

$$\int_I p_n(x) p_m d\mu(x) = \delta_{n,m}, \quad m \geq 0. \tag{1}$$

The corresponding monic orthogonal polynomials (with leading coefficient equal to 1) are $P_n(x) = p_n(x)/\gamma_n$, see [8, 32, 36]. These polynomials satisfy the following three-term recurrence relation

$$P_{n+1}(x) = (x - b_{n+1})P_n(x) - d_n P_{n-1}(x), \quad d_n \neq 0, \quad d_0 = 1, \quad n \geqslant 0, \tag{2}$$

where the recurrence coefficients are given by

$$b_n = \frac{\delta_n}{\gamma_n} - \frac{\delta_{n+1}}{\gamma_{n+1}}, \quad d_n = a_n^2, \quad a_n = \frac{\gamma_{n-1}}{\gamma_n} > 0, \qquad n \geq 1.$$

Notice that the initial conditions $P_{-1}(x) = 0$ and $P_0(x) = 1$ hold.

The converse of this result is the so-called Favard's theorem in the theory of *orthogonal polynomials on the real line* (OPRL, in short) [8]. In other words, given a sequence of monic polynomials, $\{P_n(x)\}_{n\geq 0}$, generated by (2) with recurrence coefficients $\{a_n : a_n \in \mathbb{R} \wedge a_n > 0\}_{n\geq 1}$ and $\{b_n : b_n \in \mathbb{R}\}_{n\geq 1}$, then there exists a nontrivial probability measure $d\mu$ supported on the real line so that the orthogonality conditions (1) hold. Moreover, if $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$ are bounded sequences, then $d\mu$ is unique. From now on, we will assume that the recurrence coefficients always satisfy the hypothesis of Favard's theorem.

It is very well known that the zeros of $P_n(x)$, $\{x_{n,k}\}_{k=1}^n$, are real, simple, and are located in the interior of the convex hull of the support I of the measure $d\mu$ and the zeros of P_n and P_{n+1} strictly interlace. The notation for zeros is

$$x_{n,n} < x_{n,n-1} < \cdots < x_{n,2} < x_{n,1}.$$

We suggest the reader to consult [8, 17, 32, 36], where a complete presentation of the classical theory of OPRL can be found.

From the sequence of monic orthogonal polynomials $\{P_n(x)\}_{n\geqslant 0}$ we can define the sequence of associated monic polynomials of order k [13], $\{P_n^{(k)}(x)\}_{n\geqslant 0}$, $k \geqslant 1$, by means of the shifted recurrence relation

$$P_{n+1}^{(k)}(x) = (x - b_{n+k+1})P_n^{(k)}(x) - d_{n+k}P_{n-1}^{(k)}(x), \quad n \geqslant 0,$$

with $P_{-1}^{(k)}(x) = 0$ and $P_0^{(k)}(x) = 1$. The three-term recurrence relation (2) is often represented in matrix form

$$x\mathbf{P}(x) = \mathbf{J}\mathbf{P}(x), \qquad \mathbf{P} = [P_0, P_1, \dots]^T,$$

where $\mathbf{J}$ is a semi-infinite tridiagonal matrix

$$\mathbf{J} = \begin{bmatrix} b_1 & 1 & & & \\ d_1 & b_2 & 1 & & \\ & d_2 & b_3 & 1 & \\ & & d_3 & b_4 & \ddots \\ & & & \ddots & \ddots \end{bmatrix},$$

which is called the monic Jacobi matrix [8, 19]. A useful property of the matrix $\mathbf{J}$ is that the eigenvalues of its $n \times n$ leading principal submatrices $\mathbf{J}_n$ are the zeros of the polynomial $P_n(x)$. Indeed, $P_n(x)$ is the characteristic polynomial of $\mathbf{J}_n$,

$$P_n(x) = \det(x\mathbf{I}_n - \mathbf{J}_n),$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The Stieltjes or Cauchy transformation of the orthogonality measure $d\mu$ is defined by

$$S_\mu(x) = \int_I \frac{d\mu(y)}{x-y}, \quad x \in \mathbb{C} \setminus I.$$

It has a particular interest in the OPRL theory. $S_\mu(x)$ admits the following series expansion:

$$S_\mu(x) = \sum_{k=0}^{\infty} \frac{u_k}{x^{k+1}},$$

where u_k are the moments associated with $d\mu$, i.e.,

$$u_k = \int_I x^k d\mu(x), \quad k \geq 0.$$

By a spectral transformation of the $\mathcal{S}$-function $S_\mu(x)$, we mean a new $\mathcal{S}$-function associated with a measure $d\widetilde{\mu}$, whose moments are a modification of the moments of the original measure $d\mu$. We refer to pure rational spectral transformation as a transformation of $S_\mu(x)$ given by

$$S_\sigma(x) \doteq \boldsymbol{A}(x) S_\mu(x), \quad \boldsymbol{A}(x) = \begin{bmatrix} a(x) & b(x) \\ c(x) & d(x) \end{bmatrix}, \tag{3}$$

where $a(x)$, $b(x)$, $c(x)$, and $d(x)$ are non-zero polynomials that provide a "true" asymptotic behavior to (3) (see [37]). In (3), we adopt the notation $\doteq$ introduced in [32], i.e., for the homography mapping

$$f(x) = \frac{a(x)g(x) + b(x)}{c(x)g(x) + d(x)}, \quad a(x)d(x) - b(x)c(x) \not\equiv 0,$$

we will write

$$f(x) \doteq \boldsymbol{A}(x) g(x).$$

Polynomials orthogonal with respect to a measure $d\mu$ play a central role in Harmonic Analysis (Fourier expansions of functions in the space $L^2(d\mu)$, rational approximation to functions defined by the Cauchy transformation of the measure $d\mu$ (the denominators of the diagonal Padé approximants are the corresponding orthogonal polynomials and the numerators are the associated polynomials of order 1), quadrature rules as well as their extensions, polynomial inequalities, and their applications (see [2, 21, 25], and the references therein, among others).

Orthogonal Polynomials on the Unit Circle and Spectral Transformations

Let $d\sigma$ be a nontrivial probability measure supported on the unit circle $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ parametrized by $z = e^{i\theta}$. There exists a unique sequence $\{\phi_n(z)\}_{n\geqslant 0}$ of orthonormal polynomials

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

such that

$$\int_{-\pi}^{\pi} \phi_n(e^{i\theta})\overline{\phi_m(e^{i\theta})} d\sigma(\theta) = \delta_{n,m}, \quad m \geq 0.$$

The corresponding monic polynomials are defined by $\Phi_n(z) = \phi_n(z)/\kappa_n$ and they are well known as *orthogonal polynomials on the unit circle* (OPUC, in short). These polynomials satisfy the following recurrence relations (see [14, 32, 36]):

$$\Phi_{n+1}(z) = z\Phi_n(z) - \overline{\alpha}_n \Phi_n^*(z), \quad n \geq 0, \tag{4}$$

$$\Phi_{n+1}^*(z) = \Phi_n^*(z) - \alpha_n z \Phi_n(z), \quad n \geq 0, \tag{5}$$

with initial condition $\Phi_0(z) = 1$. The polynomial $\Phi_n^*(z) = z^n \overline{\Phi}_n(z^{-1})$ is the so-called reversed polynomial and the complex numbers $\{\alpha_n\}_{n\geq 0}$ where $\alpha_n = -\overline{\Phi_{n+1}(0)}$, are known as Verblunsky, Schur, Geronimus, or reflection parameters. Let notice that $|\alpha_n| < 1$.

In this context, there is an analogous of the Favard's theorem, called in the contemporary literature as Verblunsky's theorem [32, 34].

Theorem 1 ([32, 34], Verblunsky Theorem) *Let $\{\zeta_n\}_{n\geq 0}$ be a sequence of complex numbers in $\mathbb{D}$. Then, there is a unique nontrivial probability measure $d\sigma(z)$ supported on the unit circle such that $\alpha_n = \zeta_n$.*

Based on the above theorem, the OPUC are completely determined by their Verblunsky coefficients. This fact reaffirms the need to study orthogonal polynomials associated with modifications of the original Verblunsky coefficients. In our opinion, one of the most interesting results in this direction appeared in [27]. In this work, Peherstorfer introduced and studied the so-called associated polynomials on the unit circle. Indeed, for a fixed positive integer number r, the associated polynomials of order r, $\left\{\Phi_n^{[r]}(z)\right\}_{n\geq 0}$, are generated by the shifted Verblunsky coefficients $\{\alpha_{n+r}\}_{n\geq 0}$ through the Szegő recurrences with initial condition $\Phi_0^{[r]}(z) := 1$.

If we replace in (4) the sequence $\{\alpha_n\}_{n\geq 0}$ by $\{-\alpha_n\}_{n\geq 0}$, then we obtain the sequence of second kind polynomials $\{\Omega_n(z)\}_{n\geqslant 0}$, that is a sequence of OPUC according to Verblunsky Theorem. These polynomials can be expressed in terms of the monic orthogonal polynomials with respect to the measure $d\sigma(z)$ as follows:

$$\Omega_n(z) = \int_{\partial\mathbb{D}} \frac{y+z}{y-z} \left(\Phi_n(y) - \Phi_n(z)\right) d\sigma(y), \quad n \geq 1.$$

Notice that $\deg \Omega_n(z) = n$ and $\Omega_n(z)$ is monic.

The Riesz-Herglotz transform of the measure $d\sigma$ is given by

$$F_\sigma(z) = \int_{-\pi}^{\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z}\, d\sigma(\theta).$$

Since $F_\sigma(0) = 1$ and $\Re F_\sigma(z) > 0$ on the unit open disc $\mathbb{D} = \{z \in \mathbb{C} : |z| < 1\}$, $F_\sigma(z)$ is called a Carathéodory function [15], or, simply, $\mathcal{C}$-function. Let c_k be the k-th moment associated with the measure $d\sigma$, i.e.,

$$c_k = \int_{-\pi}^{\pi} e^{-ik\theta} d\sigma(\theta).$$

$F_\sigma(z)$ can be written in terms of the moments $\{c_n\}_{n\geqslant 0}$ as follows:

$$F(z) = 1 + 2\sum_{k=1}^{\infty} c_k z^k.$$

As for the real line case, by a spectral transformation of a $\mathcal{C}$-function $F_\sigma(z)$ we mean a new $\mathcal{C}$-function associated with a measure $d\psi$, a modification of the original measure $d\sigma$. We refer to pure rational spectral transformation as a transformation of $F_\sigma(z)$ given by

$$F_\psi(z) \doteq \boldsymbol{E}(z)F_\sigma(z), \quad \boldsymbol{E}(z) = \begin{bmatrix} A(z) & B(z) \\ C(z) & D(z) \end{bmatrix}, \tag{6}$$

where $A(z)$, $B(z)$, $C(z)$, and $D(z)$ are non-zero polynomials that provide a "true" behavior to (6) around the origin (see [3]), i.e., $A(z)D(z) - C(z)B(z) \not\equiv 0$.

Szegő Transformation and Geronimus Relations

Let us assume that the measure $d\mu$ is supported on the interval $[-1, 1]$. Let introduce a measure supported on the unit circle $d\sigma$ such that

$$d\sigma(\theta) = \frac{1}{2}|d\mu(\cos\theta)|.$$

In particular, if $d\mu$ is an absolutely continuous measure, i.e., $d\mu(x) = \omega(x)dx$, we have

$$d\sigma(\theta) = \frac{1}{2}\omega(\cos\theta)|\sin\theta|d\theta.$$

This is the so-called Szegő transformation of probability measures supported on $[-1, 1]$ to probability measures supported on $\mathbb{T}$ (see [14, 15, 33, 36]). We write the relation between $d\mu$ and $d\sigma$ through the Szegő transformation as $\sigma = \mathrm{Sz}(\mu)$. Of course, under the previous considerations, we get

$$\alpha_n \in (-1, 1), \quad n \geqslant 0.$$

There is a relation between the OPRL associated with a measure $d\mu$ supported on $[-1, 1]$ and the OPUC associated with the measure $\sigma = \mathrm{Sz}(\mu)$ supported on $\mathbb{T}$, [33]

$$p_n(x) = \frac{\kappa_{2n}}{\sqrt{2(1-\alpha_{2n-1})}}\left(z^{-n}\Phi_{2n}(z) + z^n\Phi_{2n}(1/z)\right). \tag{7}$$

From (7) one can obtain a relation between the coefficients of the corresponding recurrence relations, see [33],

$$d_{n+1} = \frac{1}{4}\left(1-\alpha_{2n-1}\right)\left(1-\alpha_{2n}^2\right)\left(1+\alpha_{2n+1}\right), \quad n \geq 0, \tag{8}$$

$$b_{n+1} = \frac{1}{2}[\alpha_{2n}\left(1-\alpha_{2n-1}\right) - \alpha_{2n-2}\left(1+\alpha_{2n-1}\right)], \quad n \geq 0, \tag{9}$$

with the convention $\alpha_{-1} = -1$. Notice that $b_n \equiv 0,\ n \geq 1$, if and only if $\alpha_{2n} = 0,\ n \geq 0$.

There is also a relation between the $\mathcal{S}$-function and the $\mathcal{C}$-function associated with $d\mu$ and $d\sigma$, respectively, as follows:

$$F(z) = \frac{1-z^2}{2z} S(x),$$

or, equivalently,

$$S(x) = \frac{F(z)}{\sqrt{x^2-1}},$$

with $2x = z + z^{-1}$ and $z = x - \sqrt{x^2-1}$.

The aim of this paper is to present an overview about COPRL, COPUC, and its connection through the Szegő transformation. The structure of the paper is as follows. In section “Co-Polynomials on the Real Line”, we present an overview about algebraic and analytic aspects, structural relations, and spectral transformations of COPRL proposed in [5]. In section “Co-polynomials on the Unit Circle”, we obtain a new structural relation based on a transfer matrix approach proposed in [3] for similar perturbations in the OPUC theory. In section “Szegő Transformation and Co-polynomials”, we explore the relation between the co-polynomials on the real line (resp. on the unit circle), and the corresponding sequences of monic orthogonal polynomials obtained on the unit circle (resp. on the real line) via Szegő transformation.

Co-Polynomials on the Real Line

Let $\{g_n(x)\}_{n\geqslant 0}$ be a sequence of orthogonal polynomials satisfying the three term-recurrence relation for OPRL with new recurrence coefficients, $\{\mathfrak{b}_n\}_{n\geq 1}$ and $\{\mathfrak{d}_n\}_{n\geq 1}$, i.e.,

$$g_{n+1}(x) = (x - \mathfrak{b}_{n+1})g_n(x) - \mathfrak{d}_n g_{n-1}(x),$$

with initial conditions $g_{-1}(x) = 0$ and $g_0(x) = 1$, perturbed in a (generalized) co-dilated and/or co-recursive way, namely *co-polynomials on the real line* (COPRL). In other words, we will consider an arbitrary single modification of the recurrence coefficients as follows:

$$\mathfrak{d}_n = \lambda_k^{\delta_{n,k}} d_n, \qquad \lambda_k > 0, \qquad\qquad \text{(co-dilated case)} \tag{10}$$

$$\mathfrak{b}_n = b_n + \tau_{k+1}\delta_{n,k+1}, \quad \tau_{k+1} \in \mathbb{R}. \qquad\qquad \text{(co-recursive case)} \tag{11}$$

where k is a fixed non-negative integer number, and δ_{nk} is the Kronecker delta.

The study of the algebraic and analytic properties of the COPRL and its applications was initiated by Chihara [7] and later continued by several authors. Among others, the contributions of Marcellán [23, 30], Maroni [10, 31], Ronveaux [29, 31], and Peherstorfer [26] are remarkable. For some applications, see [11, 12, 20, 35].

In the next section, we study some inequalities for zeros of COPRL. The original contributions are contained in [5] following some ideas developed in [4].

Zeros and Inequalities

It is very well known that the orthonormal version of (2), for recurrence coefficients depending on a parameter ϵ, can be written in an operator form by using a symmetric Jacobi matrix, $\boldsymbol{J}(\epsilon)$,

$$\mathbf{J}(\epsilon) = \begin{bmatrix} b_1 & a_1 & & & \\ a_1 & b_2 & a_2 & & \\ & a_2 & b_3 & a_3 & \\ & & \ddots & \ddots & \ddots \end{bmatrix},$$

where $a_n^2 = d_n$ (for simplicity, we omit here the dependence of ϵ). In a matrix form,

$$x\boldsymbol{p} = \boldsymbol{J}(\epsilon)\boldsymbol{p},$$

where $p_n(x) = \gamma_n^{-1/2} P_n(x)$ and $\boldsymbol{p} = [p_0(x), p_1(x), \dots]^T$. According to a version of Hellmann-Feynman's theorem [17, Sect. 7.3], if $\partial \boldsymbol{J}_n(\epsilon)/\partial\epsilon$ is strictly positive (resp. negative) definite, then the zeros of the corresponding OPRL are strictly increasing (resp. decreasing) functions of ϵ. But for some cases related with COPRL, we can obtain more information on the behavior of zeros following a different approach recently proposed in [3].

In [23], using the theory of difference equations, the authors deduced the explicit expression of the COPRL associated with the perturbation (10) and/or (11) in terms of the initial OPRL and their associated polynomials of order k. However, in the previous research the questions related to the domain of validity of the connection formulas between COPRL and OPRL were omitted.

Let us define

$$D(u_n, v_n) := \begin{vmatrix} u_n & v_n \\ u_{n+1} & v_{n+1} \end{vmatrix}, \tag{12}$$

the Casorati determinant associated with two arbitrary sequences $\{u_n\}_{n\geq 1}$ and $\{v_n\}_{n\geq 1}$. From the theory of linear difference equations, we know that if the Casorati determinant is different from zero for every n, then these two sequences are said to be linearly independent [24]. Notice that $\{P^{(k)}_{n-k}(x)\}_{n\geq 0}$, is a solution of the recurrence relation (2). It is easy to verify that

$$\begin{aligned} D(P_n(x), P^{(k)}_{n-k}(x)) &= d_n \left(P_{n-1}(x)P^{(k)}_{n-k}(x) - P^{(k)}_{n-k-1}(x)P_n(x) \right), \\ &= d_n D(P_{n-1}(x), P^{(k)}_{n-k-1}(x)). \end{aligned}$$

Let X denote the set of zeros of $P_{k-1}(x)$. From the above equalities, we get

$$D(P_n(x), P^{(k)}_{n-k}(x)) = \left(\prod_{j=k}^{n} d_j \right) P_{k-1}(x), \tag{13}$$

which means that $P_n(x)$ and $P^{(k)}_{n-k}(x)$ with $n > k$, are linearly independent in $\mathbb{C} \setminus X$. If we denote by $\{P_n(x; \lambda_m, \tau_{m+1}; \ \ldots ; \lambda_k, \tau_{k+1})\}_{n\geq 0}$ the COPRL associated with the finite composition of perturbations (10) and (11) from order m to order k, $m \leq k$ then, after elementary calculations, for $m = k$ we have

Theorem 2 ([23]) *For $x \in \mathbb{C} \setminus X$ the following formulas hold:*

$$\begin{aligned} P_n(x; \lambda_k, \tau_{k+1}) &= P_n(x), & n \leq k, \\ P_n(x; \lambda_k, \tau_{k+1}) &= P_n(x) - Q_k(x)P^{(k)}_{n-k}(x), & n > k, \end{aligned}$$

where $Q_k(x) = \tau_{k+1}P_k(x) + d_k(\lambda_k - 1)P_{k-1}(x)$.

As a consequence of the last result, we get

Corollary 1 ([5]) *$P_n(x; \lambda_k, \tau_{k+1})$ and $P_n(x)$ share at most the zeros of $Q_k(x)$ and $P_{k-1}(x)$.*

From the interlacing property of two consecutive OPRL, we can easily deduce that $Q_k(x)$, $P_k(x)$, and $P_{k-1}(x)$ are coprime. But we can go a step further.

Proposition 1 ([5]) *Let assume $\lambda_k \neq 1$ and $\tau_{k+1} \neq 0$ and define $c := (\lambda_k - 1)/\tau_{k+1}$. Let $\{y_{k,j}(c)\}_{j=1}^{k}$ be the zeros of $Q_k(x)$. The following statements hold:*

i) If $c > 0$*, then*

$$x_{k-1,j-1} < y_{k,j}(c) < x_{k,j}; \quad x_{k-1,0} := -\infty.$$

Moreover, $y_{k,j}(c)$ *(for a fixed value of* j*) is a strictly increasing (resp. decreasing) function of* λ_k *(resp.* τ_{k+1}*).*

ii) If $c < 0$*, then*

$$x_{k,j} < y_{k,j}(c) < x_{k-1,j}; \quad x_{k-1,k} := \infty.$$

Also, $y_{k,j}(c)$ *(for a fixed value of* j*) is a strictly decreasing (resp. increasing) function of* λ_k *(resp.* τ_{k+1}*).*

Furthermore,

$$\lim_{\lambda_k \to 1} y_{k,j}(c) = x_{k-1,j}, \qquad \lim_{\tau_{k+1} \to \infty} y_{k,j}(c) = x_{k-1,j}.$$

The location of the extreme zeros with respect to the orthogonality interval I can be given by using [36, Theorem 3.3.4].

The next theorem has direct consequences in the interlacing and monotonicity of zeros of COPRL.

Theorem 3 ([5]) *Let* $x_{n,j+1}$ *and* $x_{n,j}$ *be two consecutive zeros of* $P_n(x)$*, then the following holds. If there are no zeros of* $Q_k(x)P_k(x)$ *in* $I_j = (x_{n,j+1}, x_{n,j})$ *that are not common with the zeros* $P_n(x; \lambda_k, \tau_{k+1})$*, then the interval* I_j *contains at most an odd number of zeros of* $P_n(x; \lambda_k, \tau_{k+1})$*. Moreover, if there are zeros of* $Q_k(x)P_k(x)$ *in* I_j *that are not common to the zeros of* $P_n(x; \lambda_k, \tau_{k+1})$*, then the interval* I_j *contains at most an even number of zeros of* $P_n(x; \lambda_k, \tau_{k+1})$*.*

Note that the previous result contains as a particular case the interlacing property obtained in [7]. For the co-recursive case, that is, $\lambda_k := 1$, we have the following interlacing property.

Corollary 2 ([5]) *Let* $l < k$ *be the number of no common zeros between* $P_n(x; 1, \tau_{k+1})$ *and* $P_n(x)$*. Denote by* $\{y_{n,j}(1, \tau_{k+1})\}_{j=1}^{l}$ *and* $\{y_{n,j}\}_{j=1}^{l}$*, these zeros. If* $\tau_{k+1} < 0$*, then*

$$y_{n,n}(1, \tau_{k+1}) < y_{n,l} < y_{n,l-1}(1, \tau_{k+1}) < y_{n,l-1} < \cdots < y_{n,1}(1, \tau_{k+1}) < y_{n,1}, \qquad (14)$$

where the role of the zeros $\{y_{n,j}(1, \tau_{k+1})\}_{j=1}^{l}$ *and* $\{y_{n,j}\}_{j=1}^{l}$ *is reversed when* $\tau_{k+1} > 0$*.*

Corollary 3 ([5]) *The zeros of the polynomial* $P_n(x; 1, \tau_{k+1}; 1, \tau_{k+2})$ *(for a fixed value of* k *and* $n > k$*) are strictly increasing functions of* τ_{k+1} *and* τ_{k+2}*.*

The previous results for the co-recursive case reduce and give more information than Hellmann-Feynman's theorem. Notice that the existence of cases for which

$\det(\partial \boldsymbol{J}_n(\epsilon)/\partial\epsilon) = 0$, mentioned in the beginning of the section, could imply strictly monotonicity of zeros. We recall that Corollary 3 was deduced in [4] from the perturbation theory for symmetric matrices.

Example 1 It is well known [8] that the monic Jacobi polynomials $\{P_n^{(\alpha,\beta)}(x)\}_{n\geq 0}$ satisfy for any real value of α and β, the recurrence relation (2) where

$$a_n^{(\alpha,\beta)} = \frac{4n(n+\alpha)(n+\beta)(n+\alpha+\beta)}{(2n+\alpha+\beta-1)(2n+\alpha+\beta+1)(2n+\alpha+\beta)^2},$$

$$b_{n+1}^{(\alpha,\beta)} = \frac{\beta^2-\alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)}.$$

Furthermore, if $\alpha, \beta > -1$ the polynomials are orthogonal with respect to the weight $(1-x)^\alpha(1+x)^\beta$ on the interval $[-1, 1]$. In order to illustrate Corollary 3, we consider a new sequence of Jacobi polynomials associated with two consecutive modification (11). Figure 1 is obtained by using Wolfram Mathematica® 9.0[1] with the aid of the function **JacobiP**[**n**, α, β, **x**] and the recurrence relation (2), and shows the polynomials $P_6^{(2,1)}(x)$ (continuous line), $P_6^{(2,1)}(x; 1, 0.1; 1, 0.2)$ (large-dashed line), and $P_6^{(2,1)}(x; 1, 0.25; 1, 0.28)$ (small-dashed line). Observe that the zeros behave in accordance with our result. In other words, the monotonicity is "strictly" and it is not something that can be guaranteed by Hellmann-Feynman's theorem.

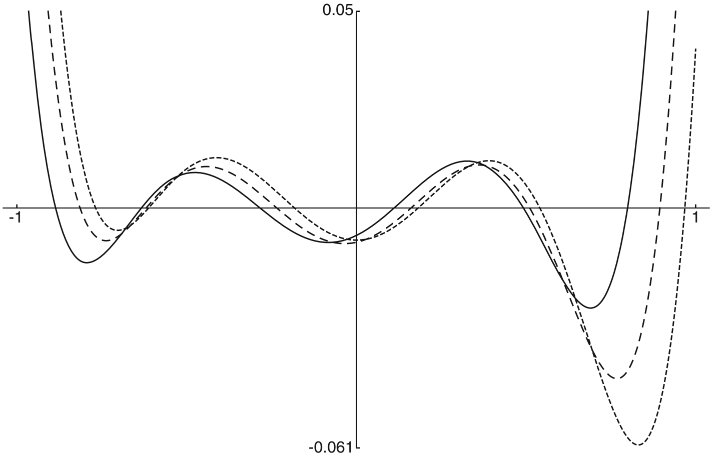


Fig. 1 Graphs of $P_6^{(2,1)}(x)$, $P_6^{(2,1)}(x; 1, 0.1; 1, 0.2)$, and $P_6^{(2,1)}(x; 1, 0.25; 1, 0.28)$

[1] Wolfram Mathematica is a registered trademark of Wolfram Research, Inc.

Theorem 4 ([5]) *With the notation of Proposition 1, let us define* $y_1 := \max\{x_{k,1},\ y_{k,1}(c)\}$. *Let denote by* $\{x_{n,j}(\lambda_k, \tau_{k+1})\}_{j=1}^{n}$ *the zeros of the polynomial* $P_n(x; \lambda_k,\ \tau_{k+1})$. *If* $c > 0$, *then*

$$x_{n,l} < x_{n,l}(\lambda_k, \tau_{k+1}),$$

for all the zeros of $P_n(x; \lambda_k, \tau_{k+1})$ *and* $P_n(x)$ *in* $\mathbb{R} \setminus [-\infty, y_1]$, *where the role of the zeros* $x_{n,l}$ *and* $x_{n,l}(\lambda_k, \tau_{k+1})$ *is reversed when* $c < 0$.

The usual tool dealing with the inequalities concerning the largest (or the smallest zero) of OPRL is the Perron-Frobenius Theorem [17, Theorem 7.4.1]. Notice that the previous result gives more information.

Example 2 The monic Laguerre polynomials $\{L_n^{(\alpha)}(x)\}_{n\geq 0}$ satisfy, for any real value of α, the recurrence relation (2) where

$$a_n^{(\alpha)} = n(n+\alpha),$$
$$b_{n+1}^{(\alpha)} = 2n + 1 + \alpha.$$

Furthermore, if $\alpha > -1$ the polynomials are orthogonal with respect to the weight $x^{\alpha}e^{-x}$ on the interval $[0, \infty)$. In order to illustrate Theorem 4, we consider a new sequence of Laguerre polynomials associated with the modifications (10) and (11). Figure 2 is obtained by using Wolfram Mathematica® 9.0 with the aid of the function **LaguerreL[n**, α, **x]** and the recurrence relation (2), and shows the polynomials $L_5^{(2)}(x)$ (continuous line) and $L_5^{(2)}(x; 1.2, 3)$ (dashed line). Observe that for this case with $c = 0.0\bar{6}$, all the zeros greater than $y_1 = 1.2268$ behave in accordance with Theorem 4. Notice that, the Perron-Frobenius Theorem can guarantee this result only for the largest zero.

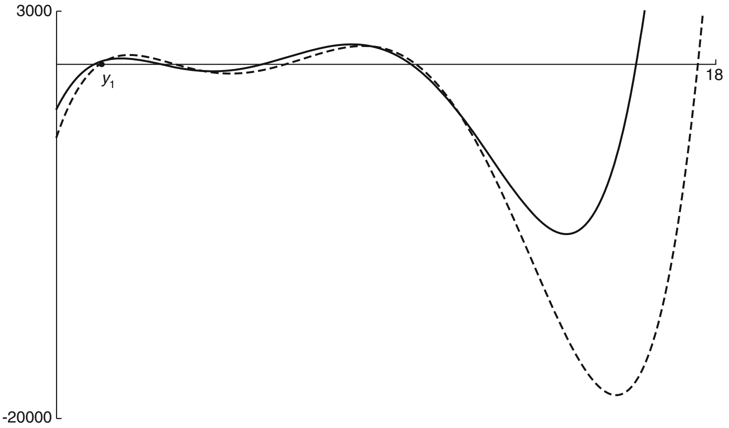


Fig. 2 Graphs of $L_5^{(2)}(x)$ and $L_5^{(2)}(x; 1.2, 3)$

A Transfer Matrix Approach

Theorem 2 has been successfully used in the study of zeros of COPRL but it has two main constraints. First, the structural relation is not useful if we are interested in the finite composition of perturbations, mainly from a computational point of view. Second, the structural relation is not valid on the whole complex plane. The aim of this section is to use a transfer matrix approach to avoid these constraints.

Set

$$\mathbf{P}_{n+1}(x) := \left[P_{n+1}(x), P_n(x)\right]^T, \quad \mathbf{A}_n := \begin{bmatrix} x - b_{n+1} & -d_n \\ 1 & 0 \end{bmatrix}.$$

From (2), we get

$$\mathbf{P}_{n+1}(x) = \mathbf{A}_n\, \mathbf{P}_n(x), \quad \mathbf{P}_0(x) := \left[P_0(x), P_{-1}(x)\right]^T,$$

as well as

$$\mathbf{P}_{n+1}(x) = (\mathbf{A}_n \cdots \mathbf{A}_0)\, \mathbf{P}_0(x), \tag{15}$$

As previously, we have

$$\mathbf{P}_{n+1}(x; \lambda_k, \tau_{k+1}) = (\mathbf{A}_n \cdots \mathbf{A}_{k+1})\, \mathbf{A}_k(\lambda_k, \tau_{k+1})\, (\mathbf{A}_{k-1} \cdots \mathbf{A}_0)\, \mathbf{P}_0(x), \tag{16}$$

where

$$\mathbf{A}_k(\lambda_k, \tau_{k+1}) = \begin{bmatrix} x - b_{k+1} - \tau_{k+1} & -\lambda_k d_k \\ 1 & 0 \end{bmatrix}.$$

Combining (15) and (16), we can deduce that the following formula holds on $\mathbb{C}$

$$\mathbf{P}_{n+1}(x; \lambda_k, \tau_k) = (\mathbf{A}_n \cdots \mathbf{A}_{k+1})\, \mathbf{A}_k(\lambda_k, \tau_{k+1}) \mathbf{A}_k^{-1}\, (\mathbf{A}_n \cdots \mathbf{A}_{k+1})^{-1}\, \mathbf{P}_{n+1}(x). \tag{17}$$

The previous equation has some computational advantage as compared to Theorem 2 and it holds in $\mathbb{C}$. But we can improve this result by using an auxiliary sequence of polynomials.

Consider the associated polynomials of order $k = 1$ (also called either first kind associated polynomials, or numerator polynomials) $\{r_n(x)\}_{n \geq 0}$, which are the unique solution of the recurrence relation

$$x r_n(x) = a_{n+1} r_{n+1}(x) + b_{n+1} r_n(x) + a_n r_{n-1}(x), \quad a_n^2 = d_n, \quad n \geqslant 0, \tag{18}$$

with initial conditions $r_{-1}(x) := -1$ and $r_0(x) := 0$ or, equivalently, $r_0(x) := 0$ and $r_1(x) := 1/a_1$. Note that $r_n(x)$ is a polynomial of degree $n - 1$. We define $R_n(x) := \gamma_n^{-1} r_n(x) = P_{n-1}^{(1)}(x)$ which is a monic polynomial.

Theorem 5 (**[5]**) *The following formulas hold in* $\mathbb{C}$*:*

$$\left(\prod_{j=1}^{k} d_j\right)\begin{bmatrix} P_{n+1}(x;\lambda_k,\tau_{k+1}) \\ -R_{n+1}(x;\lambda_k,\tau_{k+1}) \end{bmatrix} = \mathbf{M}_k \begin{bmatrix} P_{n+1}(x) \\ -R_{n+1}(x) \end{bmatrix}, \quad n > k,$$

where $\mathbf{M}_k$ *is*

$$\mathbf{M}_k = \begin{bmatrix} \prod_{j=1}^{k} d_j + Q_k(x)R_k(x) & Q_k(x)P_k(x) \\ \widehat{R}_k(x)R_k(x) & \prod_{j=1}^{k} d_j + \widehat{R}_k(x)P_k(x) \end{bmatrix},$$

with $\widehat{R}_k(x) = -\tau_{k+1}R_k(x) - (\lambda_k - 1)d_kR_{k-1}(x)$.

Next we give a relation between the COPRL associated with two modifications of different levels.

Corollary 4 (**[5]**) *Let* k, m *be two fixed non-negative integer numbers with* $m < k$. *Then, the following relation holds:*

$$\left(\prod_{j=m+1}^{k} d_j\right)\begin{bmatrix} P_{n+1}(x;\lambda_k,\tau_{k+1}) \\ -R_{n+1}(x;\lambda_k,\tau_{k+1}) \end{bmatrix} = \mathbf{M}_k\mathbf{M}_m^{-1} \begin{bmatrix} P_{n+1}(x;\lambda_m,\tau_{m+1}) \\ -R_{n+1}(x;\lambda_m,\tau_{m+1}) \end{bmatrix}, \quad n > k.$$

For a finite composition of perturbations we have the following result:

Theorem 6 (**[5]**) . *For* $0 < m \le k < \infty$ *and for* $n > m$ *the following relation holds:*

$$\left(\prod_{j=m}^{k}\prod_{l=0}^{j} d_l\right)\begin{bmatrix} P_{n+1}(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1}) \\ -R_{n+1}(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1}) \end{bmatrix} = \left(\prod_{j=m}^{k}\mathbf{M}_j\right)\begin{bmatrix} P_{n+1}(x) \\ -R_{n+1}(x) \end{bmatrix}.$$

Spectral Transformations of COPRL

Following Chihara [8], let us consider

$$S(x) = \sum_{n>0} \frac{u_n}{x^{n+1}} = \left[x - b_1 - \frac{d_1}{x - b_2 - \ldots}\right]^{-1} \tag{19}$$

and

$$S^{k+1}(x) = \left[x - b_{k+2} - \frac{d_{k+2}}{x - b_{k+3} - \ldots}\right]^{-1}, \tag{20}$$

which are the Stieltjes function associated with our initial OPRL and the Stieltjes function for the associated polynomial sequence of order $k+1$, where u_n is the sequence of the moments for the regular linear functional $\mathcal{U}$ whose monic orthogonal polynomial sequence satisfies (2).

A rational spectral transformation of the $\mathcal{S}$-function S is a new $\mathcal{S}$-function $S_R(x)$ defined by

$$S_R(x) \doteq \boldsymbol{A}\ S(x) \tag{21}$$

where $a(x)$, $b(x)$, $c(x)$, and $d(x)$ are coprime polynomials that provide a true asymptotic behavior to (21) around infinity, see [37].

Theorem 7 ([5]) *Let $S(x;\lambda_k,\tau_{k+1})$ be the $\mathcal{S}$-function associated with the perturbations* (10) *and* (11)*. Then*

$$S(x;\lambda_k,\tau_{k+1}) \doteq \begin{pmatrix} d_{k+1}R_k(x) & -R_{k+1}(x) - \hat{R}_k(x) \\ d_{k+1}P_k(x) & -P_{k+1}(x) + Q_k(x) \end{pmatrix} S^{k+1}(x). \tag{22}$$

As an immediate corollary, if we take $\tau_{k+1} = 0$ and $\lambda_k = 1$, then

$$d_{k+1}S^{k+1}(x) \doteq \begin{pmatrix} P_{k+1}(x) & -R_{k+1}(x) \\ P_k(x) & -R_k(x) \end{pmatrix} S(x). \tag{23}$$

Theorem 8 ([5]) *$S(x;\lambda_k,\tau_{k+1})$ is a pure rational spectral transformation of $S(x)$, given by*

$$S(x;\lambda_k,\tau_{k+1}) \doteq cof(\mathbf{M}_k)S(x), \tag{24}$$

where $cof(.)$ is the cofactor matrix operator.

An equivalent result was obtained in [23].

Corollary 5 ([5]) *$S(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1})$ is a pure rational spectral transformation of $S(x)$ given by*

$$S(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1}) \doteq cof\left(\prod_{j=m}^{k}\mathbf{M}_j\right) S(x). \tag{25}$$

Additional results related with spectral transformations of COPRL can be found in [23, 37].

Co-polynomials on the Unit Circle

For a fixed non-negative integer number k, let us consider the perturbed Verblunsky coefficients $\{\beta_n\}_{n\geq 0}$ given by

$$\beta_n = \eta_k \delta_{nk} + (1-\delta_{nk})\alpha_n. \qquad (k\text{–modification}) \tag{26}$$

where η_k is an arbitrary complex number. In order to achieve a new sequence of Verblunsky coefficients, from now on we assume that $|\eta_k| < 1$ with $\eta_k \neq \alpha_k$. We define a sequence of monic *co-polynomials on the unit circle* (COPUC, in short), $\{\Phi_n(z;k)\}_{n\geq 0}$, those polynomials generated using $\{\beta_n\}_{n\geq 0}$ through the Szegő recurrences. Analogously, we denote by $\{\Omega_n(z;k)\}_{n\geq 0}$ the corresponding sequence of polynomials of the second kind.

A Transfer Matrix Approach

The Szegő recurrences (4) and (5) can be rewritten as

$$\begin{bmatrix} \Phi_{n+1}(z) \\ \Phi_{n+1}^*(z) \end{bmatrix} = \boldsymbol{C}_n(z) \begin{bmatrix} \Phi_n(z) \\ \Phi_n^*(z) \end{bmatrix}, \quad \boldsymbol{C}_n(z) = \begin{bmatrix} z & -\overline{\alpha}_n \\ -\alpha_n z & 1 \end{bmatrix}, \tag{27}$$

where $\boldsymbol{C}_n(z)$ is said to be a transfer matrix (see [16, 32, 33]). Notice that if we set $\Phi_n(z) := 0$ and $\Phi_n^*(z) := 0$ for $n \leq -1$, and $\alpha_{-1} := -1$ and $\alpha_n := 0$, for $n \leq -2$, then (27) holds for all $n \in \mathbb{Z}$.

Obviously, $\Phi_n(z;k)$ (resp. $\Omega_n(z;k)$) is exactly $\Phi_n(z)$ (resp. $\Omega_n(z)$), for $n \leq k$. When $n \geq k+1$, we follow an analogue procedure to the one of [32, Chap. 3] to obtain the relation between $\Phi_n(z;k)$ (resp. $\Omega_n(z;k)$) and $\Phi_n(z)$ (resp. $\Omega_n(z)$) using a simple matrix recursion based on the transfer matrix for the COPUC.

Theorem 9 ([3]) *The following relations hold:*

$$2z^{k+2}\kappa_{k+1}^{-2} \begin{bmatrix} -\Omega_{n+1}(z;k) \\ \Phi_{n+1}(z;k) \end{bmatrix} = \boldsymbol{B}_k(z) \begin{bmatrix} -\Omega_{n+1}(z) \\ \Phi_{n+1}(z) \end{bmatrix}, \quad n \geq k+1,$$

where the corresponding transfer matrix $\boldsymbol{B}_k(z)$ *is*

$$\boldsymbol{B}_k(z) = \begin{bmatrix} r(z;k)\Omega_k^*(z) + zr^*(z;k)\Omega_k(z) & s(z;k)\Omega_k^*(z) - zs^*(z;k)\Omega_k(z) \\ r(z;k)\Phi_k^*(z) - zr^*(z;k)\Phi_k(z) & s(z;k)\Phi_k^*(z) + zs^*(z;k)\Phi_k(z) \end{bmatrix},$$

with

$$r(z;k) = (1-\alpha_k\overline{\beta}_k)z\Phi_k(z) - \overline{(\alpha_k - \beta_k)}\Phi_k^*(z),$$

$$s(z;k) = (1-\alpha_k\overline{\beta}_k)z\Omega_k(z) + \overline{(\alpha_k - \beta_k)}\Omega_k^*(z).$$

Next, we give a relation between the COPUC associated with two modifications of different level.

Corollary 6 ([3]) *For $\ell < k$, the following relation holds :*

$$z^{k-\ell}\left(\frac{\kappa_{k+1}}{\kappa_{\ell+1}}\right)^{-2}\begin{bmatrix}-\Omega_{n+1}(z;k)\\ \Phi_{n+1}(z;k)\end{bmatrix} = \boldsymbol{B}_k(z)\boldsymbol{B}_\ell^{-1}(z)\begin{bmatrix}-\Omega_{n+1}(z;\ell)\\ \Phi_{n+1}(z;\ell)\end{bmatrix}, \quad n \geq k+1.$$

For a finite composition of perturbations we have the following result:

Theorem 10 ([3]) *For $0 \leq l < \cdots < m < \infty$, the following relation holds:*

$$2^{m-l}\prod_{j=l+1}^{m+1} z^{j+1}\kappa_j^{-2}\begin{bmatrix}-\Omega_{n+1}(z;l,\ldots,m)\\ \Phi_{n+1}(z;l,\ldots,m)\end{bmatrix} = \prod_{j=l}^{m}\boldsymbol{B}_j(z)\begin{bmatrix}-\Omega_{n+1}(z)\\ \Phi_{n+1}(z)\end{bmatrix}, \quad n \geq l+1.$$

Spectral Transformations of COPUC

Let $F(z;k)$ be the $\mathcal{C}$-function, associated with the sequence of COPUC $\{\Phi_n(z;k)\}_{n\geq 0}$. We begin expressing $F(z;k)$ in terms of $F_\sigma(z)$.

Theorem 11 ([3]) *$F(z;k)$ is a pure rational spectral transformation of $F_\sigma(z)$, given by*

$$F(z;k) \doteq \boldsymbol{B}(z;k)F_\sigma(z). \tag{28}$$

From the above theorem, as in [27], we can obtain the orthogonality measure associated with $F(z;k)$.

Let $F(z;l,\ldots,m)$ be the $\mathcal{C}$-function associated with the finite composition of COPUC $\{\Phi_n(z;l,\ldots,m)\}_{n\geq 0}$. As a consequence of Theorem 10, we obtain the next corollary.

Corollary 7 ([3]) *$F(z;l,\ldots,m)$ is a pure rational spectral transformation of $F_\sigma(z)$, given by*

$$F(z;l,\ldots,m) \doteq \prod_{j=l}^{m}\boldsymbol{B}_j(z)F_\sigma(z). \tag{29}$$

The above corollary can be read in terms of quadratic irrationality. An analytic function $f(z)$ in $\mathbb{D}$ is said to satisfy quadratic irrationality condition if and only if there exist polynomials $a(z)$, $b(z)$, and $c(z)$ such that

$$a(z)f^2(z) + b(z)f(z) + c(z) = 0, \quad a(z) \neq 0.$$

Lemma 1 ([32]) *Let f and g be analytic functions in $\mathbb{D}$ such that there exist polynomials $a(z)$, $b(z)$, $c(z)$, and $d(z)$ with*

$$\boldsymbol{A}(z) = \begin{bmatrix} a(z) & b(z) \\ c(z) & d(z) \end{bmatrix}, \quad \det \boldsymbol{A}(z) \neq 0,$$

so that

$$g(z) \doteq \boldsymbol{A}(z) f(z).$$

Then g satisfies a quadratic irrationality condition if and only if f satifies such a condition.

In other words, the quadratic irrationality condition is preserved by rational spectral transformations. As a straightforward consequence of Corollary 11 and Lemma 1, we obtain the next result.

Corollary 8 ([3]) *$F(z)$ satisfies a quadratic irrationality condition if and only if $F(z; l, \cdots, m)$ is.*

Notice that Corollary 8 characterizes in terms of the Verblunsky coefficients a wide class of orthogonal polynomials such that their corresponding $\mathcal{C}$-functions satisfy a quadratic irrationality condition.

Szegő Transformation and Co-polynomials

In this section, we explore the relation between the co-polynomials on the real line (resp. on the unit circle), and the corresponding sequences of monic orthogonal polynomials obtained on the unit circle (resp. on the real line) via Szegő transformation.

Szegő Transformation and Co-polynomials on the Real Line

The modification of the Verblunsky coefficients for the corresponding OPUC associated with the perturbed recurrence coefficients through the Szegő transformation is shown in the following result:

Theorem 12 ([6]) *Let $\{\widehat{\alpha}_n\}_{n\geq 0}$ be the Verblunsky coefficients for the corresponding OPUC, associated with (10) and (11) through the Szegő transformation. Then, for a fixed non-negative integer k,*

$$\begin{aligned} \widehat{\alpha}_n &= \alpha_n, \quad 0 \leq n < 2k-1, \\ \widehat{\alpha}_{2k-1} &= \alpha_{2k-1} + M, \end{aligned}$$

$$\begin{aligned}
\widehat{\alpha}_{2k} &= \frac{(1-\alpha_{2k-1})\alpha_{2k} + 2\tau_{k+1} + M\alpha_{2k-2}}{1-\alpha_{2k-1}-M}, \\
\widehat{\alpha}_{2m+1} &= -1 + \frac{4d_{m+1}}{(1-\widehat{\alpha}_{2m-1})(1-\widehat{\alpha}^2_{2m})}, \quad n = 2m+1,\ m \geq k, \\
\widehat{\alpha}_{2m} &= \frac{2b_{m+1} + (1+\widehat{\alpha}_{2m-1})\widehat{\alpha}_{2m-2}}{1-\widehat{\alpha}_{2m-1}}, \quad n = 2m,\ m \geq k+1,
\end{aligned}$$

where $M = \dfrac{4(\lambda_k - 1)d_k}{(1-\alpha_{2k-3})(1-\alpha^2_{2k-2})}$.

Note that, through the Szegő transformation, the modifications (10) and (11) yield the modification of all the Verblunsky coefficients $\widehat{\alpha}_n$ with $n > k$.

Example 3 Let $\{S_n(x)\}_{n\geq 0}$ be a sequence of monic symmetric polynomials orthogonal with respect to an even weight function supported on a symmetric subset of $[-1, 1]$. They are generated by

$$S_{n+1}(x) = xS_n(x) - d_nS_{n-1}(x), \quad d_n \neq 0, \quad d_0 = 1, \quad n \geqslant 0,$$

with initial conditions $S_{-1}(x) = 0$ and $S_0(x) = 1$, see [8].

Let $\{\widehat{\gamma}_n\}_{n\geq 0}$ be the Verblunsky coefficients for the corresponding OPUC, associated with (10) through the Szegő transformation. Then, for a fixed non-negative integer k,

$$\begin{aligned}
\widehat{\gamma}_{2n} &= \gamma_{2n} = 0, \quad n \geq 0, \\
\widehat{\gamma}_{2n-1} &= \gamma_{2n-1}, \quad 0 \leq n < k, \\
\widehat{\gamma}_{2k-1} &= \gamma_{2k-1} + \frac{4(\lambda_k - 1)d_k}{1-\gamma_{2k-3}}, \\
\widehat{\gamma}_{2n+1} &= -1 + \frac{4d_{n+1}}{(1-\widehat{\gamma}_{2n-1})}, \quad n \geq k.
\end{aligned}$$

Notice that the modification (10) yields, from the Szegő transformation, the modification of all odd Verblunsky coefficients greater than k.

Consider the $\mathcal{S}$-function $S(x; \lambda_k, \tau_{k+1})$, associated with the COPRL [5], given by

$$S(x; \lambda_k, \tau_{k+1}) \doteq cof(\mathbf{M}_k)S(x).$$

By applying the Szegő transformation in the above expression, we have the following result:

Theorem 13 ([6]) *Let $F(z; \lambda_k, \tau_{k+1})$ be the $\mathcal{C}$-function associated with the perturbations (10) and (11) through the Szegő transformation. Then,*

$$\frac{2z}{1-z^2}F(z;\lambda_k,\tau_{k+1}) \doteq \mathrm{cof}\left(\mathbf{M}_k\left(\frac{z+z^{-1}}{2}\right)\right)\left(\frac{2z}{1-z^2}F(z)\right),$$

with $2x = z + z^{-1}$.

For the finite composition of perturbations (10) and (11), we can consider the $\mathcal{S}$-function $S(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1})$, associated with the COPRL $P_n(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1})$ [5], given by

$$S(x;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1}) \doteq cof\left(\prod_{j=m}^{k}\mathbf{M}_j\right)S(x).$$

Then, applying the Szegő transformation, we get the following result:

Theorem 14 ([6]) *Let* $F(z;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1})$ *be the* $\mathcal{C}$*-function associated with the finite composition of perturbations* (10) *and* (11) *through the Szegő transformation. Then,*

$$\frac{2z}{1-z^2}F(z;\lambda_m,\tau_{m+1};\ldots;\lambda_k,\tau_{k+1}) \doteq cof\left(\prod_{j=m}^{k}\mathbf{M}_j\left(\frac{z+z^{-1}}{2}\right)\right)\left(\frac{2z}{1-z^2}F(z)\right),$$

with $2x = z + z^{-1}$.

Szegő Transformation and Co-polynomials on the Unit Circle

It is easy to check that (26) implies, through the Szegő transformation, the modification of both sequences of recurrence coefficients associated with the OPRL. More precisely,

Theorem 15 ([3]) *Let* $\{\widehat{b}_n\}_{n\geq 1}$ *and* $\{\widehat{d}_n\}_{n\geq 1}$ *be the recurrence coefficients for the corresponding COPRL associated with (k-modification) through the Szegő transformation. Then, for* $k = 2m-1$,

$$\widehat{d}_{n+1} = \left(\frac{1+\eta_{2m-1}}{1+\alpha_{2m-1}}\right)^{\delta_{n+1,m}}\left(\frac{1-\eta_{2m-1}}{1-\alpha_{2m-1}}\right)^{\delta_{n+1,m+1}} d_{n+1},$$

$$\widehat{b}_{n+1} = b_{n+1} + \frac{1}{2}(\alpha_{2m-1}-\eta_{2m-1})(\alpha_{2m}+\alpha_{2m-2})\delta_{n+1,m+1},$$

and for $k = 2m$,

$$\begin{aligned}
\widehat{d}_{n+1} &= \left(\frac{1-\eta_{2m}^2}{1-\alpha_{2m}^2}\right)^{\delta_{n+1,m+1}} d_{n+1},\\
\widehat{b}_{n+1} &= b_{n+1} + \frac{1}{2}(\eta_{2m}-\alpha_{2m})(1-\alpha_{2m-1})\delta_{n+1,m+1} + \frac{1}{2}(\alpha_{2m}-\eta_{2m})(1+\alpha_{2m+1})\delta_{n+1,m+2}.
\end{aligned}$$

Notice that the co-dilated case (resp. co-recursive case) for OPRL yields, through the Szegő transformation, the modifications of all odd (resp. even) Verblunsky coefficients.

Example 4 Let $d\sigma$ be a nontrivial probability measure supported on $\mathbb{T}$ and let $\{\Phi_n(z)\}_{n\geq 0}$ be the corresponding OPUC. For a positive integer ℓ the sieved OPUC $\{\Phi_n^{\{\ell\}}(z)\}_{n\geq 0}$ are defined as those orthogonal polynomials associated with the Verblunsky coefficients $\{\alpha_n^{\{\ell\}}\}_{n\geq 0}$ given by

$$\alpha_n^{\{\ell\}} = \begin{cases} \alpha_{m-1} & \text{if } n+1 = m\ell,\\ 0 & \text{otherwise,}\end{cases} \tag{30}$$

for $n \geq 0$. We also denote by $\sigma^{\{\ell\}}$ the nontrivial probability measure supported on $\mathbb{T}$ associated with $\{\alpha_n^{\{\ell\}}\}_{n\geq 0}$. Note that $\{\Phi_n^{\{1\}}(z)\}_{n\geq 0}$ are the polynomials $\{\Phi_n(z)\}_{n\geq 0}$. The earliest treatment of $\{\Phi_n^{\{\ell\}}(z)\}_{n\geq 0}$ for $\ell \geq 2$ is found in [1, 18, 22]. The best general reference on this subject is the work by Petronilho [28], see also [9].

Consider the case $\ell = 2$, then from (30) we have $\{\alpha_n^{\{2\}}\}_{n\geq 0} = \{0, \alpha_0, 0, \alpha_1, \ldots\}$. Then we have the following result.

Let $\{\widehat{b}_n^{\{2\}}\}_{n\geq 1}$ and $\{\widehat{d}_n^{\{2\}}\}_{n\geq 1}$ be the recurrence coefficients for the corresponding OPRL associated with (k-modification) through the Szegő transformation. Then, for a fixed non-negative integer k,

$$\begin{aligned}
\widehat{d}_{n+1}^{\{2\}} &= \left(\frac{1+\eta_k}{1+\alpha_k}\right)^{\delta_{n+1,k+1}} \left(\frac{1-\eta_k}{1-\alpha_k}\right)^{\delta_{n+1,k+2}} d_{n+1}^{\{2\}},\\
\widehat{b}_{n+1}^{\{2\}} &= 0.
\end{aligned}$$

Note that this k-modification yields, by using the Szegő transformation, the modification of two consecutive recurrence coefficients $d_{k+1}^{\{2\}}$ and $d_{k+2}^{\{2\}}$.

Consider the $\mathcal{C}$-function $F(z;k)$, corresponding to the COPUC [3], given by

$$F(z;k) \doteq \boldsymbol{B}(z;k)F_\sigma(z).$$

Then, applying the Szegő transformation to the above equation, we have

Theorem 16 ([3]) *Let $S(x;k)$ be the $\mathcal{S}$-function for the corresponding COPRL associated with (26) through the Szegő transformation. Then,*

$$\sqrt{x^2-1}\,S(x;k) \doteq \boldsymbol{B}_k(x-\sqrt{x^2-1})\left(\left(\frac{1}{x-\sqrt{x^2-1}}-x\right)S_\mu(x)\right),$$

where $z = x - \sqrt{x^2-1}$.

Let us consider the $\mathcal{C}$-function $F(z;l,\dots,m)$ associated with the finite composition of perturbations (26) [3], given by

$$F(z;l,\dots,m) \doteq \prod_{j=l}^{m} \boldsymbol{B}_j(z)F_\sigma(z).$$

Then, applying the Szegő transformation, we get

Theorem 17 ([6]) *Let $S(x;l,\dots,m)$ be the $\mathcal{S}$-function for the corresponding OPRL associated with the finite composition of perturbations (26) through the Szegő transformation. Then,*

$$\sqrt{x^2-1}\,S(x;l,\dots,m) \doteq \prod_{j=l}^{m} \boldsymbol{B}_j(x-\sqrt{x^2-1})\left(\sqrt{x^2-1}S_\mu(x)\right),$$

where $z = x - \sqrt{x^2-1}$.

Verblunsky Coefficients and LU Factorization

If we define the sequence $\{v_k\}_{k\geqslant 0}$, given in terms of the Verblunsky coefficients by

$$v_k = \frac{1}{2}(1+\alpha_k)(1-\alpha_{k-1}), \tag{31}$$

then

$$d_{k+1} = v_{2k}v_{2k+1}, \tag{32}$$

$$b_{k+1}+1 = v_{2k-1}+v_{2k}, \tag{33}$$

and, we can find a unique factorization

$$\mathbf{J}+\mathbf{I} = \mathbf{LU},$$

where **J** is the Jacobi matrix associated with (2), **I** is the semi-infinite identity matrix, **L** is a lower bidiagonal matrix, and **U** is a upper bidiagonal matrix, with

$$\mathbf{L} = \begin{bmatrix} 1 & & & \\ v_1 & 1 & & \\ & v_3 & 1 & \\ & & \ddots & \ddots \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} v_0 & 1 & & \\ & v_2 & 1 & \\ & & v_4 & 1 \\ & & & \ddots & \ddots \end{bmatrix}.$$

Thus, from (31), we have

$$\alpha_k = -1 + \frac{2v_k}{1-\alpha_{k-1}}, \tag{34}$$

or equivalently,

$$\alpha_k = -1 + \frac{2v_k|}{|2} - \frac{2v_{k-1}|}{|2} - \cdots - \frac{2v_1|}{|2-v_0}.$$

Therefore, from the a sequence $\{v_k\}_{k\geqslant 0}$ we can determine in a very simple way the Verblunsky coefficients $\{\alpha_k\}_{k\geqslant 0}$ for the measure $d\sigma$ supported on $\mathbb{T}$.

From (32) and (33), we can find the sequence $\{v_k\}_{k\geqslant 0}$ in terms of the recurrence coefficients $\{b_k\}_{k\geqslant 1}$ and $\{d_k\}_{k\geqslant 1}$, as follows:

$$v_{2k} = b_{k+1} + 1 - \frac{d_k|}{|b_k+1} - \cdots - \frac{d_1|}{|b_1+1},$$

$$v_{2k+1} = \frac{d_{k+1}|}{|b_{k+1}+1} - \frac{d_k|}{|b_k+1} - \cdots - \frac{d_1|}{|b_1+1},$$

with $k \geq 0$.

If we perturb the entries of the Jacobi matrix **J** at the level k, then we have a new sequence $\{\tilde{v}_n\}_{n\geqslant 0}$, given by

$$\tilde{v}_{2n} = b_{n+1} + 1 - \frac{d_n|}{|b_n+1} - \cdots - \frac{d_{k+1}|}{|b_{k+1}+\tau_{k+1}+1} - \frac{\lambda_k d_k|}{|b_k+1} - \cdots - \frac{d_1|}{|b_1+1},$$

$$\tilde{v}_{2n+1} = \frac{d_{n+1}}{\tilde{v}_{2n}},$$

with $k \geq 0$. This can be summarized in the following:

Theorem 18 ([6]) *Let* $\{\tilde{v}_n\}_{n\geqslant 0}$ *be the new sequence associated with* (10) *and* (11). *Then*

$$\begin{aligned}\tilde{v}_n &= v_n, \quad 0 \le n \le 2k-1,\\ \tilde{v}_{2k} &= v_{2k} + (1-\lambda_k)v_{2k-1} + \tau_{k+1},\\ \tilde{v}_{2m+1} &= \frac{d_{m+1}}{\tilde{v}_{2m}},\ \tilde{v}_{2(m+1)} = b_{m+2} + 1 - \tilde{v}_{2m+1},\ m \ge k.\end{aligned}$$

Therefore, as we mentioned previously, we can compute the new Verblunsky coefficients directly from the sequence $\{\tilde{v}_n\}_{n\geqslant 0}$ as follows:

Theorem 19 ([6]) *Let* $\{\widehat{\alpha}_n\}_{n\geqslant 0}$ *be the Verblunsky coefficients for the corresponding OPUC associated with the perturbations* (10) *and* (11) *through Szegő transformation. Then,*

$$\begin{aligned}\widehat{\alpha}_n &= \alpha_n, \quad 0 \le n \le 2k-1,\\ \widehat{\alpha}_{2k} &= \alpha_{2k} + \frac{2[(1-\lambda_k)v_{2k-1} + \tau_{k+1}]}{1-\alpha_{2k-1}},\\ \widehat{\alpha}_n &= -1 + \frac{2\tilde{v}_n}{1-\widehat{\alpha}_{n-1}}, \quad n \ge 2k+1.\end{aligned}$$

This is an alternative way to compute the perturbed Verblunsky coefficients through the Szegő transformation using the LU factorization.

Acknowledgements The authors wish to express their thanks to Th. M. Rassias and N. J. Daras for the invitation to participate in this volume. The research of the first author is supported by the Portuguese Government through the FCT under the grant SFRH/BPD/101139/2014 and partially supported by the Brazilian Government through the CNPq under the project 470019/2013-1. The research of the first and second author is supported by Dirección General de Investigación Científica y Técnica, Ministerio de Economía y Competitividad of Spain, grant MTM2012-36732-C03-01.

References

1. V.M. Badkov, Systems of orthogonal polynomials explicitly represented by the Jacobi polynomials. Math. Notes **42**, 858–863 (1987)
2. P. Borwein, T. Erdelyi, *Polynomials and Polynomial Inequalities* (Springer, New York, 1995)
3. K. Castillo, On perturbed Szegő recurrences. J. Math. Anal. Appl. **411**, 742–752 (2014)
4. K. Castillo, Monotonicity of zeros for a class of polynomials including hypergeometric polynomials. Appl. Math. Comput. **266**, 173–193 (2015)
5. K. Castillo, F. Marcellán, J. Rivero, On co-polynomials on the real line. J. Math. Anal. Appl. **427**, 469–483 (2015)
6. K. Castillo, F. Marcellán, J. Rivero, On perturbed orthogonal polynomials on the real line and the unit circle via Szegő's transformation. Appl. Math. Comput. (2017, Accepted for publication)

7. T.S. Chihara, On co-recursive orthogonal polynomials. Proc. Am. Math. Soc. **8**, 899–905 (1957)
8. T.S. Chihara, An introduction to orthogonal polynomials, in *Mathematics and Its Applications*, vol. 13 (Gordon and Breach, New York/London/Paris, 1978)
9. M.N. de Jesus, J. Petronilho, On orthogonal polynomials obtained via polynomial mappings. J. Approx. Theory **162**, 2243–2277 (2010)
10. J. Dini, P. Maroni, A. Ronveaux, Sur une perturbation de la récurrence vérifiée par une suite de polynômes orthogonaux. Portugal. Math. **46**, 269–282 (1989)
11. W. Erb, Optimally space localized polynomials with applications in signal processing. J. Fourier Anal. Appl. **18**(1), 45–66 (2012)
12. W. Erb, Accelerated Landweber methods based on co-dilated orthogonal polynomials. Numer. Algorithms **68**, 229–260 (2015)
13. Y.L. Geronimus, On some difference equations and corresponding systems of orthogonal polynomials. Izv. Akad. Nauk SSSR, Ser. Mat. **5**, 203–210 (1943)
14. Y.L. Geronimus, *Orthogonal Polynomials: Estimates, Asymptotic Formulas and Series of Polynomials Orthogonal on the Unit Circle and on an Interval* (Consultants Bureau, New York, 1961)
15. Y.L. Geronimus, Orthogonal polynomials on a circle and their applications. Am. Math. Soc. Translat. Ser. 1 **3**, 1–78 (1962)
16. L. Golinskii, P. Nevai, Szegő difference equations, transfer matrices and orthogonal polynomials on the unit circle. Commun. Math. Phys. **223**, 223–259 (2001)
17. M.E.H. Ismail, *Classical and Quantum Orthogonal Polynomials in One Variable*. Encyclopedia in Mathematics and its Applications, vol. 98 (Cambridge University Press, Cambridge, 2005)
18. M. Ismail, X. Li, On sieved orthogonal polynomials IX: orthogonality on the unit circle. Pac. J. Math. **152**, 289–297 (1992)
19. C.G.J. Jacobi, Über die reduction der quadrastischen formen auf die kleinste anzahl glieder. J. Reine Angew. Math. **39**, 290–292 (1848)
20. J. Letessier, Some results on co-recursive associated Laguerre and Jacobi polynomials. SIAM J. Math. Anal. **25**(2), 528–548 (1994)
21. G.G. Lorentz, M.V. Gollitschek, Y. Makovoz, *Constructive Approximation* (Springer, New York, 1996)
22. F. Marcellán, G. Sansigre, Orthogonal polynomials on the unit circle: symmetrization and quadratic decomposition. J. Approx. Theory **65**, 109–119 (1991)
23. F. Marcellán, J.S. Dehesa, A. Ronveaux, On orthogonal polynomials with perturbed recurrence relations. J. Comput. Appl. Math. **30**, 203–212 (1990)
24. L.M. Milne-Thomson, *The Calculus of Finite Differences*. American Mathematical Society (Chelsea Publishing, Providence, 2000)
25. G.V. Milovanovic, M.Th. Rassias (eds.), *Analytic Number Theory, Approximation Theory and Special Functions* (Springer, New York, 2014)
26. F. Peherstorfer, Finite perturbations of orthogonal polynomials. J. Comput. Appl. Math. **44**, 275–302 (1992)
27. F. Peherstorfer, A special class of polynomials orthogonal on the unit circle including the associated polynomials. Constr. Approx. **12**, 161–185 (1996)
28. J. Petronilho, Orthogonal polynomials on the unit circle via a polynomial mapping on the real line. J. Comput. Appl. Math. **216**, 98–127 (2008)
29. A. Ronveaux, Fourth-order differential equations for numerator polynomials. J. Phys. A Math. Gen. **21**(15), L749 (1988)
30. A. Ronveaux, F. Marcellán, Co-recursive orthogonal polynomials and fourth-order differential equation. J. Comput. Appl. Math. **25**(1), 105–109 (1989)
31. A. Ronveaux, S. Belmehdi, J. Dini, P. Maroni, Fourth-order differential equation for the co-modified of semi-classical orthogonal polynomials. J. Comput. Appl. Math. **29**(2), 225–231 (1990)

32. B. Simon, *Orthogonal polynomials on the unit circle, Part 1: Classical theory*. Colloquium Publications Series, vol. 54 (American Mathematical Society, Providence, 2005)
33. B. Simon, *Orthogonal polynomials on the unit circle, Part 2: Spectral theory*. Colloquium Publications Series, vol. 54 (American Mathematical Society, Providence, 2005)
34. B. Simon, *Szegő's Theorem and Its Descendants: Spectral Theory for L^2 Perturbations of Orthogonal Polynomials* (Princeton University Press, Princeton, 2011)
35. H.A. Slim, On co-recursive orthogonal polynomials and their application to potential scattering. J. Math. Anal. Appl. **136**, 1–19 (1988)
36. G. Szegő, *Orthogonal Polynomials*, 4th edn. Colloquium Publications Series, vol. 23 (American Mathematical Society, Providence, 1975)
37. A. Zhedanov, Rational spectral transformations and orthogonal polynomials. J. Comput. Appl. Math. **85**, 67–86 (1997)

Electromagnetic Compatibility (EMC) in Challenging Environments

C. Christopoulos

Abstract The paper describes the essential aspects of ElectroMagnetic Compatibility (EMC) as applied to the response of critical systems to severe ElectroMagnetic (EM) threats. The significance of deterministic and stochastic models is outlined together with the role of numerical modelling and physical testing in the analysis and synthesis of complex systems. It is emphasised that the exploitation of the synergies between modelling and testing is the best way to approach the EM design of complex systems.

Keywords Electromagnetic interference • Electromagnetic compatibility • Electromagnetic hardening of systems

Introduction

Modern engineering systems, including defence systems, incorporate a multitude of electrical and electronic systems for command, control, communication and actuation purposes. Increasingly, electromechanical systems are being replaced by electronic systems since this offers advantages in terms of weight, cost and flexibility. Electromechanical systems are reliable as long as physical integrity is maintained. It is difficult to compromise their operation other than by direct physical damage. In contrast, electronic based systems, which rely on the transmission of data streams or power pulses for control and actuation purposes, can be made to malfunction or be permanently damaged if by remote action messages encoded in electrical signals are interfered with. This may be caused either by deliberate action (electronic warfare) or, inadvertently by other users of the electromagnetic (EM) spectrum (ElectroMagnetic Interference-EMI). Hence, defence systems in particular are designed to be able to sustain severe electromagnetic threats so that their operation, or at least their failure in a safe mode, is assured. By defence

C. Christopoulos (✉)
Emeritus Professor of Electrical Engineering, University of Nottingham, Nottingham, UK
e-mail: christos.christopoulos@nottingham.ac.uk

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_5

systems we mean a very wide range of assets be they land, sea or air based, and also important infrastructure such as essential power and communications networks, and logistics facilities.

ElectroMagnetic Compatibility (EMC) is the discipline concerned with the analysis and design of systems which are electromagnetically compliant with each other, are able to sustain a certain amount of interference and in addition do not contribute undue amounts of EM energy to the environment (EM pollution) [1]. The maximum EM energy levels that each systems is designed to withstand without losing functionality (its immunity or susceptibility level) depends on its nature and importance and varies widely between commercial and military systems. Similarly, permitted EM energy levels which may be emitted by systems are specified in commercial and military standards. Such limits are specified over broad frequency ranges and it is fair to say that increasingly the range of frequencies over which EMC standards need to be enforced is continuously extending. We should work on the basis that we need to design electromagnetically compliant systems from dc to light!

In the case of military systems the requirements are very strict as in most cases we deal with safety critical systems where malfunction may have catastrophic results (very high immunity is required). The emission aspect is also important as defence systems which broadcast an electromagnetic signature are easily detected and with some effort important information can be obtained by a hostile agent on the operation of the system, internal information exchanges, etc. [2, 3].

It is therefore imperative that all designers and users of defence systems are familiar with the essentials of EMI and EMC and in particular with severe EM threats as are likely to be experienced in adverse environments and in hostile situations. In this article we aim to offer some guidance to the main issues and remedies which may be employed to reduce risks and harden systems to potential EM threats. It is not possible to offer comprehensive guidance but interested readers may access references for more details [1, 4].

EMC in the Context of Electromagnetic System Design

In this section we address the reasons that EMC is becoming increasingly important to the EM design of systems. These are technological and economic in nature.

Starting with the technological issues, it is undoubtedly the case that the replacement of cumbersome pneumatic and hydraulic controls by compact electronic controllers brings benefits in terms of size, weight, long-term reliability, servicing, versatility and flexibility. An example is the initiative on more electric aircraft where weight is at a premium [5, 6]. The immense processing power of modern electronic controllers allows for sophisticated control and monitoring strategies to be implemented, and regularly updated with minimum cost and disturbance to operations. On the negative side, the operation of electronic controllers may be affected by electrical signals which deliberately or inadvertently impinge on these controllers either guided by conductors (conducted interference) or airborne (radiated interference). If this were to happen then, potentially, systems will malfunction or even fail completely and may require extensive repairs and replacement. These problems are not new—practitioners were aware of the potential vulnerability of electronic systems to EMI since the 1930s. What is new in the last 30 years is the extensive use of electronics in engineering, the lowering of logic level to a few volts and the ubiquitous presence of EM spectrum users. These factors taken together have lowered the EM energy level which may cause logic malfunctions and increased the EM energy levels that are injected into the environment as the result of the operation of commercial and military systems. Thus the challenges come from two directions—the increased vulnerability of systems on the one hand and the more electromagnetically hostile environment on the other.

It is also important to emphasise that EM emission comes in different ways, e.g. narrow band (e.g. radar), broad band (e.g. LEMP, NEMP) as will be seen in the next section. Thus, whilst a system may appear on first reflection to operate over a restricted frequency range (e.g. an electrical motor drive connected to a mains supply of 50, 60 or 300 Hz), in reality we are dealing with a system containing microprocessor equipment driven by clock frequencies in the GHz range, switched mode power supplies driven by clocks in the MHz range, which, including their harmonics, generate EMI energy well into the microwave range. In terms of vulnerability such systems are susceptible to EMI over an equally wide frequency range. The designer and user of such equipment must be aware of the susceptibility and emission from such systems not just over the restricted frequency range represented by the functional (operational) signals but also over the broad band of frequencies where the system is vulnerable and/or injects EMI into the environment. This is the most challenging aspect of EMC as it requires the characterisation of systems across a very wide frequency band.

Equally important and challenging are the cost implications of EMC or lack of it. It is difficult to obtain accurate figures on EMC costs but a minimum 2% of the development cost of new products may be attributed to EMC. The exact figure depends on the nature of the product, the degree of innovation and the manner in which EMC is embedded into the design process. Basically, we can adopt one of the two strategies.

First, we may take the optimistic view that our design will not suffer from EMC problems and therefore design without reference to EMC. We thus spend no money or time on EMC related issues before a prototype has been produced. On the production of a prototype we then test and more often than not we will find that we have EMC problems. Fixing them will require spending money and time. It is a truism that retrofitting is a costly way to address problems as many potential remedies are simply not available so late in the design process. Significant costs and unplanned delays are the characteristics of this approach.

Second, we may approach EMC as a fundamental aspect of design and therefore spend time and money at the start of the design phase addressing potential EMC issues. At the production of a prototype we test and inevitably there may still be some EMC issues to address. These, in a well-managed design, will be relatively minor and may be addressed promptly and without major expense. The result is invariably a compliant design arrived at in a shorter time and at a lower cost compared with the first approach. The implications of this are that EMC should be incorporated into design right from the start to minimise costs and allow scope for seeking the optimum technical solutions without undue constrains imposed by already solidified design choices.

We are thus brought to recognise that EMC must be part of a concurrent design process, essentially communicating with a whole host of other design requirements, e.g. mechanical, thermal, etc. Only in this way can we hope that EMC can be ensured in an optimal way in a complex design environment. Naturally, this requires the availability of sophisticated design tools (analytical, numerical, experimental, etc.), highly trained personnel, awareness of the level of EM threats and the regulatory and standards regime applicable in each case. In addition to proper design for EMC, a continuing effort should be made to manage EMC throughout the lifetime of the designed system. Often, a good design is compromised by changes made as part of maintenance or for operational reasons by persons who are unfamiliar with EMC and focus instead on convenience during normal operations. This is clearly highly undesirable and counters good EMC practices embedded during the design phase.

We will have something to say on some (but not all) of these issues in the forthcoming sections.

EMI Sources, Coupling Paths and Potential Victims

In order for EMC issues to arise there must be an EM threat (a source of EMI), one or more coupling paths via which EM energy can travel from the source of EMI to the engineering system, and components or sub-systems which are vulnerable to EM threats (potential victims). If at least one of these three factors affecting the EM integrity of systems is absent, then we do not have an EMC problem. It follows therefore that we can reduce the risks of malfunctions due to EM threats by intervening in one or more of these three aspects of EMC. We may seek to reduce the EM energy generated at source (if at all possible), weaken coupling paths (e.g. by introducing EM shielding) or increase the immunity of components so that they can sustain a certain amount of EMI without malfunction or permanent damage (hardening). Each of these interventions has its own technological challenges and cost implications and therefore designers must be able to weigh against each other the various proposed remedies and interventions aimed at hardening systems to EM threats and assuring compliance to applicable standards and in-house requirements. A well-designed system should meet all applicable civilian and military EMC standards and any further additional requirements which specific systems may demand because of special operational circumstances. Typical military standards are described in [7, 8]. It has to be accepted that that the nature of EM emissions is such that they do not respect boundaries or a black-box approach. EMC is best pursued by adopting a holistic global approach, working closely with other design disciplines, and being able to influence the design team in recognising the importance and challenges of designing EM compliant systems [9]. To illustrate the nature and extent of the challenges we offer below a shortlist as an illustration only of the various facets of the EMC problem. In reality, clear-cut boundaries do not always exist between the various areas indicated below:

1. Source region (e.g. the extent, frequency range, duration and intensity of the EM threat). Site surveys may be necessary or access to other research results to assess fully the nature of the EM threat. If the system under consideration is very close to the source region (e.g. for NEMP), then highly non-linear phenomena take place making the entire study very difficult.
2. External interaction (e.g. the calculation of induced current flows on the outer surfaces-shields of the system, on wire penetrations, on grounding wires, etc.).
3. Internal interaction (e.g. current flows on internal wiring, penetration through apertures, wire-to-wire and field-to-wire coupling inside the system).
4. Device susceptibility (e.g. calculation of device pin currents and voltages, calculation of failure probabilities of particular sub-systems).

Naturally, analysis of the threats is only one aspect and should it indicate an unacceptable probability of failure then the designer is required to develop a synthesis of counter measures (e.g. filtering, electrical isolation, balancing, shielding, better grounding, etc.) capable of reducing the risks.

Sources of EMI

Any serious study of electromagnetic interference requires an assessment and basic understanding of EM threats. EMI can come from natural or man-made sources.

Natural sources contribute to the EM environment at low-frequencies (LF), e.g. through slow fluctuations of the earth's magnetic field, through lightning discharges and at high-frequencies through extra-terrestrial processes. We will describe in more detail the nature of the lightning EM pulse (LEMP) as it is severe enough to affect even well-designed systems. But this should not be taken to mean that other natural threats such as the slow variation of magnetic fields during geomagnetic storms is harmless—there is ample evidence that they may affect the operation of major infrastructure (e.g. power systems) by tripping protective relays and interrupting energy supplies over large territories.

Man-made sources of EMI are the result of human activity and have been increasing steadily as more IT, communications and power electronic systems are incorporated into industrial and military processes. Civilian EMC standards set upper limits for emissions by individual equipment but there is no doubt that background noise over large areas of the frequency spectrum is steadily on the increase. Results from some recent surveys of commercial environments are given in [10]. An understanding of civilian EMC specifications is also useful in

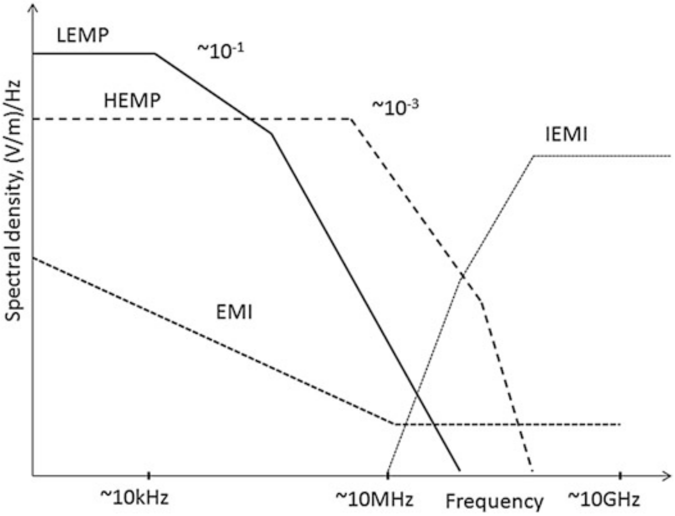


Fig. 1 Schematic of the EM interference spectrum indicating particular threats

the military field as increasingly commercial off the self (COTS) equipment are employed in military applications. However, we will limit our discussion here to two examples of severe EM threat-the lightning EM pulse (LEMP) and nuclear EM pulse (NEPM). A schematic depicting spectrum users and EM threats adapted from [11] is shown in Fig. 1. The basic phenomenology of the lightning discharge is the attachment of an ionised channel to a structure and hence the transfer of large amounts of electric charge. The resulting pulsed high currents have a fast rise-time of the order of 1 μs and decay time approximately 50 μs. Several such discharges may take place in rapid succession. In terms of the response of engineering systems to LEMP of importance are the attachment point (in cases of a direct hit) and the pulse shape, peak value and repetition rate of the pulse train. In order to design and test systems which are able to function properly under LEMP threats standards specify pulse shapes which reproduce as far as possible actual LEMP threats as shown in Fig. 2. The current waveform for each segment A to D is specified in detail in [12]. Facilities are available in major industries and agencies which permit the injection of the appropriate current shapes for testing and validation purposes. A particular concern in recent years is the widespread introduction (especially in airborne systems) of carbon fibre composites (CFCs) in preference to metals as this offers weight savings and good mechanical strength. From the EMC point of view, however, this creates difficulties as CFCs afford limited EM shielding compared to metals. Moreover, the laminated, layered, anisotropic nature of CFC panels, electrical conditions and current sharing at junctions and at riveted joints make it difficult to predict current flows under lightning strike conditions. Whole system modelling under these conditions is very challenging and the subject of continuing research and development work. In case of a nuclear detonation (especially a

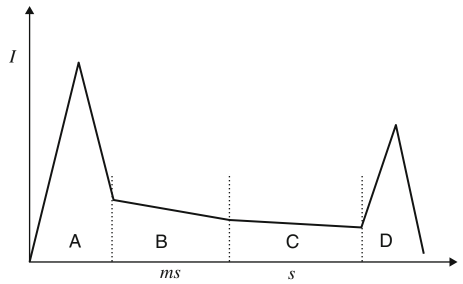


Fig. 2 Schematic of lightning test waveforms. Detailed descriptions of components A-D may be found in [12]

high-altitude one) in addition to the physical and other phenomena, an EM wave is established driven by currents generated by the impact of γ-rays on matter and the production of fast electrons (Compton Effect). The pulsed wave has typically a rise-time of a few nanoseconds and decay time of the order of μs and a peak electric field of the order of 50 kV/m [13]. Such a broad spectrum and high intensity field propagating over distances on a continental scale can severely impact on important infrastructure and military assets to such a degree as to make any defensive or offensive response problematic. Thus important systems have to be designed to cope with such an eventuality. The spectral densities of LEMP and NEMP threats are discussed in more detail in [14]. Material on high power electromagnetics and intentional EMI (IEMI) in general may be found in [15, 16]. The threat of IEMI to infrastructure is addressed in [17–19].

Coupling Paths

Coupling of EM radiation to potential victims can take place in several ways. At low-frequencies and typically below 30 MHz EMI enters systems through wire or other conducting penetrations (mains, control and communication cables, Cu water pipes, structural steel work, etc.). At higher frequencies, penetration takes place either in wire guided mode or through space much like the radiation from antenna. In the former case we talk about *conducted interference* and in the latter about *radiated interference*. Radiated interference can also be distinguished into near-field and far-field coupling. In the former case coupling can be understood in terms of mutual capacitance and inductance and in the latter as true radiative coupling as experienced far from a radiating antenna. The number of coupling paths can be large and sometimes unpredictable in large complex systems and care must be taken to establish the critical paths as any potential remedy will depend critically on the exact nature of coupling. It is not possible in a short article to paint a comprehensive picture of all coupling mechanisms but in order to sensitise readers to the issues involved we will give some illustrative examples.

Wires, cables, tracks, in close proximity to each other are electromagnetically coupled through mutual inductance and capacitances (cross-talk). As the frequency increases, coupling gets more complicated and transmission line behaviour dominates when the electrical length of the wiring is comparable to the wavelength (electrically long wires). At even higher frequencies, wire-to-wire coupling resembles that between antennas and requires a full-field treatment based on solving Maxwell's equations. Central to the understanding of wire-borne interference is the idea of differential-mode (DM) and common-mode (CM) signals. The concept applies to any multi-conductor systems the simplest case being between two conductors. Here the current in each conductor is equal and opposite (DM currents). However, in practice such a system strictly does not exist—there are always conducting objects in the proximity, e.g. ground, chassis, etc. At high-frequencies in particular, the currents on the two conductor are not equal and opposite but an additional current component shared equally and flowing in the same direction in the two wires is present (CM current). This current "returns" through conducting objects in the proximity of the wires and its origin can be simply understood at least at low-frequencies in terms of current flows from the two conductors through stray inductances and capacitance to the nearby objects. The situation is depicted in Fig. 3.

The total current flowing in a wiring system is made up of two modes for two wires and a third object [4].

$$
\begin{aligned}
I_1 &= I_d + I_c \\
I_2 &= -I_d + I_c
\end{aligned}
$$

The common- and differential-mode currents I_c and I_d, respectively, may be obtained in terms of the total currents from,

$$
\begin{aligned}
I_c &= (I_1 + I_2)/2 \\
I_d &= (I_1 - I_2)/2
\end{aligned}
$$

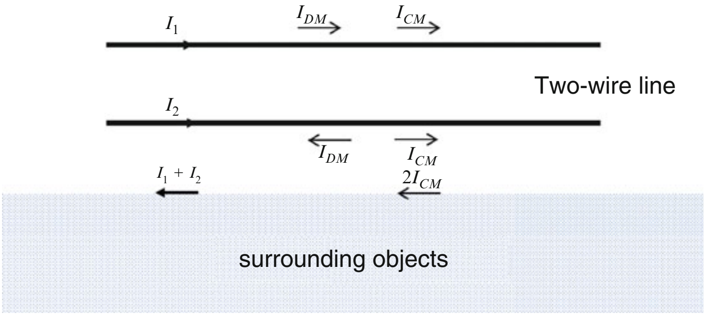


Fig. 3 The various current components in a two-wire transmission line in the proximity of other objects (e.g. equipment cabinet)

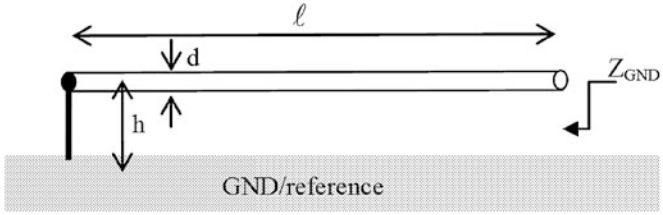


Fig. 4 A simple grounding arrangement to illustrate the nature of the grounding impedance at various frequencies

In normal operation, the DM currents are the ones required by design and the CM currents are the result of unavoidable and undesirable stray coupling to other circuits. Moreover, radiation (emission) from wiring is primarily due to CM currents (antenna mode) as there is partial cancellation of radiation from the two opposing DM currents. No understanding of EMC problems caused by emission from wires can be reached without recognising and controlling the flow of common mode currents. Generally speaking, the task of the EMC engineer is not simply to cope with the DM currents but more importantly to control the flow of stray currents, return CM currents, grounding wire currents, etc. As these are not normally explicitly identified or calculated during design it is imperative that they form part of a thorough EMC study.

For the purposes of further illustration, we address here the problem of grounding. We attempt to estimate the magnitude of the grounding impedance at the point of connection of the grounding wire to the equipment. Without loss of generality and in order to facilitate calculation we idealise somewhat the grounding configuration as shown in Fig. 4. We estimate the grounding impedance at low- medium- and high-frequencies. At LF the grounding impedance consists essentially of the resistance of the grounding wire which is normally very low of the order of $m\Omega$. As the frequency of the currents flowing in the grounding wire rises the inductance of the wire contributes to a voltage drop and hence its impact begins to dominate. The situation for these two cases (LF and MF) is described by equations,

$$Z_{gnd} \simeq R$$
$$Z_{gnd} \simeq R + j\omega L$$

At even higher frequencies, when the length of the grounding wire becomes comparable to the wavelength the grounding assembly resembles more to a transmission line (a cylindrical wire above ground in our case) and hence the grounding impedance is that of a short-circuited transmission line as shown below,

$$Z_{gnd} = jZ_0 \tan(2\pi\ell/\lambda)$$

where Z_0 is the characteristic impedance of the line and ℓ/λ is the ratio of the wire length to the wavelength. We see from this expression that the grounding impedance varies greatly with this ratio and can assume very high values when the ratio is equal to 0.25! Such a high grounding impedance will naturally cancel out any benefits or protection the grounding system is meant to provide. It is therefore imperative for EMC to look carefully at all stray current flows, including grounding currents, as they can affect greatly system EMC response especially at HF. Particular care is needed in the grounding of electrically large installations where instead of single grounding wires a wire grid with a periodicity a fraction of the wavelength at the highest frequency of interest is required to avoid resonances as suggested by the discussion above. It must also be emphasised that components sharing a common grounding structure will suffer from cross-talk through potential fluctuation at the point of connection to the grounding structure.

More information on the different types of coupling may be found in the literature [1, 4, 9, 20].

Victims of Interference

EMI may affect the operation of integrated circuits (ICs) by excessive cross-talk and thus alteration in logic states, or in more severe cases by direct damage (electrical breakdown across solid-state junctions). EMC designers aim to minimise interference and add suitable protections so damage is limited except in exceptional circumstances. It is also the responsibility of general functional design to implement defensive measures so that in the case of interference systems fail in a safe mode and are able to reset/restart automatically if this is appropriate for the safe operation of the system. Proper software design minimises the impact of interference and hence enhances immunity. Guidelines and design practices to enhance the immunity of ICs may be found in [21, 22].

Detailed information on emission and immunity levels of various components and systems may be found in the appropriate standards but it is useful to give some general guidance on interference levels which are likely to cause problems. The relevant parameters of interfering signals are frequency, bandwidth, magnitude and polarisation.

A reasonable expectation is that electric field levels of the order of 100 V/m may cause upsets and permanent damage can be expected at the 10 kV/m level. These are, however, ball park numbers the exact values depending on technology used, design and construction practices. Coupling to systems through airborne interference is likely to be strongest when the wavelength of the incident radiation is comparable to the physical dimension of the potential victim as in this case we can expect resonances. Common control equipment are handheld hence of the order of 10 s of cm in size. The wavelength in air at 1 GHz is 30 cm hence the frequencies most likely to couple in a resonant manner to handheld systems are of the order of 1 GHz. For conducted interference taking account of the typical lengths of wires we can expect maximum coupling in the 10 s of MHz region. It appears therefore that if we wish to cause maximum interference and even damage we need to employ frequencies which resonate with the potential victim. However, in general, we cannot be certain of the resonances of potential victims and hence we cannot select with reasonable certainty the most appropriate frequency for maximum interference. In such cases we may prefer a broadband interference signal hoping that we will hit some system resonances over the frequency range. However, we can launch much higher intensity interference signals at a single frequency rather than over a broad range of frequencies. Some typical examples of narrow- middle- and broad-band signals are shown in Fig. 5.

We see that a narrow pulse in the time-domain gives a broad spectrum in the frequency domain and thus a better chance of hitting a resonance, albeit at a lower intensity level. Portable equipment for the launching of hostile signals have been developed and some are described in the literature [15, 16].

Shielding, Non-linear Protection and Filtering

In military systems and in high-value important civilian infrastructure systems it is necessary to incorporate sufficient protection and countermeasures against EM interference whether unintentional or hostile. The extent of measures taken will depend on the importance of the system to be protected the consequences of failure and the likelihood of an EM hostile environment. In most systems, a degree of EM shielding is considered an essential design measure and we therefore start with a description of the fundamentals of shielding. We then briefly address a more complete protection arrangement against severe threats.

EM Shielding

A highly conducting enclosure with no apertures or wire penetrations is a very effective EM shield at all frequencies except for low-frequency magnetic fields [1]. In practice, all enclosures will have wire penetrations, some apertures and may not be made of perfect conductors. We deal in this section with perfectly conducting enclosures with apertures and in the next with protection from wire penetrations. The case of enclosures made of materials which are not highly conducting (e.g. CFCs) although important cannot be addressed in this article.

A schematic of the basic shielded enclosure with an aperture is shown in Fig. 6a. This canonical configuration may represent a vehicle, aircraft or missile structure and it allows us to illustrate in an elegant and efficient way the main attributes of shielding. We consider the case of an EM wave incident on the enclosure with the aperture. For simplicity we assume this to be a plane wave incident at a right angle on the front face of the enclosure as shown in Fig. 6a. The first issue to address is

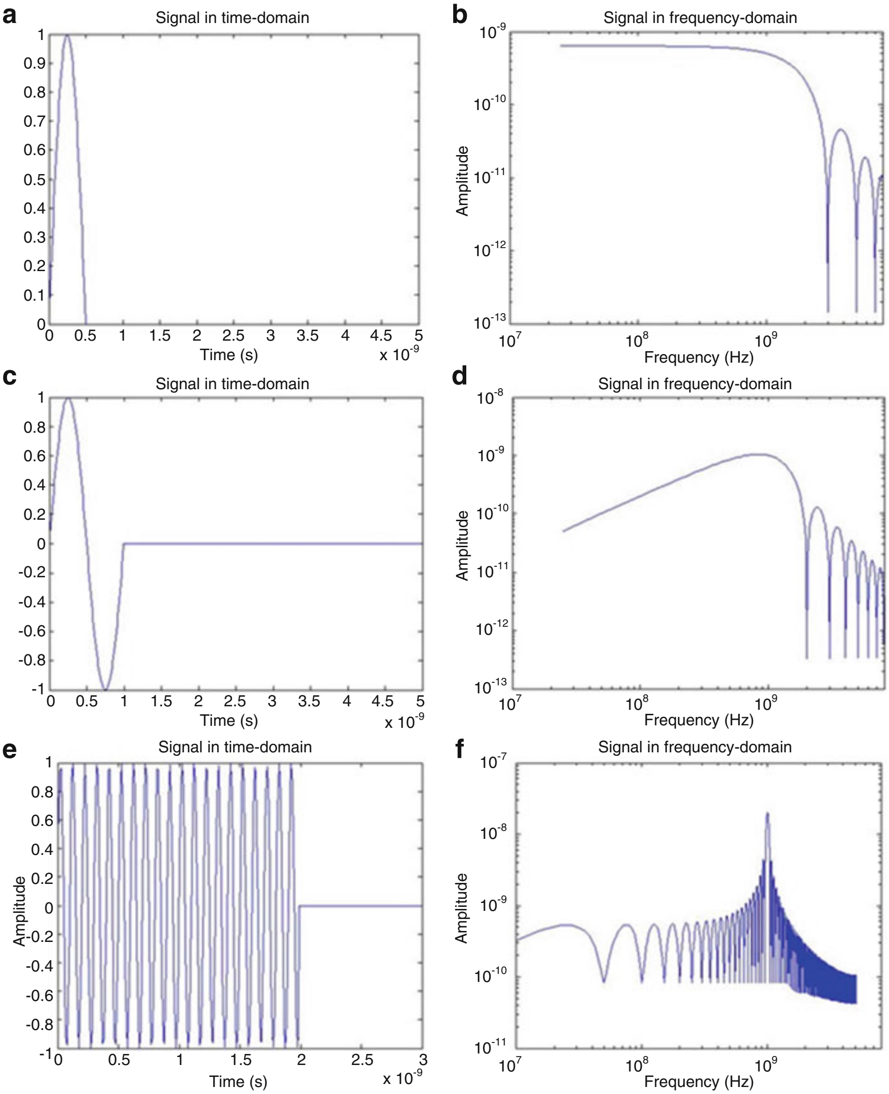


Fig. 5 Half-cycle of a 1 GHz sinusoidal signal (**a**) and its spectrum (**b**). A full-cycle 1 GHz sinusoidal signal (**c**) and its spectrum (**d**). Twenty cycles of a 1 GHz sinusoidal signal (**e**) and its spectrum (**f**)

the worst case of electric field polarisation for EMI penetration. The two extreme cases are for the electric field vector to be parallel or vertical to the long side of the narrow aperture. The answer, although perhaps counterintuitive, is straightforward from EM theory. Penetration is easiest if the electric field vector is perpendicular to the long side of the aperture. For the alternative parallel case boundary conditions

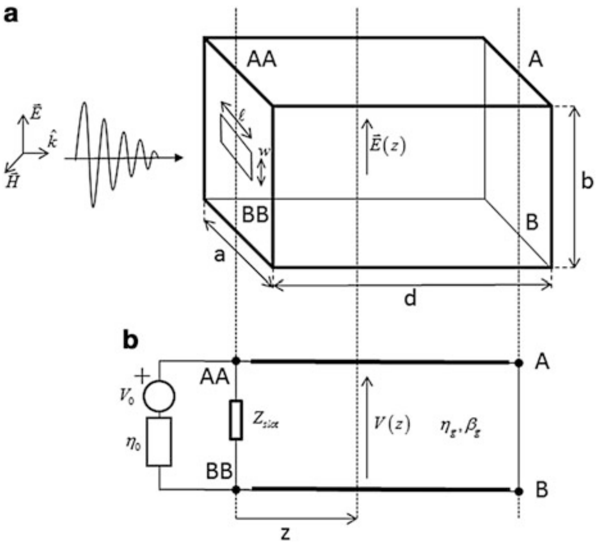


Fig. 6 Typical arrangement of a cabinet with an aperture (**a**) and its equivalent circuit for calculating shielding effectiveness (**b**)

dictate that the electric field parallel to a perfectly conducting boundary must be zero hence over the narrow aperture the electric filed is forced almost to zero. We therefore focus on the worst case shown in Fig. 6a.

The EMC designer is primarily interested in the estimation of the Shielding Effectiveness (SE) of the enclosure namely the ratio (in dB) between the electric field at a point inside the enclosure with and without the enclosure. In effect, the SE is a measure of how effective the enclosure is in reducing the penetration of an external field into the enclosure. A full rigorous analytical solution to such problems is not possible, instead, full-field numerical solutions are required [23]. However, in order to illustrate the main aspects of shielding we describe here a simplified approach based on representing the enclosure by an equivalent transmission line (TL) as shown in Fig. 6b [24, 25]. The electric field inside the enclosure is represented by the voltage across the TL and thus the shielding effectiveness, calculated at a distance z from the aperture, may be obtained from,

$$SE = 20\log\left(\frac{V_0}{2V(z)}\right)$$

In the equivalent TL the two parameters are,

$$\eta_g = {}^{\eta_0}\Big/\sqrt{1-\left({}^{\lambda}\!/_{2a}\right)^2}$$

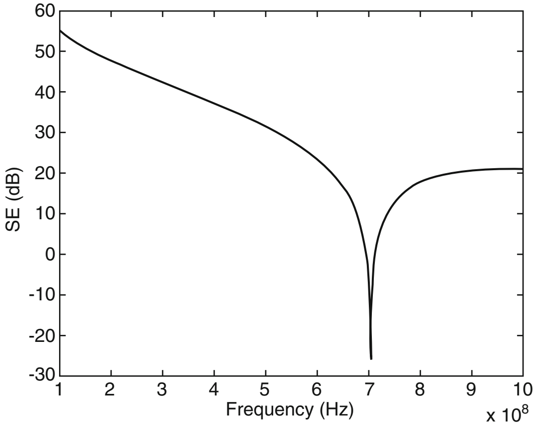


Fig. 7 The shielding effectiveness of the cabinet in Fig. 6a using the equivalent in Fig. 6b. a = 0.3 m, b = 0.12 m, d = 0.3 m, l = 0.1 m, w = 5 mm

$$\beta_g = \beta_0 \sqrt{1 - \left({}^{\lambda}\!/_{2a}\right)^2}$$

An example of an SE calculation using this model is shown in Fig. 7. The following conclusions may be drawn from this figure:

1. The SE is strongly frequency dependent and generally decreases with increasing frequency.
2. At certain frequencies (around 700 MHz in our example) the SE may be negative indicating the enclosure makes matters worse. These frequencies correspond to enclosure resonances and hence users must avoid exposure to signals with substantial energy at these specific frequencies.
3. Although the SE at one location inside the enclosure is shown in Fig. 7 it is clear by the nature of the model that SE also depends on the position inside the enclosure thus suggesting avoiding the placing of sensitive equipment at locations of low SE.

In [26] further examples are shown, including comparisons with measurements, and in the presence of PCBs as loads inside the enclosure. These indicate that if the enclosure is loaded, then the drop in SE at resonances is less pronounced or even disappears. This merely means that EM energy from the interfering signal is absorbed by the PCB and thus further investigation is needed to establish the consequences on PCB normal operation of this absorption. Several refinements to this simple model have been made to increase its utility and allow rapid and inexpensive assessments of SE at the early stages of design. Details are available in the literature.

Non-linear Protection, Filtering

In this section we aim to give a broad outline of the range of measures that need to be taken to protect important assets against severe EM threats.

Severe threats involve potentially high amounts of EM energy (high voltages and currents), fast rise-times and long durations. Most protective measures are particularly suited to address one of these features and alone cannot cope with the entire range of potential threats. Some protection is suited to high energy signals, other to fast rise-time signal, etc. Thus, we need a staged protection each stage aimed at achieving a particular threat reduction. A schematic of a typical arrangement is shown in Fig. 8. The main principle here is the removal of EMI on entry to the protected area and the binding together of ground/reference conductors. Power signals are initially attenuated by a combination of non-linear resistors and lightning arresters as shown in Fig. 9a. This arrangement removes high-voltage peaks but is not able to deal with fast rise-time transients which require solid-state devices (zener diodes) as shown in the same figure. The resulting signal which has now most of the EMI energy removed pass through a filter which removes common-mode signals as shown in Fig. 9b. A similar arrangement for data signals based on a combination of solid-state (for fast signals) and lightning arresters (for slow high-energy signals) is shown in Fig. 9c.

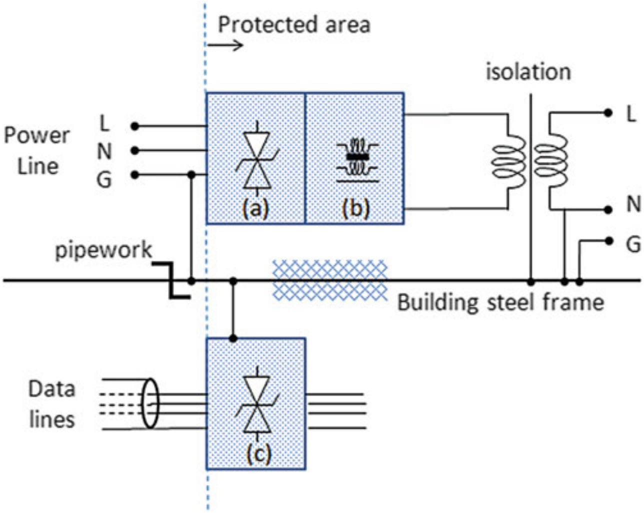


Fig. 8 Schematic of protection arrangements against severe threats

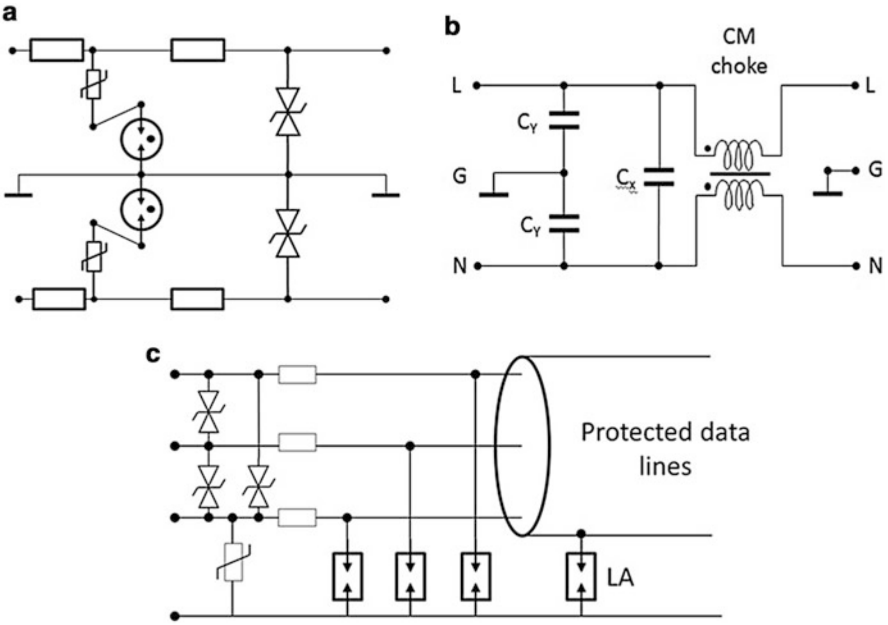


Fig. 9 (**a**) Clipping of high-voltage signals with non-linear resistors and gas discharge devices. The remaining fast rise-time signals are absorbed by zener diodes. (**b**) Common-mode suppression filter. (**c**) Data lines are protected by zener diodes. Overvoltage signals to common are attenuated by lightning arresters

Analysis and Design Tools: Simulation and Testing

The current trends for rapid, right-first-time designs and the complexity of modern systems make traditional design techniques cumbersome and instead the use of numerical computer-aided techniques is much favoured. Numerical models of physical systems are thus created in software and provided they incorporate correctly the relevant physics of the system they can be used to predict behaviour and test design modifications well in advance of the production of a prototype [27, 28]. There are, however, those who believe that a numerical model cannot represent fully all aspects of system behaviour and that actual physical testing is required. The view advocated in this paper is that both approaches are needed and that the best guarantee of successful design is the complementary use of both thus maximising their advantages and minimising their disadvantages. Tables 1 and 2 summarise the advantages and disadvantages of both approaches. It can be seen that numerical techniques can alert the designer to potential EMC problems very early in the design stage thus offering a wide scope for remedies. But, there is no guarantee that a numerical model encompasses all relevant interactions, e.g. an imperfection

Table 1 Advantages and disadvantages of testing

Many aspects of the test environment are difficult to isolate and control (resonances and damping of screened rooms, spurious reflections in open-are test sites, etc.)
Proximity effects affect the calibration of measurement antennas
Extrapolations from near- to far-field are not always correct
Very large objects require very expensive test facilities
Testing is only possible when a prototype is available by which time some design changes are not possible or too expensive
Some experiments are too costly (e.g. inaccessible places) or dangerous to be performed
Testing can reveal inadvertent errors and omissions in manufacture and assembly

Table 2 Advantages and disadvantages of numerical modelling

Easy to isolate design factors and study the sensitivity of the design to various parameters ("what is" experiments)
Possible to study response well before a prototype is available when design changes are easier to make
Can have full diagnostics even in remote and hostile places
Can do studies which in real life would be too dangerous or costly to do
It is easy to overlook significant factors or to miss manufacturing problems
Can be overwhelmed with information!

during manufacture or assembly and thus actual physical testing is always desirable. Testing on its own, however, reveals problems very late in the design when the options for re-design are limited and very costly.

There is naturally a desire to minimise testing as it is very time-consuming and costly. Moreover, in modern complex systems (e.g. military aircraft with several miles of wiring, connectors, etc.) it is virtually impossible to incorporate diagnostics at thousands of inaccessible points. Numerically, however, this is not a problem. Part of testing therefore, in addition of demonstrating compliance with standards, is to validate the numerical model. If measurements at a few selected critical points agree with the model predictions there is then a reasonable expectation that the model operates correctly and the rest of its predictions can be accepted with confidence. Thus the testing cycle is reduced in duration and cost with obvious benefits. A further advantage is that with a validated numerical model any future modifications (e.g. introduction of a new component or sub-system) need only be tested in software thus avoiding the cost and complexity of further physical testing.

Uncertainty and Lifetime Compliance

Most engineers most of the time are exposed to problems of *deterministic* type. This means that all the parameters of the system are precisely known and under these conditions the response to a known stimulus is required. In EMC this typically means that all system geometrical and material details are known, we assume an incident EM pulse (the threat) and we wish to calculate the voltage at a critical pin to determine whether there will be a failure or not. In real life, however, this situation is rarely the case. Full geometrical details are rarely known and vary somewhat from one batch to another. The same is true for material properties and even for the EM threat which may have uncertain polarisation, rise-time, etc. We are then asked to calculate the response of a system where several parameters are not fixed but typically have a mean value and a spread around this mean. In the language of mathematics the system has a number of random variables and hence the response too is a random variable. These types of problem are termed as *stochastic*. It is clearly important to be able to address stochastic problem and not merely deterministic ones. Even if we start with a deterministic system, through its lifetime parameters will change (e.g. component tolerances, ageing of components, replacement of parts with new "equivalent" replacement components, etc.). We need to be able to assure the performance of the system throughout its lifetime and for a range of eventualities brought about by ageing, deterioration in service, selective introduction of new parts and new technologies etc. A stochastic model of the system helps us to minimise further testing provided any parameter changes remain within the statistical envelope used in the original calculation. This offers economy and confidence in design.

A first approach to a stochastic model is to do a number of trials (simulations) of the system with a range of parameters taken from the spread of the system random variables. In this way we will eventually build up the statistics of the response and thus the mean, standard deviation, etc., of the response. This is the so-called Monte Carlo approach and requires typically 1000s of trials. Since in EMC studies each trial requires running a complete simulation this represents an immense simulation task which in most cases is unrealistic. It is, however, possible to obtain statistical information on system response by performing a small number of trials and combining the results of the calculated responses, with appropriate weight factors, to obtain the mean value, variance and other higher moments of the response random variable [29, 30]. The complete probability density function of the response may also be reconstructed in this way. These statistical techniques are similar to the approximations of an integral from a few selected values of the integrand over the interval of integration (Gaussian Quadrature). For problems with one Gaussian random variable $x = \bar{X} + \tilde{x}$ where $\bar{X}$ is the mean value of the random variable and $\tilde{x}$ is a zero-mean random variable, three evaluations of the response g(.) are required to obtain the first two moments,

$$\bar{\tilde{g}} = E\left\{g\left(\bar{X} + \hat{x}\right)\right\} = \frac{2}{3}g\left(\bar{X}\right) + \frac{1}{6}g\left(\bar{X} + \sigma\sqrt{3}\right) + \frac{1}{6}g\left(\bar{X} - \sigma\sqrt{3}\right)$$

$$\tilde{\sigma}_g^2 = \frac{2}{3}\left[g(\bar{X}) - \bar{\tilde{g}}\right]^2 + \frac{1}{6}\left[g(\bar{X} + \sigma\sqrt{3}) - \bar{\tilde{g}}\right]^2 + \frac{1}{6}\left[g(\bar{X} - \sigma\sqrt{3}) - \bar{\tilde{g}}\right]^2$$

The three evaluations points and corresponding weight factors are shown in Table 3. In the case of say two random variables nine evaluations are required.

Table 3 Derivation of the first two moments (mean and variance) of a random output using responses at three values of the single random input and corresponding weight factors

Evaluation points	$\bar{X} - \sigma\sqrt{3}$	$\bar{X}$	$\bar{X} + \sigma\sqrt{3}$
Weights	1/6	2/3	1/6

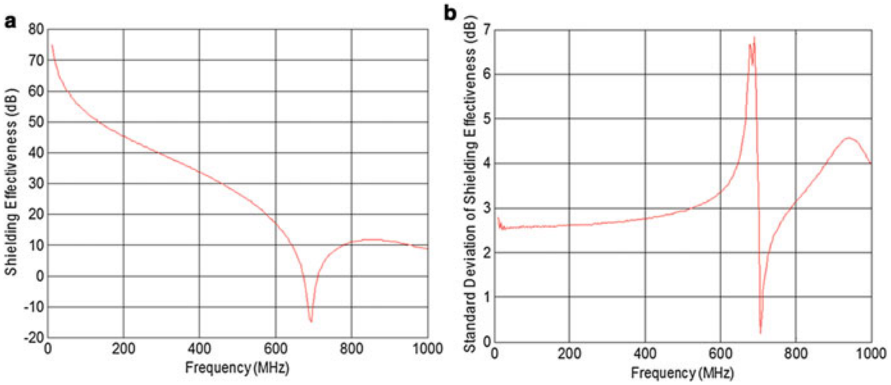


Fig. 10 The mean (**a**) and standard deviation (**b**) of the shielding effectiveness of a cabinet with two parameters which are Gaussian random variables

As the number of random variables increases then the number of evaluations also increases rapidly but special sampling techniques can be used to reduce the number of evaluations to realistic levels at the expense of prediction accuracy [31]. An example for the mean value and variance of the shielding effectiveness (SE) of a cabinet is shown in Fig. 10 [32]. Here, the statistics of the SE are obtained over a wide frequency range from nine evaluations of the two random variables (length and width of the aperture).

Conclusions

An assessment of some of the most severe EM threats and possible mitigation techniques has been presented in this paper.

The nature of EM interactions requires a deep understanding of source regions, coupling paths and potential victims, so that the most effective and inexpensive measures are adopted to minimise the possibility of EMC problems. This requires a wide grasp of many electrical engineering disciplines (electromagnetics, signals, communications, power electronics) but also areas such as heat transfer which can make competing demands on design.

It has been proposed that a combination of numerical models, simulations and experimental testing techniques offer the best approach to solving EMC related problems. It is important that designers exploit fully the capabilities offered by each approach and their synergies. A hierarchy of models is available to address all stages in the design which together with physical testing can give confidence in the analysis and synthesis of systems. Models should be used intelligently to aid understanding and support the creative thinking of designers. It is also important to appreciate that strictly speaking all problems are ultimately stochastic and therefore techniques should be developed which can address efficiently this aspect of EMC.

Designing systems which are initially compliant with EMC requirements is only one important step. It is also necessary that EMC is managed through the entire expected lifetime of systems by putting in place measures to check that EMC is maintained. Paramount in this respect is to educate users, and maintenance personnel on the principles of EMC and the manner in which designed features which on first reflection do not affect normal operations have nevertheless important EMC implications.

References

1. C. Christopoulos, *Principles and Techniques of Electromagnetic Compatibility*, 2nd edn. (CRC Press, Boca Raton, 2007)
2. TEMPEST, spying on information systems through leaking emanations, including unintentional radio or electrical signal. https://en.wikipedia/wiki/Tempest_(codename). Accessed 29 July 2015
3. W. van Eck, Electromagnetic radiation from video display units: an eavesdropping risk? Comput. Secur. **4**, 269–286 (1985)
4. C.R. Paul, *Introduction to Electromagnetic Compatibility*, 2nd edn. (Wiley, New Jersey, 2006)
5. J.A. Rosero et al., Moving towards a more electric aircraft. IEEE Aerosp. Electron. Syst. Mag. **22**(3), 3–9 (2007)
6. C. Champion, Towards more electric aircraft. Skyline Clean Sky Newsl. **6**, 4–5 (2012)
7. Department of Defence, Washington DC, Requirements for the Control of Electromagnetic Emissions and Susceptibility. MIL-STD 461D (1993)
8. Ministry of Defence, Glasgow, UK, Electromagnetic Compatibility. DEF-STAN 59–41 (1988)
9. F.M. Tesche, M.V. Ianoz, T. Karlsson, *EMC Analysis Methods and Computational Models* (Wiley, New York, 1997)
10. F. Leferink, Conducted interference, challenges and interference cases. IEEE EMC Mag. **4**(1), 78–85 (2015)
11. D.V. Giri, F.M. Tesche, Classification of intentional electromagnetic environments. IEEE Trans. EMC **46**(3), 322–328 (2004)
12. EM Environmental Effects Requirements for Systems. MIL-STD-464 (1997)
13. C.L. Longmire, On the electromagnetic pulse produced by nuclear explosions. IEEE Trans. EMC **29**, 3–13 (1978)
14. R.L. Gardner et al., Comparison of lightning and public domain HEMP waveforms on the surface of an aircraft, in *6th International Zurich Symposium on EMC*, pp. 175–180 (1985)
15. D.V. Giri, *High-power Electromagnetic Radiators: No-lethal Weapons and Other Applications* (Harvard University Press, Cambridge, 2004)
16. W.A. Radasky et al., Introduction to the special issue on high-power electromagnetics (HPEM) and intentional electromagnetic interference (IEMI). IEEE Trans. EMC **46**(3), 314–321 (2004)

17. O.-H. Arnesen, R. Hoad, Overview of the European project 'HIPOW'. IEEE EMC Mag. **3**(4), 64–67 (2014)
18. Van de Beek et al., Overview of the European project STRUCTURES. IEEE EMC Mag. **3**(4),70–79 (2014)
19. V. Deniau, Overview of the European project security of railways in Europe against EM attacks (SECRET). IEEE EMC Mag. **3**(4), 80–85 (2014)
20. K.S.H. Lee (ed.), Interaction Notes: Principles, Techniques and Reference Data. Report AFWL-TR-80-402 (1980)
21. J.-M. Redoute, A. Richelli, A methodological approach to EMI resistant analog integrated design. IEEE EMC Mag. **4**(2), 66–74 (2015)
22. F. Fiori, EMI susceptibility: the Achilles' heel of smart power ICs. IEEE EMC Mag. **4**(2), 75–79 (2015)
23. C. Christopoulos, *The Transmission-Line Modeling Method TLM* (IEEE Press, New York, 1995)
24. M.P. Robinson et al., Shielding effectiveness of a rectangular enclosure. Electron Lett. **32**(19), 1559–1560 (1996)
25. M.P. Robinson et al., Analytical formulation of the shielding effectiveness of enclosures with apertures. IEEE Trans. EMC **40**(3), 240–248 (1998)
26. D.W.P. Thomas et al., Characterization of the shielding effectiveness of loaded equipment cabinets, in *IET Conf Publ, EMC York, 99*, pp. 89–94 (1999)
27. C. Christopoulos, Modeling and simulation for EMC-part I. IEEE EMC Mag. **4**(1), 47–56 (2015)
28. C. Christopoulos, Modeling and simulation for EMC-part II. IEEE EMC Mag. **4**(2), 63–72 (2015)
29. L. De Menezes et al., Efficient computation of stochastic electromagnetic problems using unscented transforms. IET Sci. Meas. Technol. **2**(2), 88–95 (2008)
30. D.W.P. Thomas et al., Estimation of the probability distributions for cable coupling using unscented transforms. Ann. Telecommun. **66**, 475–482 (2011)
31. I. Lee et al., Dimension reduction method for reliability-based robust design optimization. Comput. Struct. **86**(13–14), 1550–1562 (2008)
32. L. De Menezes, D.W.P. Thomas, C. Christopoulos, Statistics of the shielding effectiveness of cabinets, in *Proceedings of the ESA Workshop on Aerospace EMC*, Florence, Italy, 6 p. (2009)

Cybersecurity Investments with Nonlinear Budget Constraints: Analysis of the Marginal Expected Utilities

Patrizia Daniele, Antonino Maugeri, and Anna Nagurney

Abstract In this paper, we consider a recently introduced cybersecurity investment supply chain game theory model consisting of retailers and consumers at demand markets with the retailers being faced with nonlinear budget constraints on their cybersecurity investments. We construct a novel reformulation of the derived variational inequality formulation of the governing Nash equilibrium conditions. The reformulation then allows us to exploit and analyze the Lagrange multipliers associated with the bounds on the product transactions and the cybersecurity levels associated with the retailers to gain insights into the economic market forces. We provide an analysis of the marginal expected transaction utilities and of the marginal expected cybersecurity investment utilities. We then establish some stability results for the financial damages associated with a cyberattack faced by the retailers. The theoretical framework is subsequently applied to numerical examples to illustrate its applicability.

Keywords Cybersecurity • Investments • Supply chains • Game theory • Nash equilibrium • Variational inequalities • Lagrange multipliers • Stability

MSC 2010: 49K40, 65K10, 65K15, 90C33, 90C46.

Introduction

Cybercrime is a major global issue with cyberattacks adversely affecting firms, governments, other organizations, and consumers [14]. For example, it has been estimated that cyberattacks cost firms $400 billion annually [22]. In a recent study

P. Daniele (✉) • A. Maugeri
Department of Mathematics and Computer Science, University of Catania, Catania, Italy
e-mail: daniele@dmi.unict.it

A. Nagurney
Isenberg School of Management, University of Massachusetts, Amherst, MA 01003, USA

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_6

[19] that surveyed 959 top executives in such industries as banking, insurance, energy, retail, pharmaceuticals, healthcare, and automotive, it was found that 63% reported that their companies experienced significant attacks daily or weekly. Cyberattacks can result not only in direct financial losses and/or the loss of data, but also in an organization's highly valued asset—its reputation. It is quite understandable, hence, that worldwide spending on cybersecurity was approximately $75 billion in 2015, with the expectation that, by 2020, companies around the globe will be spending around $170 billion annually (see [12]).

Organizations, as noted by Ostvold and Walker [19], are part of ecosystems and the decisions that they make individually, including those in terms of cyberinvestments, may affect other organizations. Indeed, as discussed in [16], who developed a supply chain game theory model for cybersecurity investments, the level of a cybersecurity investment of a retailer may affect not only his vulnerability to cyberattacks but also that of the network of the supply chain consisting of retailers and consumers who engage in electronic transactions. Effective modeling of the complexity of cyberattacks and cybersecurity investments using operations research techniques, including game theory, can assist in the analysis of complex behaviors and provide, ultimately, tools and insights for policymakers.

For example, [13] developed a multiproduct network economic model of cybercrime with a focus on financial services, since that industrial sector is a major target of cyberattacks. The model captured the perishability of the value of financial products to cybercriminals in terms of the depreciation in prices that the hacked products command over time in the black market. Nagurney and Nagurney [14], subsequently, constructed a supply chain game theory model in which sellers maximize their expected profits while determining both their product transactions with consumers and their cybersecurity investments. However, network vulnerability was not captured. Nagurney et al. [16] then showed how the model in [14] could be extended to quantify and compute network vulnerability. The studies [14] and [16] were inspired, in part, by the contributions in [20]. The supply chain game theory network framework of [14] and [16] is, nevertheless, more general than that of [20] since the firms, which are retailers, are not assumed to be identical, and the demand side for products of the supply chain network is also captured. In addition, the firms can have distinct cybersecurity investment cost functions and are faced with distinct damages, if attacked. Such features provide greater modeling flexibility as well as realism.

More recently, [15], building on the prior supply chain network cybersecurity investment modeling and analysis work noted above, introduced a novel game theory model in which the budget constraints for cybersecurity investments of retailers, which are nonlinear, are explicitly included, and conducted a spectrum of sensitivity analysis exercises. Consumers reflect their preferences for the product through the demand price functions, which depend on the product demands and on the average security of the network. The methodology utilized for the formulation, analysis, and solution of the game theory models in [13, 14, 16], and [15] was that of the theory of variational inequalities. We refer the reader to [11] for a survey of game theory, as applied to network security and privacy, and to [9] for some background on optimization models for cybersecurity investments. For a collection of papers on cryptography and network security, see the edited volume [6].

In this paper, we return to the cybersecurity investment supply chain game theory model with nonlinear budget constraints of [15]. We provide an alternative formulation of the variational inequality derived therein in order to provide a deeper qualitative and economic analysis with a focus on the Lagrange multipliers associated with the constraints. The constraints in the model in [15] include not only the nonlinear budget constraints but also lower and upper bounds on the cybersecurity levels as well as on the product transactions.

It is worth mentioning that a wide spectrum of papers has been devoted to the analysis of the behavior of the solutions to a variational inequality which models equilibrium problems by means of the Lagrange multipliers. For instance, we cite the papers [1, 3, 5] for the financial equilibrium problem, the paper [2] for the random traffic equilibrium problem, the papers [7, 8] for the elastic-plastic torsion problem, and the paper [4] for the unilateral problems. This paper is the first to analyze a cybersecurity investment supply chain game theory model with nonlinear budget constraints by means of Lagrange multipliers.

This paper is organized as follows. In section "The Model", we briefly recall, for completeness and easy reference, the supply chain network game theory model for cybersecurity investments with nonlinear budget constraints developed in [15] and provide the variational inequality formulation of the Nash equilibrium conditions. The model consists of retailers and consumers at demand markets with the former competing on their product transactions as well as their cybersecurity levels. In section "Equivalent Formulation of the Variational Inequality", we construct an alternative formulation of that variational inequality. We then provide an analysis of the marginal expected transaction utilities and of the marginal expected cybersecurity investment utilities. In addition, we present some stability results for the marginal expected cybersecurity investment utilities with respect to changes in the financial damages sustained in a cyberattack. Section "A Numerical Example" illustrates how the framework developed in section "Equivalent Formulation of the Variational Inequality" can be applied in the context of numerical examples. We summarize our results and present our conclusions in section "Conclusions".

The Model

We now recall the supply chain game theory model of cybersecurity investments with nonlinear budget constraints introduced in [15] (see also [21] for other equilibrium models with nonlinear constraints). The supply chain network, consisting of retailers and consumers at demand markets, is depicted in Fig. 1. Each retailer i; $i = 1, \dots, m$, can transact with demand market j; $j = 1, \dots, n$, with Q_{ij} denoting the product transaction from i to j. Also, each retailer i; $i = 1, \dots, m$, determines his cybersecurity or, simply, security, level s_i; $i = 1, \dots, m$. We group the product transactions for retailer i; $i = 1, \dots, m$, into the n-dimensional vector Q_i and then we group all such retailer transaction vectors into the mn-dimensional vector Q. The security levels of the retailers are grouped into the m-dimensional vector s.

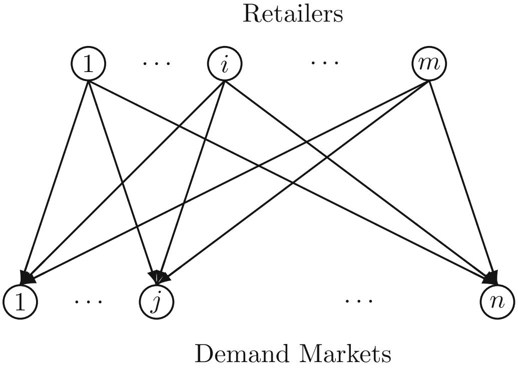


Fig. 1 The bipartite structure of the supply chain network game theory model

The cybersecurity level in the supply chain network is the average security and is denoted by $\bar{s}$, where $\bar{s} = \sum_{i=1}^{m} \frac{s_i}{m}$.

The retailers seek to maximize their individual expected utilities, consisting of expected profits, and compete in a noncooperative game in terms of strategies consisting of their respective product transactions and security levels. The governing equilibrium concept is that of Nash equilibrium [17, 18].

The demand at each demand market j, d_j, must satisfy:

$$d_j = \sum_{i=1}^{m} Q_{ij}, \quad j = 1, \ldots, n. \tag{1}$$

We group the demands at the demand markets into the n-dimensional vector d.

The product transactions are subject to upper bounds and must be nonnegative so that we have the following constraints:

$$0 \leq Q_{ij} \leq \bar{Q}_{ij}, \quad i = 1, \ldots, m; j = 1, \ldots, n. \tag{2}$$

The cybersecurity level of each retailer i must satisfy the following constraint:

$$0 \leq s_i \leq u_{s_i}, \quad i = 1, \ldots, m, \tag{3}$$

where $u_{s_i} < 1$ for all i; $i = 1, \ldots, m$. The larger the value of s_i, the higher the security level, with perfect security reflected in a value of 1. However, since, as noted in [15], we do not expect perfect security to be attainable, we have $u_{s_i} < 1$; $i = 1, \ldots, m$. If $s_i = 0$, this means that retailer i has no security.

The demand price of the product at demand market j, $\rho_j(d, \bar{s})$; $j = 1, \ldots, n$, is a function of the vector of demands and the network security. We can expect consumers to be willing to pay more for higher network security. In view of the conservation of flow equations above, we can define $\hat{\rho}_j(Q,\bar{s}) \equiv \rho_j(d,\bar{s})$; $j = 1,\ldots,n$. We assume that the demand price functions are continuously differentiable.

There is an investment cost function h_i; $i = 1,\ldots,m$, associated with achieving a security level s_i with the function assumed to be increasing, continuously differentiable and convex. 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. An example of an $h_i(s_i)$ function that satisfies these properties and that is utilized here (see also [15]) is

$$h_i(s_i) = \alpha_i \left(\frac{1}{\sqrt{(1-s_i)}} - 1 \right) \text{ with } \alpha_i > 0.$$

The term α_i enables distinct retailers to have different investment cost functions based on their size and needs. Such functions have been introduced by Shetty et al. [20] and also utilized by Nagurney et al. [16]. However, in those models, there are no cybersecurity budget constraints and the cybersecurity investment cost functions only appear in the objective functions of the decision-makers.

In the model with nonlinear budget constraints as in [15] each retailer is faced with a limited budget for cybersecurity investment. Hence, the following nonlinear budget constraints must be satisfied:

$$\alpha_i \left(\frac{1}{\sqrt{(1-s_i)}} - 1 \right) \leq B_i; \quad i = 1,\ldots,m, \tag{4}$$

that is, each retailer can't exceed his allocated cybersecurity budget.

The profit f_i of retailer i; $i = 1,\ldots,m$ (in the absence of a cyberattack and cybersecurity investment), is the difference between his revenue $\sum_{j=1}^{n} \hat{\rho}_j(Q,s)Q_{ij}$ and his costs associated, respectively, with production and transportation: $c_i \sum_{j=1}^{n} Q_{ij} + \sum_{j=1}^{n} c_{ij}(Q_{ij})$, that is,

$$f_i(Q,s) = \sum_{j=1}^{n} \hat{\rho}_j(Q,s)Q_{ij} - c_i \sum_{j=1}^{n} Q_{ij} - \sum_{j=1}^{n} c_{ij}(Q_{ij}). \tag{5}$$

If there is a successful cyberattack on a retailer i; $i = 1,\ldots,m$, retailer i incurs an expected financial damage given by

$$D_i p_i,$$

where D_i, the damage incurred by retailer i, takes on a positive value, and p_i is the probability of a successful cyberattack on retailer i, where:

$$p_i = (1 - s_i)(1 - \bar{s}), \quad i = 1, \dots, m, \tag{6}$$

with the term $(1 - \bar{s})$ denoting the probability of a cyberattack on the supply chain network and the term $(1 - s_i)$ denoting the probability of success of such an attack on retailer i.

Each retailer i; $i = 1, \dots, m$, hence, seeks to maximize his expected utility, $E(U_i)$, corresponding to his expected profit given by:

$$E(U_i) = (1 - p_i)f_i(Q, s) + p_i(f_i(Q, s) - D_i) - h_i(s_i) = f_i(Q, s) - p_i D_i - h_i(s_i). \tag{7}$$

Let $\mathbb{K}^i$ denote the feasible set corresponding to retailer i, where $\mathbb{K}^i \equiv \{(Q_i, s_i) | 0 \leq Q_{ij} \leq \bar{Q}_{ij}, \forall j,\ 0 \leq s_i \leq u_{s_i}$ and the budget constraint holds for $i\}$ and define $\mathbb{K} \equiv \prod_{i=1}^{m} \mathbb{K}^i$.

We now recall the following definition from [15]:

Definition 1 (A Supply Chain Nash Equilibrium in Product Transactions and Security Levels) A product transaction and security level pattern $(Q^*, s^*) \in \mathbb{K}$ is said to constitute a supply chain Nash equilibrium if for each retailer i; $i = 1, \dots, m$,

$$E(U_i(Q_i^*, s_i^*, \hat{Q}_i^*, \hat{s}_i^*)) \geq E(U_i(Q_i, s_i, \hat{Q}_i^*, \hat{s}_i^*)), \quad \forall (Q_i, s_i) \in \mathbb{K}^i, \tag{8}$$

where

$$\hat{Q}_i^* \equiv (Q_1^*, \dots, Q_{i-1}^*, Q_{i+1}^*, \dots, Q_m^*); \quad \text{and} \quad \hat{s}_i^* \equiv (s_1^*, \dots, s_{i-1}^*, s_{i+1}^*, \dots, s_m^*).$$

Hence, according to (8), a supply chain Nash equilibrium is established if no retailer can unilaterally improve upon his expected utility (expected profit) by choosing an alternative vector of product transactions and security level.

The following theorem was established in [15]:

Theorem 1 (Variational Inequality Formulation) *Assume that, for each retailer i; $i = 1, \dots, m$, the expected profit function $E(U_i(Q, s))$ is concave with respect to the variables $\{Q_{i1}, \dots, Q_{in}\}$, and s_i, and is continuously differentiable. Then $(Q^*, s^*) \in \mathbb{K}$ is a supply chain Nash equilibrium according to Definition 1 if and only if it satisfies the variational inequality*

$$-\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}} \times \left(Q_{ij} - Q_{ij}^*\right) - \sum_{i=1}^{m} \frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} \times \left(s_i - s_i^*\right) \geq 0,$$
$$\forall (Q, s) \in \mathbb{K} \tag{9}$$

or, equivalently, $(Q^*, s^*) \in \mathbb{K}$ *is a supply chain Nash equilibrium product transaction and security level pattern if and only if it satisfies the variational inequality*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\left[c_i + \frac{\partial c_{ij}(Q_{ij}^*)}{\partial Q_{ij}} - \hat{\rho}_j(Q^*, s^*) - \sum_{k=1}^{n}\frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial Q_{ij}} \times Q_{ik}^*\right] \times (Q_{ij} - Q_{ij}^*)$$
$$+\sum_{i=1}^{m}\left[\frac{\partial h_i(s_i^*)}{\partial s_i} - \left(1 - \sum_{k=1}^{m}\frac{s_k^*}{m} + \frac{1-s_i^*}{m}\right)D_i - \sum_{k=1}^{n}\frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} \times Q_{ik}^*\right]$$
$$\times (s_i - s_i^*) \geq 0, \quad \forall (Q, s) \in \mathbb{K}. \tag{10}$$

Equivalent Formulation of the Variational Inequality

The aim of this section is to find an alternative formulation of the variational inequality (9) governing the Nash equilibrium for the cybersecurity supply chain game theory model with nonlinear budget constraints by means of the Lagrange multipliers associated with the constraints defining the feasible set $\mathbb{K}$. To this end, we remark that $\mathbb{K}$ can be rewritten in the following way:

$$\mathbb{K} = \Big\{(Q, s) \in \mathbb{R}^{mn+n} : -Q_{ij} \leq 0,\ Q_{ij} - \overline{Q}_{ij} \leq 0,\ -s_i \leq 0,\ s_i - u_{s_i} \leq 0,$$
$$h_i(s_i) - B_i \leq 0,\ i = 1, \ldots, m,\ j = 1, \ldots, n\Big\}, \tag{11}$$

and that variational inequality (9) can be equivalently rewritten as a minimization problem. Indeed, by setting:

$$V(Q, s) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}}\left(Q_{ij} - Q_{ij}^*\right) - \sum_{i=1}^{m}\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i}\left(s_i - s_i^*\right),$$

we have:

$$V(Q, s) \geq 0 \text{ in } \mathbb{K} \text{ and } \min_{\mathbb{K}} V(Q, s) = V(Q^*, s^*) = 0. \tag{12}$$

Then, we can consider the following Lagrange function:

$$\mathcal{L}(Q, s, \lambda^1, \lambda^2, \mu^1, \mu^2, \lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}}\left(Q_{ij} - Q_{ij}^*\right)$$

$$
\begin{aligned}
&- \sum_{i=1}^{m} \frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} \left(s_i - s_i^*\right) \\
&+ \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_{ij}^1 (-Q_{ij}) \\
&+ \sum_{i=1}^{m} \sum_{j=1}^{n} \lambda_{ij}^2 (Q_{ij} - \overline{Q}_{ij}) + \sum_{i=1}^{m} \mu_i^1 (-s_i) \\
&+ \sum_{i=1}^{m} \mu_i^2 (s_i - u_{s_i}) + \sum_{i=1}^{m} \lambda_i (h_i(s_i) - B_i), \qquad (13)
\end{aligned}
$$

where $(Q, s) \in \mathbb{R}^{mn+n}$, $\lambda^1, \lambda^2 \in \mathbb{R}^{mn}_+$, $\mu^1, \mu^2 \in \mathbb{R}^m_+$, $\lambda \in \mathbb{R}^m_+$. Since for the convex set $\mathbb{K}$ the Slater condition is verified and (Q^*, s^*) is a minimal solution to problem (12), by virtue of well-known theorems (see [10]), there exist $\overline{\lambda}^1$, $\overline{\lambda}^2 \in \mathbb{R}^{mn}_+$, $\overline{\mu}^1$, $\overline{\mu}^2$, $\overline{\lambda} \in \mathbb{R}^m_+$ such that the vector $(Q^*, s^*, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda})$ is a saddle point of the Lagrange function (13); namely,

$$
\begin{aligned}
\mathcal{L}(Q^*, s^*, \lambda^1, \lambda^2, \mu^1, \mu^2, \lambda) &\le \mathcal{L}(Q^*, s^*, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda}) \\
&\le \mathcal{L}(Q, s, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda}) \qquad (14)
\end{aligned}
$$

$\forall (Q, s) \in \mathbb{K}$, $\forall \lambda^1, \lambda^2 \in \mathbb{R}^{mn}_+$, $\forall \mu^1, \mu^2, \lambda \in \mathbb{R}^m_+$ and

$$
\begin{aligned}
&\overline{\lambda}_{ij}^1(-Q_{ij}^*) = 0, \quad \overline{\lambda}_{ij}^2(Q_{ij}^* - \overline{Q}_{ij}) = 0, \quad i = 1, \ldots, m,\ j = 1, \ldots, n, \\
& \qquad (15) \\
&\overline{\mu}_i^1(-s_i^*) = 0, \quad \overline{\mu}_i^2(s_i^* - u_{s_i}) = 0, \quad \overline{\lambda}_i(h_i(s_i^*) - B_i) = 0, \quad i = 1, \ldots, m.
\end{aligned}
$$

From the right-hand side of (14) it follows that $(Q^*, s^*) \in \mathbb{R}^{mn+n}_+$ is a minimal point of $\mathcal{L}(Q, s, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda})$ in the whole space $\mathbb{R}^{mn+n}$ and, hence, for all $i = 1, \ldots, m$, and $j = 1, \ldots, n$, we get:

$$
\frac{\partial \mathcal{L}(Q^*, s^*, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda})}{\partial Q_{ij}} = -\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}} - \overline{\lambda}_{ij}^1 + \overline{\lambda}_{ij}^2 = 0 \qquad (16)
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}(Q^*, s^*, \overline{\lambda}^1, \overline{\lambda}^2, \overline{\mu}^1, \overline{\mu}^2, \overline{\lambda})}{\partial s_i} &= -\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} \\
&\quad - \overline{\mu}_i^1 + \overline{\mu}_i^2 + \overline{\lambda}_i \frac{\partial h_i(s_i^*)}{\partial s_i} = 0 \qquad (17)
\end{aligned}
$$

together with conditions (15).

Conditions (15)–(17) represent an equivalent formulation of variational inequality (9).

It is easy to see that from (16) and (17) the variational inequality (9) follows. Indeed, multiplying (16) by $(Q_{ij} - Q^*_{ij})$ we obtain:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}}(Q_{ij} - Q^*_{ij}) - \overline{\lambda}^1_{ij}(Q_{ij} - Q^*_{ij}) + \overline{\lambda}^2_{ij}(Q_{ij} - Q^*_{ij}) = 0$$

and, taking into account (15), we have:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}}(Q_{ij} - Q^*_{ij}) = \overline{\lambda}^1_{ij}Q_{ij} - \overline{\lambda}^2_{ij}(Q_{ij} - \overline{Q}_{ij}) \geq 0.$$

Analogously, multiplying (17) by $(s_i - s^*_i)$, we get:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i}(s_i - s^*_i) - \overline{\mu}^1_i(s_i - s^*_i) + \overline{\mu}^2_i(s_i - s^*_i) + \overline{\lambda}_i\frac{\partial h_i(s^*_i)}{\partial s_i}(s_i - s^*_i) = 0.$$

From (15), we have:

$$\overline{\mu}^1_i(-s^*_i) = 0, \quad \overline{\mu}^2_i s^*_i = \overline{\mu}^2_i u_{s_i}.$$

Moreover, if $\overline{\lambda}_i > 0$, then $h_i(s^*_i) = B_i = \max h_i(s_i)$, but $h_i(s_i)$ is a nondecreasing function; hence, it attains its maximum value at $s^*_i = u_{s_i}$. Therefore, we get:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i}(s_i - s^*_i) = \overline{\mu}^1_i s_i - \overline{\mu}^2_i(s_i - u_{s_i}) - \overline{\lambda}_i\frac{\partial h_i(s^*_i)}{\partial s_i}(s_i - u_{s_i}) \geq 0$$

because $h_i(s_i)$ is a nonnegative convex function such that $h_i(0) = 0$. Then $h_i(s_i)$ attains the minimum value at 0. Hence, $\frac{\partial h_i(0)}{\partial s_i} \geq 0$ and, since $\frac{\partial h_i(s_i)}{\partial s_i}$ is increasing, it results in:

$$0 \leq \frac{\partial h_i(0)}{\partial s_i} \leq \frac{\partial h_i(s_i)}{\partial s_i}, \quad \forall 0 \leq s_i \leq u_{s_i}.$$

The term $\frac{\partial E(U_i(Q^*, s^*))}{\partial Q_{ij}}$ is called the *marginal expected transaction utility*, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and the term $\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i}$ is called the *marginal expected cybersecurity investment utility*, $i = 1, \ldots, m$. Our aim is to study such marginal expected utilities by means of (15)–(17).

Analysis of Marginal Expected Transaction Utilities

From (16) we get

$$-\frac{\partial E(U_i(Q^*,s^*))}{\partial Q_{ij}} - \overline{\lambda}^1_{ij} + \overline{\lambda}^2_{ij} = 0, \quad i = 1,\dots,m,\ j = 1,\dots,n.$$

So, if $0 < Q^*_{ij} < \overline{Q}_{ij}$, then we get (see also (10))

$$-\frac{\partial E(U_i(Q^*,s^*))}{\partial Q_{ij}} = c_i + \frac{\partial c_{ij}(Q^*_{ij})}{\partial Q_{ij}} - \hat{\rho}_j(Q^*,s^*) - \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k}{\partial Q_{ij}} \times Q^*_{ik} = 0, \qquad (18)$$

$$i = 1,\dots,m,\ j = 1,\dots,n,$$

whereas if $\overline{\lambda}^1_{ij} > 0$, and, hence, $Q^*_{ij} = 0$, and $\overline{\lambda}^2_{ij} = 0$, we get

$$-\frac{\partial E(U_i(Q^*,s^*))}{\partial Q_{ij}} = c_i + \frac{\partial c_{ij}(Q^*_{ij})}{\partial Q_{ij}} - \hat{\rho}_j(Q^*,s^*) - \sum_{\substack{k=1\\k\neq i}}^{m} \frac{\partial \hat{\rho}_k}{\partial Q_{ij}} \times Q^*_{ik} = \overline{\lambda}^1_{ij}, \qquad (19)$$

$$i = 1,\dots,m,\ j = 1,\dots,n,$$

and if $\overline{\lambda}^2_{ij} > 0$, and, hence, $Q^*_{ij} = \overline{Q}_{ij}$, and $\overline{\lambda}^1_{ij} = 0$, we have

$$-\frac{\partial E(U_i(Q^*,s^*))}{\partial Q_{ij}} = c_i + \frac{\partial c_{ij}(Q^*_{ij})}{\partial Q_{ij}} - \hat{\rho}_j(Q^*,s^*) - \sum_{\substack{k=1\\k\neq i}}^{m} \frac{\partial \hat{\rho}_k}{\partial Q_{ij}} \times Q^*_{ik} = -\overline{\lambda}^2_{ij}, \qquad (20)$$

$$i = 1,\dots,m,\ j = 1,\dots,n.$$

Now let us analyze the meaning of equalities (18)–(20). From equality (18), which holds when $0 < Q^*_{ij} < \overline{Q}_{ij}$, we see that for retailer i, who transfers the product Q^*_{ij} to the demand market j, the marginal expected transaction utility is zero; namely, the marginal expected transaction cost $c_i + \dfrac{\partial c_{ij}(Q^*_{ij})}{\partial Q_{ij}}$ is equal to the marginal expected transaction revenue $\hat{\rho}_j(Q^*,s^*) + \sum\limits_{\substack{k=1\\k\neq i}}^{m} \dfrac{\partial \hat{\rho}_k}{\partial Q_{ij}} \times Q^*_{ik}$.

In equality (19), minus the marginal expected transaction utility is equal to $\overline{\lambda}^1_{ij}$; namely, the marginal expected transaction cost is greater than the marginal expected transaction revenue. Retailer j has a marginal loss given by $\overline{\lambda}^1_{ij}$.

In contrast, in case (20), in which $Q_{ij} = \overline{Q}_{ij}$ and $\overline{\lambda}_{ij}^{2} > 0$, minus the marginal expected transaction utility is equal to $-\overline{\lambda}_{ij}^{2}$; namely, the marginal expected revenue is greater than the expected transaction cost. Retailer j has a marginal gain given by $\overline{\lambda}_{ij}^{2}$.

In conclusion, we remark that the Lagrange variables $\overline{\lambda}_{ij}^{1}$, $\overline{\lambda}_{ij}^{2}$ give a precise evaluation of the behavior of the market with respect to the supply chain product transactions.

Analysis of Marginal Expected Cybersecurity Investment Utilities

From (17) we have:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} - \overline{\mu}_i^1 + \overline{\mu}_i^2 + \overline{\lambda}_i \frac{\partial h_i(s^*)}{\partial s_i} = 0, \quad i = 1, \ldots, m. \tag{21}$$

If $0 < s_i^* < u_{s_i}$, then $\overline{\mu}_i^1 = \overline{\mu}_i^2 = 0$ and we have (see also (10))

$$\begin{aligned} &\frac{\partial h_i(s_i^*)}{\partial s_i} + \overline{\lambda}_i \frac{\partial h_i(s_i^*)}{\partial s_i} \\ &= \left(1 - \sum_{k=1}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i + \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} \times Q_{ik}^*. \end{aligned} \tag{22}$$

Since $0 < s_i^* < u_{s_i}$, $h(s_i^*)$ cannot be the upper bound B_i; hence, $\overline{\lambda}_i$ is zero and (22) becomes:

$$\frac{\partial h_i(s_i^*)}{\partial s_i} = \left(1 - \sum_{k=1}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i + \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} \times Q_{ik}^*. \tag{23}$$

Equality (23) shows that the marginal expected cybersecurity cost is equal to the marginal expected cybersecurity investment revenue plus the term $\left(1 - \sum_{k=1}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i$; namely, the marginal expected cybersecurity investment revenue is equal to $\frac{\partial h_i(s_i^*)}{\partial s_i} - \left(1 - \sum_{k=1}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i$. This is reasonable because $\left(1 - \sum_{k=1}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i$ is the marginal expected damage expense.

If $\overline{\mu}_i^1 > 0$ and, hence, $s_i^* = 0$, and $\overline{\mu}_i^2 = 0$, we get:

$$-\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} = \frac{\partial h_i(0)}{\partial s_i} - \left(1 - \sum_{\substack{k=1\\k\neq i}}^{m} \frac{s_k^*}{m} + \frac{1-s_i^*}{m}\right) D_i - \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} Q_{ik}^* = \overline{\mu}_i^1. \tag{24}$$

In (24) minus the marginal expected cybersecurity investment utility is equal to $\overline{\mu}_i^1$; hence, the marginal expected cybersecurity cost is greater than the marginal expected cybersecurity investment revenue plus the marginal damage expense. Then the marginal expected cybersecurity investment revenue is less than the marginal expected cybersecurity cost minus the marginal damage expense. We note that case (24) can occur if $\dfrac{\partial h_i(0)}{\partial s_i}$ is strictly positive.

In contrast, if $\overline{\mu}_i^2 > 0$ and, hence, $s_i^* = u_{s_i}$, retailer j has a marginal gain given by $\overline{\mu}_i^2$, because

$$\begin{aligned}-\frac{\partial E(U_i(Q^*, u_{s_i}))}{\partial s_i} = &-\left(1 - \sum_{\substack{k=1\\k\neq i}}^{m} \frac{u_{s_k}}{m} + \frac{1-u_{s_i}}{m}\right) D_i \\ &- \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} \times Q_{ik}^* \\ &+ \frac{\partial h_i(u_{s_i})}{\partial s_i} + \overline{\lambda}_i \frac{\partial h_i(u_{s_i})}{\partial s_i} = -\overline{\mu}_i^2. \end{aligned} \tag{25}$$

We note that $\overline{\lambda}_i$ could also be positive, since, with $s_i^* = u_{s_i}$, $h_i(s_i)$ could reach the upper bound B_i. In (25) minus the marginal expected cybersecurity investment utility is equal to $-\overline{\mu}_i^2$. Hence, the marginal expected cybersecurity cost is less than the marginal expected cybersecurity investment revenue plus the marginal damage expense. Then the marginal expected cybersecurity investment revenue is greater than the marginal expected cybersecurity cost minus the marginal damage expense.

From (25) we see the importance of the Lagrange variables $\overline{\mu}_i^1$, $\overline{\mu}_i^2$ which describe the effects of the marginal expected cybersecurity investment utilities.

Remarks on the Stability of the Marginal Expected Cybersecurity Investment Utilities

Let us consider the three cases related to the marginal expected cybersecurity investment utilities studied in section "Analysis of Marginal Expected Cybersecurity Investment Utilities". Each of these cases holds for certain values of the damage D_i. Let us consider the value D_i for which the first case (23) occurs. We see that in this case there is a unique value of D_i for which (23) holds and if we vary such a value, also the value s_i^* in (23) varies. Now let us consider the value D_i for which (24) holds and let us call D_i^* the value of D_i for which we have

$$
\begin{aligned}
&-\frac{\partial E(U_i(Q^*, s^*))}{\partial s_i} \\
&= \frac{\partial h_i(0)}{\partial s_i} - \left(1 - \sum_{\substack{k=1 \\ k \neq i}}^{m} \frac{s_k^*}{m} + \frac{1 - s_i^*}{m}\right) D_i^* - \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} Q_{ik}^* = 0.
\end{aligned}
$$

Then for $0 < D_i < D_i^*$ the solution (Q^*, s^*) to variational inequality (9) remains unchanged because (24) still holds for these new values of D_i and the marginal expected cybersecurity investment utility remains negative, but it is increasing with respect to D_i. Analogously, if we consider the value D_i for which (25) holds and call D_i^* the value such that

$$
\begin{aligned}
-\frac{\partial E(U_i(Q^*, u_{s_i}))}{\partial s_i} = &- \left(1 - \sum_{\substack{k=1 \\ k \neq i}}^{m} \frac{u_{s_k}}{m} + \frac{1 - u_{s_i}}{m}\right) D_i^* \\
&- \sum_{k=1}^{m} \frac{\partial \hat{\rho}_k(Q^*, s^*)}{\partial s_i} \times Q_{ik}^* \\
&+ \frac{\partial h_i(u_{s_i})}{\partial s_i} + \overline{\lambda}_i \frac{\partial h_i(u_{s_i})}{\partial s_i} = 0,
\end{aligned}
$$

we see that for $D_i > D_i^*$ the solution (Q^*, s^*) to (9) remains unchanged because (25) still holds and the marginal expected cybersecurity investment utility remains positive and is increasing with respect to D_i.

A Numerical Example

The first example consists of two retailers and two demand markets as depicted in Fig. 2.

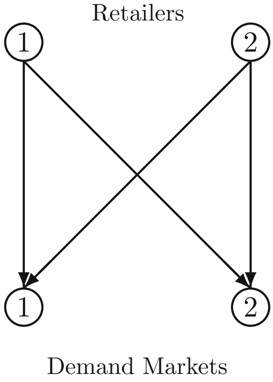


Fig. 2 Network topology for Example 1

It is inspired by related examples as in [15]. So, the cost function data are:

$$\begin{aligned} c_1 &= 5, & c_2 &= 10, \\ c_{11}(Q_{11}) &= 0.5Q_{11}^2 + Q_{11}, & c_{12}(Q_{12}) &= 0.25Q_{12}^2 + Q_{12}, \\ c_{21}(Q_{21}) &= 0.5Q_{21}^2 + Q_{21}, & c_{22}(Q_{22}) &= 0.25Q_{22}^2 + Q_{22}. \end{aligned}$$

The demand price functions are:

$$\rho_1(d,\overline{s}) = -d_1 + 0.1\frac{s_1+s_2}{2} + 100, \quad \rho_2(d,\overline{s}) = -0.5d_2 + 0.2\frac{s_1+s_2}{2} + 200.$$

The damage parameters are: $D_1 = 200$ and $D_2 = 210$ with the investment functions taking the form:

$$h_1(s_1) = \frac{1}{\sqrt{1-s_1}} - 1, \quad h_2(s_2) = \frac{1}{\sqrt{1-s_2}} - 1.$$

The damage parameters are in millions of $US, the expected profits (and revenues) and the costs are also in millions of $US. The prices are in thousands of dollars and the product transactions are in thousands. The budgets for the two retailers are identical with $B_1 = B_2 = 2.5$ (in millions of $US). In this case the bounds on the security levels are $u_{s_1} = u_{s_2} = 0.91$ and the capacities $\overline{Q_{ij}}$ are set to 100 for all i, j.

For $i = 1, 2$ we obtain:

$$\begin{aligned} -\frac{\partial E(U_i(Q,s))}{\partial Q_{i1}} &= 2Q_{i1} + Q_{11} + Q_{21} - 0.1\frac{s_1+s_2}{2} + c_i - 99, \\ -\frac{\partial E(U_i(Q,s))}{\partial Q_{i2}} &= Q_{i2} + 0.5Q_{12} + 0.5Q_{22} - 0.2\frac{s_1+s_2}{2} + c_i - 199, \end{aligned}$$

$$-\frac{\partial E(U_i(Q,s))}{\partial s_i} = -\frac{1}{20}Q_{i1} - \frac{1}{10}Q_{i2} - \left(1 - \frac{s_1+s_2}{2} + \frac{1-s_i}{2}\right)D_i + \frac{1}{2\sqrt{(1-s_i)^3}}.$$

Now, we want to find the equilibrium solution, taking into account the different values assumed by $\lambda^1,\ \lambda^2,\ \mu^1,\ \mu^2$ and λ, and searching, among them, the feasible ones. After some algebraic calculations, we realize that for $i = 1,2$ and $j = 1,2$ we get the solution when $\overline{\lambda}^1_{ij} = \overline{\lambda}^2_{ij} = \overline{\mu}^1_i = \overline{\lambda}_i = 0$, and $\overline{\mu}^2_i > 0$. Hence, $s^*_1 = s^*_2 = 0.91$ (which is the maximum value). In this case, the marginal expected transaction utilities are zero, whereas the marginal expected cybersecurity investment utilities are positive; namely, there is a marginal gain, given by $\overline{\mu}^2_i,\ i = 1,2$. Solving the system:

$$\begin{cases} \dfrac{\partial \mathcal{L}(Q^*,s^*,\overline{\lambda}^1,\overline{\lambda}^2,\overline{\mu}^1,\overline{\mu}^2,\overline{\lambda})}{\partial Q_{i1}} = 0 \\[2ex] \dfrac{\partial \mathcal{L}(Q^*,s^*,\overline{\lambda}^1,\overline{\lambda}^2,\overline{\mu}^1,\overline{\mu}^2,\overline{\lambda})}{\partial Q_{i2}} = 0 \quad i = 1,2, \\[2ex] \dfrac{\partial \mathcal{L}(Q^*,s^*,\overline{\lambda}^1,\overline{\lambda}^2,\overline{\mu}^1,\overline{\mu}^2,\overline{\lambda})}{\partial s_i} = 0 \end{cases}$$

namely:

$$\begin{cases} 3Q^*_{11} + Q^*_{21} - 0.1\dfrac{s^*_1+s^*_2}{2} + c_1 - 99 - \overline{\lambda}^1_{11} + \overline{\lambda}^2_{11} & = 0 \\[2ex] Q^*_{11} + 3Q^*_{21} - 0.1\dfrac{s^*_1+s^*_2}{2} + c_2 - 99 - \overline{\lambda}^1_{21} + \overline{\lambda}^2_{21} & = 0 \\[2ex] 1.5Q^*_{12} + 0.5Q^*_{22} - 0.2\dfrac{s^*_1+s^*_2}{2} + c_1 - 199 - \overline{\lambda}^1_{12} + \overline{\lambda}^2_{12} & = 0 \\ 0.5Q^*_{12} + 1.5Q^*_{22} - 0.2\dfrac{s^*_1+s^*_2}{2} + c_2 - 199 - \overline{\lambda}^1_{22} + \overline{\lambda}^2_{22} & = 0 \\[2ex] -\dfrac{1}{20}Q^*_{11} - \dfrac{1}{10}Q^*_{12} - \dfrac{3-2s^*_1-s^*_2}{2}D_1 + \dfrac{1+\overline{\lambda}_1}{2\sqrt{(1-s^*_1)^3}} - \overline{\mu}^1_1 + \overline{\mu}^2_1 & = 0 \\[2ex] -\dfrac{1}{20}Q^*_{21} - \dfrac{1}{10}Q^*_{22} - \dfrac{3-s^*_1-2s^*_2}{2}D_2 + \dfrac{1+\overline{\lambda}_2}{2\sqrt{(1-s^*_2)^3}} - \overline{\mu}^1_2 + \overline{\mu}^2_2 & = 0, \end{cases}$$

and therefore, assuming for $i = 1, 2$, $j = 1, 2$, $\overline{\lambda}^1_{ij} = \overline{\lambda}^2_{ij} = \overline{\mu}^1_i = \overline{\lambda}_i = 0$, and $\overline{\mu}^2_i > 0$, hence $s^*_1 = s^*_2 = 0.91$, and $D_1 = 200$ and $D_2 = 210$, we have:

$$\begin{cases} 3Q^*_{11} + Q^*_{21} = 94.091 \\ Q^*_{11} + 3Q^*_{21} = 89.091 \\ 1.5Q^*_{12} + 0.5Q^*_{22} = 195.82 \\ 0.5Q^*_{12} + 1.5Q^*_{22} = 190.82 \\ \overline{\mu}^2_1 = \frac{1}{20}Q^*_{11} + \frac{1}{10}Q^*_{12} + \frac{3 - 3 \times 0.91}{2} 200 - \frac{1}{2\sqrt{(1-0.91)^3}} \\ \overline{\mu}^2_2 = \frac{1}{20}Q^*_{21} + \frac{1}{10}Q^*_{12} + \frac{3 - 3 \times 0.91}{2} 210 - \frac{1}{2\sqrt{(1-0.91)^3}}. \end{cases}$$

The solution to the previous system is:

$$Q^*_{11} = 24.148, \quad Q^*_{21} = 21.586, \quad Q^*_{12} = 99.16, \quad Q^*_{22} = 94.16,$$
$$\overline{\mu}^2_1 = 19.6055, \quad \overline{\mu}^2_2 = 20.3273,$$

where $\overline{\mu}^2_1$ and μ^2_2 are the positive marginal expected gains.

For this example the stability results of Sect. hold. We are in the third case and if we double the value of the damage for the first retailer and assume now $D_1 = 400$, then the new value of the Lagrange multiplier is $\overline{\mu}^2_1 = 46.6055$.

Conclusions

Cyberattacks are negatively globally impacting numerous sectors of economies as well as governments and even citizens, and resulting in financial damages, disruptions, loss of services, etc. Hence, organizations, including companies from financial service firms to retailers, as well as utilities, are investing in cybersecurity. In this paper, we revisit a recently introduced cybersecurity investment supply chain game theory model described in [15] consisting of retailers and consumers at demand markets in which nonlinear budget constraints of the retailers associated with cybersecurity investments are explicitly included. The retailers compete in both product transactions and cybersecurity levels seeking to maximize their expected utilities, that is, expected profits, which capture both the expected revenues and the expected damages in the case of a cyberattack, which can differ from retailer to retailer. The consumers display their preferences through the demand price functions which are functions of the market demands for the product as well as the average security level of the network, which depends on all the retailers' investment levels. The governing equilibrium concept in this model of noncooperative behavior is that of Nash equilibrium.

In this paper, we provide a novel alternative formulation of the variational inequality formulation derived in [15]. The alternative formulation enables a deep analysis of the Lagrange multipliers associated with both the bounds on the product transactions between retailers and demand markets and the security levels of the retailers, with accompanying insights into the economic market forces. Specifically, we provide an analysis of both the marginal expected transaction utilities and the marginal expected cybersecurity investment utilities of the retailers. We also obtain stability results for the marginal expected cybersecurity investment utilities with respect to changes in the values of the retailers' financial damages.

The novel theoretical framework is then further illustrated through a numerical example for which the equilibrium product transaction and cybersecurity investment patterns are computed, along with the Lagrange multipliers. In addition, stability results are also given for the case where the first retailer's damage due to a cyberattack doubles.

The results in this paper add to the growing literature of operations research and game theory techniques for cybersecurity modeling and analysis.

Acknowledgements The research of the first author was partially supported by Istituto Nazionale di Alta Matematica Francesco Severi (Progetto di Ricerca GNAMPA 2015: *Nuove frontiere dei problemi di equlibrio su rete: dallo sviluppo sostenibile alla dinamica dei disastri ambientali ai crimini informatici*). The research of the third author was supported, in part, by the National Science Foundation under Grant No. 1551444. This support is gratefully acknowledged.

References

1. A. Barbagallo, P. Daniele, S. Giuffré, A. Maugeri, Variational approach for a general financial equilibrium problem: the deficit formula, the balance law and the liability formula. A path to the economy recovery. Eur. J. Oper. Res. **237**(1), 231–244 (2014)
2. P. Daniele, S. Giuffré, Random variational inequalities and the random traffic equilibrium problem. J. Optim. Theory Appl. **167**(1), 363–381 (2015)
3. P. Daniele, S. Giuffré, M Lorino, Functional inequalities, regularity and computation of the deficit and surplus variables in the financial equilibrium problem. J. Global Optim. **65**(3), 575–596 (2016)
4. P. Daniele, S. Giuffré, A. Maugeri, F. Raciti, Duality theory and applications to unilateral problems. J. Optim. Theory Appl. **162**, 718–734 (2014)
5. P. Daniele, S. Giuffré, M. Lorino, A. Maugeri, C. Mirabella, Functional inequalities and analysis of contagion in the financial networks, in *Handbook of Functional Equations*. Springer Optimization and its Applications, vol. 95 (Springer, New York, 2014), pp. 129–146
6. N.J. Daras, M.T. Rassias, *Computation, Cryptography, and Network Security* (Springer International Publishing, Switzerland, 2015)

7. S. Giuffré, A. Maugeri, A measure-type lagrange multiplier for the elastic-plastic torsion. Nonlinear Anal. **102**, 23–29 (2014)
8. S. Giuffré, A. Maugeri, D. Puglisi, Lagrange multipliers in elastic-plastic torsion problem for nonlinear monotone operators. J. Differ. Equ. **259**(3), 817–837 (2015)
9. L.A. Gordon, M.P. Loeb, M.P.W. Lucyshyn, L. Zhou, 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. J. Jahn, *Introduction to the Theory of Nonlinear Optimization* (Springer, Berlin, 1994)
11. M.H. Manshaei, T. Alpcan, T. Basar, J.P. Hubaux, Game theory meets network security and privacy. ACM Comput. Surv. **45**(3), Article No. 25 (2013)
12. S. Morgan, Cybersecurity Market Reaches $75 Billon in 2015; Expected to Reach $170 Billion by 2020, Forbes, 20 December (2015)
13. A. Nagurney, A multiproduct network economic model of cybercrime in financial services. Serv. Sci. **7**(1), 70–81 (2015)
14. A. Nagurney, L.S. Nagurney, A game theory model of cybersecurity investments with information asymmetry. Netnomics **16**(1–2), 127–148 (2015)
15. A. Nagurney, P. Daniele, S. Shukla, A supply chain network game theory model of cybersecurity investments with nonlinear budget constraints, Ann. Oper. Res. **248**(1), 405–427 (2017)
16. A. Nagurney, L.S. Nagurney, S. Shukla, A supply chain game theory framework for cybersecurity investments under network vulnerability, in *Computation, Cryptography, and Network Security*, ed. by N.J. Daras, M.T. Rassias (Springer International Publishing, Switzerland, 2015), pp. 381–398
17. J.F. Nash, Equilibrium points in n-person games. Proc. Natl. Acad. Sci. U. S. A. **36**, 48–49 (1950)
18. J.F. Nash, Noncooperative games. Ann. Math. **54**, 286–298 (1951)
19. R. Ostvold, B. Walker, *Business Resilience in the Face of Cyber Risk* https://www.accenture.com/t20150726T222401_w_/us-en/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Strategy_7/Accenture-Business-Resilience-in-the-face-of-cyber-risk.pdf
20. N. Shetty, G. Schwartz, M. Felegehazy, J. Walrand, Competitive cyber-insurance and internet security, in *Proceedings of the Eighth Workshop on the Economics of Information Security (WEIS 2009)*, University College London, England, 24–25 June (2009)
21. F. Toyasaki, P. Daniele, T. Wakolbinger, A variational inequality formulation of equilibrium models for end-of-life products with nonlinear constraints. Eur. J. Oper. Res. **236**(1), 340–350 (2014)
22. W. Yakowicz, in *Companies Lose $400 Billion to Hackers Each Year*. Inc., 8 September (2015)

Ellipsoid Targeting with Overlap

Nicholas J. Daras

Abstract First, we investigate the possibility of destruction of a passive point target. Subsequently, we study the problem of determination of best targeting points in an area within which stationary or mobile targets are distributed uniformly or normally. Partial results are given in the case in which the number of targeting points is less than seven or four, respectively. Thereafter, we study the case where there is no information on the enemy distribution. Then, the targeting should be organized in such a way that the surface defined by the kill radii of the missiles fully covers each point within a desired region of space-time. The problem is equivalent to the problem of packing ellipsoids of different sizes and shapes into an ellipsoidal container in $\mathbb{R}^4$ so as to minimize a measure of overlap between ellipsoids is considered.

Keywords Function of damage • Distribution of targets • Ellipsoid targeting with overlap • Sphere packing • Ellipsoid packing • Overlap measure

AMS subject classifications 2010: 65K10, 90C22, 65C50

The Probability of Destruction of a Point Target

Let $\mathbf{P}$ be the probability that a single shot destroys the point target. Then the probability that the target is destroyed by at least one shot of $\mathbf{n}$ independently aimed shots is given by $\mathbf{P}\,(\mathbf{n}\ \mathbf{s}\mathit{hots}) = \mathbf{1}-(\mathbf{1}-\mathbf{P})^{\mathbf{n}}$. Thus, to determine $\mathbf{P}\,(\mathbf{n}\ \mathbf{s}\mathit{hots})$ it suffices to know $\mathbf{P}$.

Assuming that the point target is located at the center $(\mathbf{0},\mathbf{0})$ of the plan, and denoting by $d\,(\mathbf{x},\mathbf{y})$ the probability density function of destruction of the target, if

N.J. Daras (✉)
Department of Mathematics, Hellenic Military Academy, 166 73, Vari Attikis, Greece
e-mail: ndaras@sse.gr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_7

the explosion point of the offensive weapon is the point $(\mathbf{x},\mathbf{y})$ and by $p(\mathbf{x},\mathbf{y})$ the probability density function in which the explosion point of the offensive weapon will be the point $(\mathbf{x},\mathbf{y})$, it is immediately verified that

$$\mathbf{P}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d(\mathbf{x},\mathbf{y})\,p(\mathbf{x},\mathbf{y})\;\mathbf{d}x\,\mathbf{d}y,$$

Generally, the **function of damage** $d(\mathbf{x},\mathbf{y})$ is *circularly symmetric*, i.e. the probability of destruction is a function of the variable

$$\mathbf{r}{=}\mathbf{r}(\mathbf{x},\mathbf{y})=\left(\mathbf{x}^2{+}\mathbf{y}^2\right)^{1/2}$$

If, in particular,

$$d(\mathbf{x},\mathbf{y})=\begin{cases}\mathbf{1},\ \mathbf{w}henever\ \mathbf{r}\le\mathbf{R}\\ \mathbf{0},\ \mathbf{w}henever\ \mathbf{r}>\mathbf{R}\end{cases}$$

($d(\mathbf{x},\mathbf{y})$ is the **cookie-cutter damage function**), then

$$\mathbf{P}=\iint_{\sqrt{\mathbf{x}^2+\mathbf{y}^2}\le\mathbf{R}} p(\mathbf{x},\mathbf{y})\;\mathbf{d}x\,\mathbf{d}y.$$

In what follows

1. We will be interested in the problem of determining the probability **P** depending on the various choices of the probability density of the explosion
$$p(\mathbf{x},\mathbf{y})\,.$$
2. Generally, it is assumed that the explosion density is described as a normal bivariate probability density function:
$$p(\mathbf{x},\mathbf{y})=\frac{\mathbf{1}}{\mathbf{2}\pi\;\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}}\;\mathbf{e}xp\left(-\frac{(\mathbf{x}-\mathbf{x_0})^2}{\mathbf{2}\,\sigma_{\mathbf{x}}^2}-\frac{(\mathbf{y}-\mathbf{y_0})^2}{\mathbf{2}\,\sigma_{\mathbf{y}}^2}\right).$$

We must distinguish 4 cases ([6, p. 17–27] and [15, p.34–42]):

1st Case The coordinate variations are equal and the expected point of explosion is the zero point:

$$\mathbf{x_0}{=}\mathbf{y_0}{=}\mathbf{0};\,\sigma_{\mathbf{x}}^2{=}\sigma_{\mathbf{y}}^2{=}\sigma^2.$$

Then it is easily verified that

$$\mathbf{P}{=}\mathbf{P}\left(\frac{\mathbf{R}}{\sigma}\right)=\int_{\mathbf{0}}^{\mathbf{R}}\frac{\mathbf{r}}{\sigma^2}\;\mathbf{e}xp\left(-\frac{\mathbf{r}^2}{\mathbf{2}\sigma^2}\right)\;\mathbf{d}r={\mathbf{1}}-\left(-\frac{\mathbf{R}^2}{\mathbf{2}\sigma^2}\right).$$

2nd Case The coordinate variations are equal and the coordinates of the expected point of the explosion are not equal to zero:

$$\mathbf{x_0 \neq 0, y_0 \neq 0; \sigma_x^2=\sigma_y^2=\sigma^2}.$$

Then

$$\mathbf{P=P\left(\frac{R}{\sigma},\frac{r_0}{\sigma}\right)=\frac{1}{\sigma^2}e}xp\left(-\frac{\mathbf{r_0^2}}{\mathbf{2\sigma^2}}\right)\int_0^{\mathbf{R}}\mathbf{r\ e}xp\left(-\frac{\mathbf{r^2}}{\mathbf{2\sigma^2}}\right)\mathbf{I_0}\left(\frac{\mathbf{r_0}}{\mathbf{\sigma^2}}\mathbf{r}\right)\ \mathbf{d}r$$

where $\mathbf{I_0\ (z)} = \sum_{\nu=0}^{\infty}\frac{1}{(2^\nu \nu!)^2}\mathbf{z}^\nu$ and $\mathbf{r_0}=\left(\mathbf{x_0^2+y_0^2}\right)^{1/2}$.

Alternatively, the following approximate formula is available [1, page 940]

$$\mathbf{P=P\left(\frac{R}{\sigma},\frac{r_0}{\sigma}\right)} \cong \begin{cases} \frac{\mathbf{2R^2}}{\mathbf{4+R^2}}\mathbf{e}xp\left(-\frac{\mathbf{r_0^2}}{\mathbf{4+R^2}}\right), \mathbf{i}f\ \mathbf{R\leq 1} \\ \frac{1}{2}\left(\mathbf{1+e}rf\left(\frac{\mathbf{x_1}}{\sqrt{2}}\right)\right), \mathbf{i}f\ \mathbf{1<R\leq 5} \\ \frac{1}{2}\left(\mathbf{1+e}rf\left(\frac{\mathbf{x_2}}{\sqrt{2}}\right)\right), \mathbf{i}f\ \mathbf{5<R} \end{cases}$$

where

$$\mathbf{x_1}=\frac{\sqrt[3]{\mathbf{R^2/(2+r_0)}}\mathbf{-1+\ (2/9)\ \left(2+2r_0^2\right)/\left(2+r_0^2\right)^2}}{\sqrt{\mathbf{(2/9)\ \left(2+2r_0^2\right)/\left(2+r_0^2\right)^2}}},$$

$$\mathbf{x_2=R-\left|r_0^2-1\right|^{1/2}},$$

$$\mathbf{e}rf\ \mathbf{(z)}=\frac{\mathbf{2}}{\sqrt{\pi}}\int_0^{\mathbf{z}}\mathbf{e^{-t^2}\ d}t.$$

3rd Case The coordinate variations are not equal and the coordinates of the expected point of the explosion are equal to zero:

$$\mathbf{x_0=y_0=0; \sigma_x^2 \neq \sigma_y^2}.$$

Then

$$\mathbf{P=P\left(\frac{R}{\sigma_{m\mathit{ax}}},c\right)=\frac{1}{c}}\int_0^{\mathbf{R}/\sigma^2_{\mathbf{m}ax}}\mathbf{r\ e}xp\left(\mathbf{-r^2\frac{1+c^2}{4c^2}}\right)\mathbf{I_0}\left(\mathbf{r^2\frac{1-c^2}{4c^2}}\right)\ \mathbf{d}r,$$

where $\mathbf{I_0\ (z)} = \sum_{\nu=0}^{\infty}\frac{1}{(2^\nu \nu!)^2}\mathbf{z}^\nu$, $\sigma_{\mathbf{m}ax}$: $=\mathbf{m}ax\left\{\sigma_{\mathbf{x}},\sigma_{\mathbf{y}}\right\}$ and $\mathbf{c}$: $=\frac{\mathbf{m}in\left\{\sigma_x^2,\sigma_y^2\right\}}{\mathbf{m}ax\left\{\sigma_x^2,\sigma_y^2\right\}}$.

Alternatively, the following approximate formula is available [10].

$$\mathbf{P=P\left(\frac{R}{\sigma_{m\mathit{ax}}},c\right)} \cong \frac{\mathbf{3}}{\sqrt{\mathbf{2}}}\ \frac{\sqrt[3]{\mathbf{R^2\ \left(\sigma_x^2+\sigma_y^2\right)}}\mathbf{-1+2\left(\sigma_x^2+\sigma_y^4\right)/9\left(\sigma_x^2+\sigma_y^2\right)^2}}{\sqrt{\mathbf{\sigma_x^2+\sigma_y^4}}\mathbf{/\left(\sigma_x^2+\sigma_y^2\right)}}.$$

4th Case The coordinate variations are not equal and the coordinates of the expected point of the detonation are not equal to zero:

$$\mathbf{x_0 \neq 0, y_0 \neq 0; \sigma_x^2 \neq \sigma_y^2.}$$

Then

$$\mathbf{P{=}P\left(R,c\right){=}\frac{1}{2\pi c}\,e}xp\left(-\frac{\mathbf{1}}{\mathbf{2}}\left\{\frac{\mathbf{x_0^2}}{\sigma_{\mathbf{x}}^{\mathbf{2}}}+\frac{\mathbf{y_0^2}}{\sigma_{\mathbf{y}}^{\mathbf{2}}}\right\}\right)\sum_{\mathbf{m=0}}^{\infty}\mathbf{B_mP_m}\left(\mathbf{R^2\frac{1+c^2}{4c^2}}\right).$$

where $\mathbf{P_m}\left(\boldsymbol{\lambda}\right)=\sum_{\nu=\mathbf{m+1}}^{\infty}\frac{\mathbf{e}^{-\lambda}\lambda^{\nu}}{\nu!}$ and $\mathbf{c}\!:=\frac{\mathbf{m}in\left\{\sigma_{\mathbf{x}}^{\mathbf{2}},\sigma_{\mathbf{y}}^{\mathbf{2}}\right\}}{\mathbf{m}ax\left\{\sigma_{\mathbf{x}}^{\mathbf{2}},\sigma_{\mathbf{y}}^{\mathbf{2}}\right\}}$. Here we have used the notation

$$\mathbf{B_m}\!:=\frac{\mathbf{1}}{\mathbf{2}}\mathbf{m!}\left(\frac{\mathbf{4c^2}}{\mathbf{1+c^2}}\right)^{\mathbf{m+1}}\sum_{\mathbf{i=0}}^{\mathbf{m}}\mathbf{D_{m,i}}$$

with

$$\mathbf{D_{m,i}}\!:=\frac{\mathbf{1}}{\mathbf{i!}}\left(\frac{\mathbf{1-c^2}}{\mathbf{4c^2}}\right)^{\mathbf{i}}\sum_{\mathbf{j=0}}^{\mathbf{m-i}}\frac{\left(\frac{\mathbf{x_0^2}}{\sigma_{\mathbf{x}}^{\mathbf{2}}}\right)^{\mathbf{j}}\left(\frac{\mathbf{y_0^2}}{\mathbf{c^2}\sigma_{\mathbf{x}}^{\mathbf{2}}}\right)^{\mathbf{m-i-j}}}{\mathbf{(2j)\,!\ (2m-2i-2j)\,!}}.$$

Alternatively, a third approximation formula is available [10]:

$$\mathbf{P{=}P\left(R,c\right)} \cong \frac{\sqrt[3]{\mathbf{R^2}\left(\sigma_{\mathbf{x}}^{\mathbf{2}}+\sigma_{\mathbf{y}}^{\mathbf{2}}\right)\mathbf{t}}-\left(\mathbf{1}-\left\{\mathbf{v/9t^2}\right\}\right)}{\sqrt{\mathbf{v/9t^2}}}$$

where $\mathbf{t{=}1{+}}\frac{\mathbf{x_0^2+y_0^2}}{\sigma_{\mathbf{x}}^{\mathbf{2}}+\sigma_{\mathbf{y}}^{\mathbf{2}}}$ and $\mathbf{v}{=}2\frac{\sigma_{\mathbf{x}}^{\mathbf{4}}+\sigma_{\mathbf{y}}^{\mathbf{4}}+\mathbf{2}\sigma_{\mathbf{x}}^{\mathbf{2}}\mathbf{x_0^2}+\mathbf{2}\sigma_{\mathbf{y}}^{\mathbf{2}}\mathbf{y_0^2}}{\sigma_{\mathbf{x}}^{\mathbf{2}}+\sigma_{\mathbf{y}}^{\mathbf{2}}}$.

Targeting Into Area in Which the Targets are Circularly Uniformly Distributed

Having regard to the mathematical formulas of the previous section, let us now suppose that after many observations, it was found by using statistical methods that *the opponent targets (moving or stationary) are circularly uniformly distributed into a two-dimensional region***.** The 1st problem which we will discuss is to determine the optimal target points, within the opponent region, against which should aim **n** missiles.

The case **n=2** is very simple.

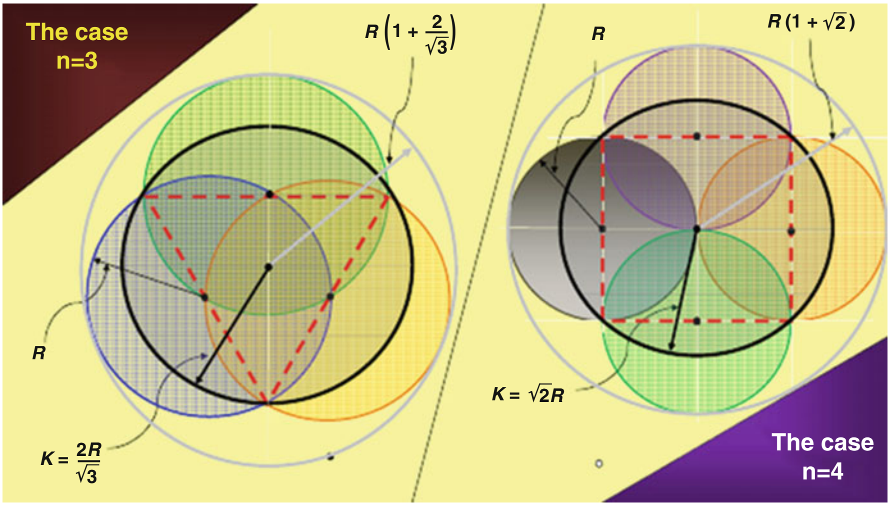


Fig. 1 Graphical representation of the cases $n = 3$ and $n = 4$

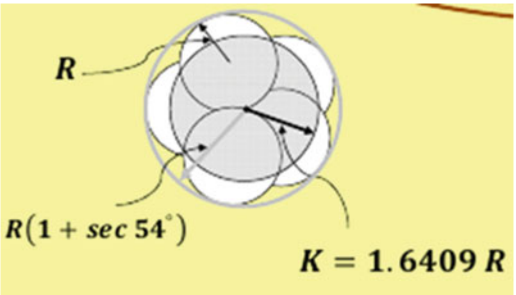


Fig. 2 Graphical representation of the radius of the larger circle of targets which is completely covered five weapons

Proposition 1 ([7]) *The largest circle within which all targets can be covered completely from two weapons has a radius equal to* **K=R**.

It follows that the use of two arms with a view to cover an entire disk shows no advantage compared to using only one weapon.

The cases **n=3** and **n=4** are a bit more complicated (Figs. 1 and 2).

Proposition 2 ([7]) *Three weapons should be targeted at middle points of the sides of an equilateral triangle entered in a circle of radius* $\mathbf{K=\frac{2R}{\sqrt{3}}}$.

Four arms should be targeted at middle points of the sides of a rectangle entered in a circle of radius $\mathbf{K=\sqrt{2}R}$.

Proposition 3 ([14]) *The radius of the larger circle of targets which is completely covered by five weapons is $K = 1.6409\ R$.*

The next result gives the smaller radii containing targets and covering weapons.

Proposition 4 ([7]) *The smaller disks containing targets and covering the kill radii of three, four, five, six or seven weapons have respective radii equal to*

$$\mathbf{R}\left(\mathbf{1}+\frac{\mathbf{2}}{\sqrt{\mathbf{3}}}\right),\mathbf{R}\left(\mathbf{1}+\sqrt{\mathbf{2}}\right),\mathbf{R}\left(\mathbf{1}+\mathbf{s}ec\ \mathbf{54}^{\circ}\right),\mathbf{3R}or\ \ \mathbf{3R}.$$

Remark 1 There are two crucial open questions. First, what happens when the weapons used are more than seven? And, secondly, if the dimension of the region, into which the observations are made, is equal to three (: observations in the three-dimensional space) or four (: observations in the space-time), what are the points (in the three-dimensional space or in the space-time) on which you need to target **n** weapons?

Targeting into Area in Which the Targets are Circularly Normally Distributed

By analogy with section, let us suppose that after many observations, it was found by using statistical methods that *the opponent targets (moving or stationary) are circularly normally distributed into a two-dimensional region*. The 2nd problem which we will now discuss is the following. Having regard to the mathematical formulas of the first section, determine optimal target points, within the opponent region, against which should aim **n** missiles.

Proposition 5 ([6, 7]) *Suppose the weapons explode randomly according to a uniform distribution and the target is distributed according to a circularly normal probability density function, with variance $\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{T}}$, into a disk having a radius equal to* **D** *(>***R***).*

1. *The rate of overall target value that is expected to be destroyed by* **n** *weapons is given by:*

$$\mathbf{E_n}\left[\mathbf{f}\right] = \left(\mathbf{1}-\mathbf{e}xp\left(-\frac{\mathbf{D^2}}{\mathbf{2}\sigma^{\mathbf{2}}_{\mathbf{T}}}\right)\right)\left(\mathbf{1}-\mathbf{e}xp\left(-\frac{\mathbf{n}\,\mathbf{R}^2}{\mathbf{D}^2}\right)\right).$$

2. *The value* $\mathbf{D_o}_{pt}$ *of the radius* **D** *that maximizes* $\mathbf{E_n}\left[\mathbf{f}\right]$ *is given by the formula:*

$$\left(\frac{\mathbf{D_o}_{pt}}{\sigma_{\mathbf{T}}}\right)^{\mathbf{2}}=\sqrt{\mathbf{2n}}\left(\frac{\mathbf{R}}{\sigma_{\mathbf{T}}}\right)(\Rightarrow \mathbf{E_n}\left[\mathbf{f}\right]=\left(\mathbf{1}-\mathbf{e}xp\left(-\sqrt{\frac{\mathbf{n}}{\mathbf{2}}}\frac{\mathbf{R}}{\sigma_{\mathbf{T}}}\right)\right)).$$

The case in which the weapons explode with accurate targeting is given below.

Proposition 6 *Suppose the weapons* ***explode with accurate targeting*** *and the target is distributed according to a circularly normal probability density function, with variance* $\boldsymbol{\sigma}_{\mathbf{T}}^{\mathbf{2}}$*, into a disk having a radius equal to* $\mathbf{D}$ *(*$>\mathbf{R}$*).*

i Marsaglia [13] If the kill radius of each of the two weapons is equal to $\mathbf{R}{=}\mathbf{R_0}\boldsymbol{\sigma}_{\mathbf{T}}$*, then the two weapons should explode on opposite sides from one another around the center of the target, at a distance* $\mathbf{d}{=}\mathbf{d_0}\boldsymbol{\sigma}_{\mathbf{T}}$ *from the center, wherein the constants* $\mathbf{R_0}$ *and* $\mathbf{d_0}$ *must satisfy the equation*

$$\int_{-\mathbf{d_0}/\mathbf{R_0}}^{\mathbf{1}} \frac{\mathbf{y\ e^{-d_0R_0y}}}{\sqrt{\mathbf{1-y^2}}}\ \mathbf{d}y{=}\mathbf{0}.$$

ii Gilliland [9]

1. *If* $(\mathbf{R}/\boldsymbol{\sigma}_{\mathbf{T}}) \le \mathbf{1}$*, then two weapons must explode in the two opposing sides symmetric relative to the center* $(\mathbf{0},\mathbf{0})$*, the point of explosion of each weapon at distance* $\mathbf{R}$ *from the center.*
2. *If* $(\mathbf{R}/\boldsymbol{\sigma}_{\mathbf{T}}) \le \mathbf{2}/\left(\mathbf{1}+\sqrt{\mathbf{3}}\right)$*, then three weapons should explode at the vertices of an equilateral triangle with sides equal to* $\mathbf{2R}$ *and center in* $(\mathbf{0},\mathbf{0})$*.*
3. *If* $(\mathbf{R}/\boldsymbol{\sigma}_{\mathbf{T}}) \le \mathbf{1}/\sqrt{\mathbf{3}}$*, then four weapons should explode on the vertices of a rhombus having sides equal to* $\mathbf{2R}$*, small diagonal equal to* $\mathbf{2R}$*, and center in* $(\mathbf{0},\mathbf{0})$*.*

Remark 2 There are two crucial open questions. First, *what happens when the weapons are not accurate when targeting*? And, secondly, *if the dimension of the region, into which the observations are made, is equal to three* (: observations in the three-dimensional space) *or four* (: observations in the space-time), *what are the points* (in the three-dimensional space or in the space-time) *on which you need to target* $\mathbf{n}$ *weapons*?

Ellipsoid Targeting with Overlap: Statement of the Problem

Question If, after a shot, there is no successful intercept enemy and if the enemy who remains invulnerable changes positions constantly, how should we react?

From a practical point of view, it seems unprofitable to attempt to make new observations to derive new stochastic conclusions on how is the enemy allocation, and then try new shots, aiming at selected points according to the above specifications.

Immediate Reaction Make simultaneous shots of many weapons that will ***fully*** cover the whole area and throughout a period of time. **It should be identified targeting points into the space of 4 or more generally of** $\mathbf{n}$ (>4) **dimensions,**

so that the balls with centers at the targeting points and with appropriate radii depending on the (same or different) kill radii of the used arms**, fully cover the entire enemy area, during a selected period**.

From a general mathematical point of view, it should be sought a minimum number of overlapping ellipsoids into the space of four or more dimensions, covering the entire enemy space during a given entire period. This is the **problem of packing ellipsoids** of different sizes and shapes into an ellipsoidal container so as to minimize a measure of overlap between ellipsoids considered [2, 8, 12].

We shall use some notation and preliminaries.

1. We denote by $\partial \mathbf{f}$ the ***Clarke subdifferential*** [5] of the function $\mathbf{f}:\mathscr{X}\to\mathbb{R}$, where $\mathscr{X}$ is a finite-dimensional vector space over the real numbers endowed with inner product $\langle\ ,\ \rangle$. (The usual Euclidean space $\mathbb{R}^{\mathbf{n}}$ with inner product $\langle\mathbf{x},\mathbf{y}\rangle := \mathbf{x}^{\mathbf{T}}\mathbf{y}$ and the space of symmetric matrices $\mathbf{S}\mathbb{R}^{\mathbf{n}\times\mathbf{n}}$ with inner product $\langle\mathbf{X},\mathbf{Y}\rangle := \mathit{trace}\,(\mathbf{XY})$ are two examples of particular interest in this presentation.)
2. In defining this quantity, we assume Lipschitz continuity of $\mathbf{f}$ at $\mathbf{x}$, and define the ***Clarke directional derivative*** as follows [3]:

$$\mathbf{f}^{\mathbf{0}}\,(\mathbf{x};\mathbf{h}) := \mathit{limsup}_{\,\mathbf{y}\longrightarrow\mathbf{x}\ \mathbf{a}\mathit{nd}\ \mathbf{t}\searrow\mathbf{0}}\ \frac{\mathbf{f}\,(\mathbf{y}+\mathbf{t}h)-\mathbf{f}\,(\mathbf{y})}{\mathbf{t}}$$

3. The ***Clarke subdifferential*** is then

$$\partial\mathbf{f}\,(\mathbf{x}) := \left\{\mathbf{v}\in\mathscr{X}:\langle\ \mathbf{v},\mathbf{h}\rangle \le \mathbf{f}^{\mathbf{0}}\,(\mathbf{x};\mathbf{h})\ \mathit{for\ all}\,\mathbf{h}\in\mathscr{X}\right\}. \tag{1}$$

If $\mathbf{f}$ is convex, the Clarke subdifferential coincides with the usual subdifferential from convex analysis:

$$\partial\mathbf{f}\,(\mathbf{x}) := \{\mathbf{v}\in\mathscr{X}:\mathbf{f}\,(\mathbf{y})\ge\mathbf{f}\,(\mathbf{x})+\langle\ \mathbf{v},\mathbf{y}-\mathbf{x}\rangle\ \mathit{for\ all}\,\mathbf{y}\in\mathit{dom}\,\mathbf{f}\}.$$

Definition 1 An ***ellipsoid*** $\mathscr{E}\subset\mathbb{R}^{\mathbf{n}}$ $(\mathbf{n}\ge\mathbf{2})$ can be specified in terms of its center $\mathbf{c}\in\mathbb{R}^{\mathbf{n}}$ and a symmetric positive definite eccentricity matrix $\mathbf{S}$:

$$\mathscr{E}=\left\{\mathbf{x}\in\mathbb{R}^{\mathbf{n}}:(\mathbf{x}-\mathbf{c})^{\mathbf{T}}\mathbf{S}^{-\mathbf{2}}\,(\mathbf{x}-\mathbf{c})\le\mathbf{1}\right\}=\{\mathbf{c}+\mathbf{S}u:\|\mathbf{u}\|_{\mathbf{2}}\le\mathbf{1}\} \tag{2}$$

It is easy that the ellipsoid $\mathscr{E}$ can equivalently be written as

$$\mathscr{E}=\left\{\mathbf{x}\in\mathbb{R}^{\mathbf{n}}:(\mathbf{x}-\mathbf{c})^{\mathbf{T}}\,\Sigma^{-1}\,(\mathbf{x}-\mathbf{c})\le\mathbf{1}\right\}. \tag{3}$$

where $\Sigma := \mathbf{S}^{\mathbf{2}}$ is also a symmetric positive definite matrix.

Remark 3 Note that the eigenvalues of $\mathbf{S}$ are the lengths $\mathbf{r_1}$, $\mathbf{r_2},\ldots,$ $\mathbf{r_n}$ of the principal semi-axes of $\mathscr{E}$. Further, the eigenvalues of Σ are $\mathbf{r}_{\mathbf{1}}^{\mathbf{2}},\ldots,$ $\mathbf{r}_{\mathbf{n}}^{\mathbf{2}}$ and the matrices $\mathbf{S}$ and Σ have the form

$$\mathbf{S} = \mathbf{Q}\begin{pmatrix} \mathbf{r_1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{r_n} \end{pmatrix}\mathbf{Q^T} \textit{ and } \Sigma = \mathbf{Q}\begin{pmatrix} \mathbf{r_1^2} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{r_n^2} \end{pmatrix}\mathbf{Q^T}$$

for some ***orthogonal matrix*** **Q**, which determines the orientation of the ellipsoid.

In view of the above, we can proceed to the rigorous mathematical restatement of the problem we have set.

Formulation of the Problem Given the semi-axis lengths $\mathbf{r_{i,1}}, \ldots, \mathbf{r_{i,n}}$ for a collection of **N** ellipsoids

$$\mathscr{E}_\mathbf{i}, \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N},$$

we want to specify centers $\mathbf{c_i}$ and matrices $\mathbf{S_i}$ for these ellipsoids, such that

(a) $\mathscr{E}_\mathbf{i} \subset \mathscr{E}$, for some fixed ellipsoidal container $\mathscr{E}$;
(b) The eigenvalues of $\mathbf{S_i}$ are $\mathbf{r_{i,1}}, \ldots, \mathbf{r_{i,n}}$, for $\mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N}$;
(c) Some measure of volumes of the pairwise overlaps

$$\mathscr{E}_\mathbf{i} \bigcap \mathscr{E}_\mathbf{j}, \mathbf{i}\ ,\mathbf{j} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N}, \mathbf{i}\ \neq\ \mathbf{j},$$

is minimized.

A Partial Case: The Sphere Packing Problem

We will first deal with the simplest case in which all enclosed shapes are spheres of arbitrary dimension [11, 16] and present a successive approximation algorithm that is shown to accumulate or converge to a stationary point of the formulation. To do so, we may proceed as in [17].

When the inscribed objects are spheres, the variables in the problem are the centers $\mathbf{c_i} \in \mathbb{R}^\mathbf{n}$, $\mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N}$, which we aggregate as a matrix:

$$\mathbf{c}\ :=\ (\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_N}). \tag{4}$$

Assuming that the corresponding radii $\mathbf{r_i}$, $\mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N}$ are given, we express the containment condition for each sphere as follows:

$$\mathscr{E}_\mathbf{i} \subset \mathscr{E}\ \Leftrightarrow \mathbf{c_i} \in \mathbf{K_i} \tag{5}$$

where $\mathbf{K_i}$ is a closed, bounded, convex set with nonempty interior. Obviously, if $\mathscr{E}$ is a sphere of radius **R** centered at **0**, then $\mathbf{K_i} = \{\mathbf{c_i}\colon \|\mathbf{c_i}\| \leq \mathbf{R} - \mathbf{r_i}\}$. It is reasonable to consider as a natural and simple measure for the *overlap* between two spheres $\mathscr{E}_\mathbf{i}$

and $\mathscr{E}_{\mathbf{j}}$ is the diameter of the largest sphere inscribed into the intersection, which we denote by an auxiliary variable:

$$\boldsymbol{\xi}_{\mathbf{i,j}} := \mathbf{m}ax\left\{\mathbf{0}, \left(\mathbf{r_i}+\mathbf{r_j}\right) - \left\|\mathbf{c_i}-\mathbf{c_j}\right\|_{\mathbf{2}}\right\}, \boldsymbol{\xi} := \left(\boldsymbol{\xi}_{\mathbf{i,j}}\right)_{\mathbf{1}\leq\mathbf{i}<j\leq N}. \tag{6}$$

With this notation, our minimum-overlap problem can thus be formulated as follows:

$$\mathbf{m}in_{\mathbf{c},\boldsymbol{\xi}}\ \mathbf{H}\left(\boldsymbol{\xi}\right) \tag{7a}$$

subject to

$$\left(\mathbf{r_i}+\mathbf{r_j}\right) - \left\|\mathbf{c_i}-\mathbf{c_j}\right\|_{\mathbf{2}} \leq \boldsymbol{\xi}_{\mathbf{i,j}}\ for\ \mathbf{1} \leq \mathbf{i} < j \leq N \tag{7b}$$

$$\mathbf{0}\ \leq \boldsymbol{\xi} \tag{7c}$$

$$\mathbf{c_i} \in \mathbf{K_i};\ for\ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N}, \tag{7d}$$

where (7c) denotes the entry wise condition $\boldsymbol{\xi}_{\mathbf{i,j}} \geq \mathbf{0}$, $\mathbf{1} \leq \mathbf{i} < \mathbf{j} \leq \mathbf{N}$, and the objective $\mathbf{H}{:}\mathbb{R}_{+}^{\mathbf{n(n-1)/2}} \to \mathbb{R}_{+}$ is chosen to be a convex and continuous function satisfying the following norm properties.

(a) $\mathbf{H(0)}\ = \mathbf{0}$,
(b) $\mathbf{H}\left(\boldsymbol{\xi}\right) > \mathbf{0}$ whenever $\boldsymbol{\xi} \neq \mathbf{0}$;
(c) $\mathbf{0} \leq \widetilde{\boldsymbol{\xi}} \leq \boldsymbol{\xi} \Rightarrow \mathbf{H}\left(\widetilde{\boldsymbol{\xi}}\right) \leq \mathbf{H}\left(\boldsymbol{\xi}\right)$.

Linearization of (7) around iterate $\widetilde{\mathbf{c}}$ restates the problem as follows:

$$\mathbf{P}(\widetilde{\mathbf{c}})\ := min_{\widetilde{\mathbf{c}},\boldsymbol{\xi}}\mathbf{H}\left(\widetilde{\boldsymbol{\xi}}\right) \tag{8a}$$

subject to

$$\left(\mathbf{r_i}+\mathbf{r_j}\right) - \mathbf{z}_{\mathbf{i,j}}^{\mathbf{T}}\left(\mathbf{c_i}-\mathbf{c_j}\right) \leq \widetilde{\boldsymbol{\xi}}_{\mathbf{i,j}}\, for\, \mathbf{1} \leq \mathbf{i}{<}\mathbf{j} \leq \mathbf{N} \tag{8b}$$

$$\mathbf{0}\ \leq \widetilde{\boldsymbol{\xi}} \tag{8c}$$

$$\mathbf{c_i} \in \mathbf{K_i};\ for\ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{8d}$$

where

$$\mathbf{z_{i,j}} := \begin{cases} \frac{\left(\widetilde{\mathbf{c}}_{\mathbf{i}}-\widetilde{\mathbf{c}}_{\mathbf{j}}\right)^{\mathbf{T}}}{\left\|\widetilde{\mathbf{c}}_{\mathbf{i}}-\widetilde{\mathbf{c}}_{\mathbf{j}}\right\|},\ when\ \widetilde{\mathbf{c}}_i \neq \widetilde{\mathbf{c}}_j \\ \mathbf{0},\ \ otherwise \end{cases}$$

Algorithm 1 Packing Spheres by Minimizing Overlap

```
Given $\mathbf{r_i}$, $\mathbf{i = 1, 2, \ldots, N}$ and $\mathbf{K_i}$ closed, convex, bounded with nonempty interior;
Choose $\mathbf{c^{(0)}} \in \mathbf{K_1 \times K_2 \times \cdots \times K_N}$;
for $\mathbf{k = 0, 1, 2, \ldots}$ do
    Solve $\mathbf{P}\left(\mathbf{c^{(k)}}\right)$ defined by (8) to obtain $(\mathbf{c^{(k+1)}};\ \boldsymbol{\xi}^{(\mathbf{k+1})})$;
    if $\mathbf{H}\left(\boldsymbol{\xi}^{(\mathbf{k+1})}\right) = \mathbf{H}\left(\boldsymbol{\xi}^{(\mathbf{k})}\right)$ then
        stop and return $\mathbf{c^{(k)}}$;
    end if
    Set $\boldsymbol{\xi}_{\mathbf{i,j}}^{(\mathbf{k+1})} = max\left\{\mathbf{0}, (\mathbf{r_i}+\mathbf{r_j}) - \left\|\mathbf{c_i^{(k+1)}} - \mathbf{c_j^{(k+1)}}\right\|_{\mathbf{2}}\right\}$ for $\mathbf{1 \le i < j \le N}$;
end for
```

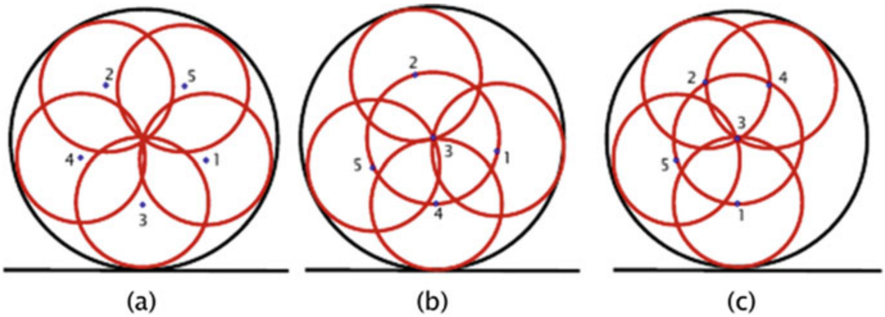


Fig. 3 Solutions obtained by Algorithm 1 for packing circles of radius $\mathbf{0.5}$ into a circle of radius $\mathbf{1}$, showing final overlap measures for each. (**a**) Global solution: $o = 0.4122147478$. (**b**) Local solution $o = 0.5$. (**c**) Local solution $o = 0.5$

This problem is convex, with affine constraints except for the inclusion (8d), which is satisfied strictly since each set $\mathbf{K_i}$ is closed, bounded and convex, with nonempty interior. To solve (8), we apply the next iterative algorithm, given by Uhler and Wright in [17]).

The convergence behavior of Algorithm 1 is described in the following.

Theorem 1 ([17]) *Suppose that the sets* $\mathbf{K_i}$ *in* (7) *are closed, bounded, and convex, with a nonempty interior, and that Assumption 1 holds. Then Algorithm 1 either terminates at a stationary point for* (7)*, or else generates an infinite sequence* $\left(\mathbf{c^{(k)}}\right)_{\mathbf{k=0,1,2,\ldots}}$ *for which all accumulation points* $\widehat{\mathbf{c}}$ *are either stationary points for* (7)*, or else have* $\widehat{\mathbf{c}}_\mathbf{i} = \widehat{\mathbf{c}}_\mathbf{j}$ *for some pair* $(\mathbf{i}, \mathbf{j})$ *with* $\mathbf{1 = i < j = N}$.

To be more attractive, let us give three examples cited in [17].

Example 1 (Five Circles) Consider the problem of packing five circles of radius $\mathbf{r_j = 0.5}$ into an enclosing circle of radius $\mathbf{R = 1}$. A few iterations of Algorithm 1 with $\mathbf{H}(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|_\infty$ and random starting points reveal a global solution in Fig. 3a and a family of local solutions in Fig. 3b and c.

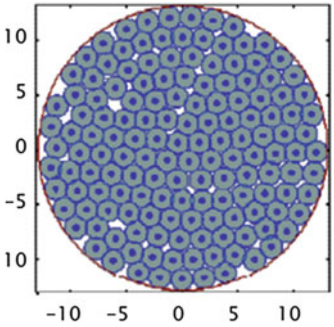


Fig. 4 Circle packing in a circular enclosure. A nearly hexagonal arrangement is seen in the interior

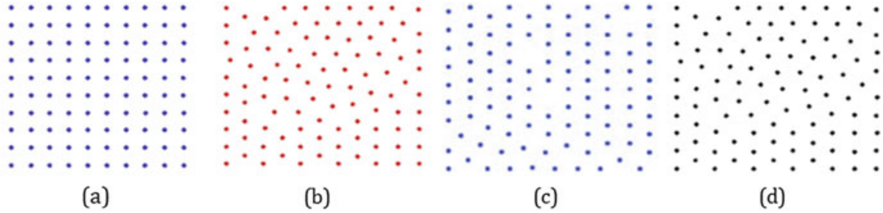


Fig. 5 Local minima obtained by Algorithm 1 for packing circles into a square, showing final overlap measures for each. (**a**) $o = 0.1192514295$. (**b**) $o = 0.1188906843$. (**c**) $o = 0.1181440939$. (**d**) $o = 0.1179656050$

Observe that, in the local solutions, one of the packed circles is positioned in the center of the enclosing circle and the remaining four circles are arranged around the boundary, in such a way that the maximum overlap between any pair of circles is $\mathbf{0.5}$.

Example 2 (Uniform Circles in $\mathbb{R}^2$) Application of Algorithm 1 with **N=150** circles, each of area $\boldsymbol{\pi}$, and a circular container of size $\mathbf{150\sqrt{12}}$, results in a total circle area-to-container area ratio which is equal to the optimal packing density (see Fig. 4). The hexagonal arrangement of the circles is clearly visible in the interior of the container.

Running tests in which **100** circles are packed into a square container and starting points are generated by arranging centers in a **10**×**10** square lattice are shown in Fig. 5, where, for clarity, only the centers appear, omitting the circles and a random perturbation to each center may be added. When no perturbations are added to the starting configuration, the algorithm does not move from the initial square configuration shown in Fig. 5a. When random initial enough large perturbations are applied, many different local minima are obtained (Fig. 5b–d).

Note that all of these have a maximum overlap less than the square configuration, and that hexagonal structure is recognizable in large parts of the domain, with square structure and disorder in intermediate regions.

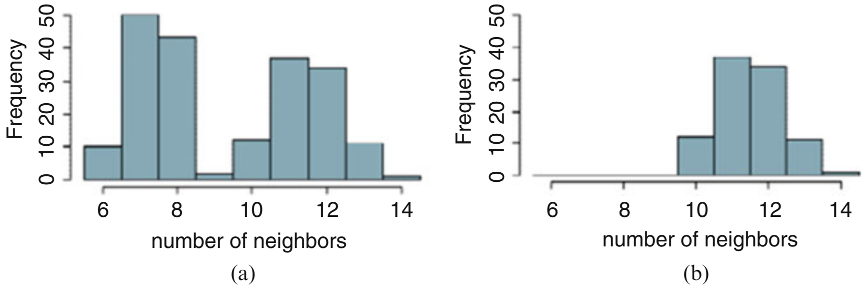


Fig. 6 Neighbor counts for packing of 100 three-dimensional spheres in a spherical container. (**a**) Distribution of neighbor counts. (**b**) Distribution of neighbor counts after removing spheres at the periphery

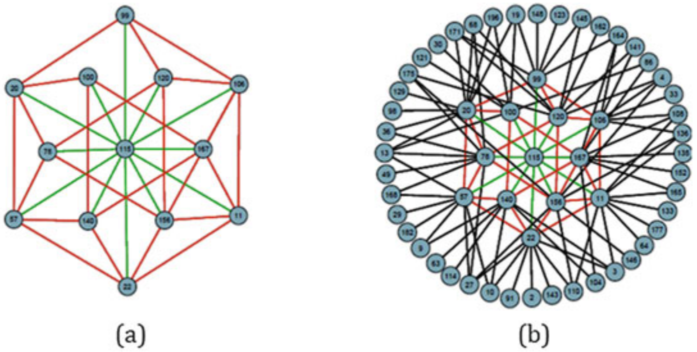


Fig. 7 Graph representing the sphere arrangement for one central sphere and neighboring spheres up to a distance of 2. Vertices correspond to spheres and edges represent overlap or "touching". (**a**) Neighbor graph for distance 1. (**b**) Neighbor graph for distance 2

Example 3 (Uniform Spheres in $\mathbb{R}^3$*)* Algorithm 1 converges to a solution in a finite minimum-overlap arrangement with $\mathbf{200}$ spheres enclosed in a larger sphere. Choosing the small spheres to have volume $\boldsymbol{\pi}$ and the containing sphere to have volume $\mathbf{200\sqrt{18}}$, the corresponding density of $\boldsymbol{\pi}/\sqrt{\mathbf{18}}$ is optimal in infinite space. At the solution obtained by Algorithm 1, we count the number of spheres that touch or intersect each sphere. This statistic provides an indication of the type of packing attained. The histogram for the number of neighboring spheres is shown in Fig. 6a. A more instructive diagram is obtained by removing from consideration those spheres that touch the enclosing sphere. After doing so, we obtain the histogram in Fig. 6b.

For further evidence, C. Uhler and S. J. Wright construct a graph where the vertices correspond to spheres and the edges represent overlap or "touching" with neighboring spheres [17]. They chose one centrally located sphere and graphed its contacts with neighbors (Fig. 7a) and neighbors-of-neighbors (Fig. 7b).

The General Case: The Ellipsoid Packing Problem

Let us now turn to a discussion of a bi-level optimization procedure for packing ellipsoids into an ellipsoidal container in a way that minimizes the maximum overlap of any pair of ellipsoids. It is not as obvious how to measure the overlap between two ellipsoids as between two spheres, since it depends on the orientation of the ellipsoids as well as the location of their centers.

In [17] C. Uhler and S. J. Wright proposed to measure the overlap by the sum of principal semi-axes of the largest ellipsoid that can be inscribed in the intersection of the two ellipsoids. To do so, they studied an alternative problem to the problem of finding the ellipsoid of largest volume [18] inscribed in an intersection of ellipsoids, first considered by Boyd and Vandenberghe in Sect. 8.4.2 of their book [4]. In fact, defining the ellipsoid $\mathscr{E}_{\mathbf{i}} = \{\mathbf{c_i} + \mathbf{S_i u} : \|\mathbf{u}\|_2 \leq \mathbf{1}\}$ and parameterizing the inscribed ellipsoid similarly by $\mathscr{E}_{\mathbf{i,j}} = \left\{\mathbf{c_{i,j}} + \mathbf{S_{i,j} u} : \|\mathbf{u}\|_2 \leq \mathbf{1}\right\}$, C. Uhler and S. J. Wright formulated the problem of measuring the maximal overlap as follows:

$$\widehat{\mathbf{O}}\left(\mathbf{c_i}, \mathbf{c_j}, \Sigma_{\mathbf{i}}, \Sigma_{\mathbf{j}}\right) := max_{\mathbf{S_{i,j}} \succcurlyeq \mathbf{0},\ \mathbf{c_{i,j}}, \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}}\ trace\left(\mathbf{S_{i,j}}\right)$$

subject to

$$\begin{pmatrix} -\boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}\mathbb{I} & \mathbf{O} & \mathbf{S_{i,j}} \\ \mathbf{O} & \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}} - \mathbf{1} & \left(\mathbf{c_{i,j}} - \mathbf{c_i}\right)^{\mathbf{T}} \\ \mathbf{S_{i,j}} & \mathbf{c_{i,j}} - \mathbf{c_i} & -\Sigma_{\mathbf{i}} \end{pmatrix} \preccurlyeq \mathbf{0} \tag{9a}$$

subject to

$$\begin{pmatrix} -\boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}\mathbb{I} & \mathbf{O} & \mathbf{S_{i,j}} \\ \mathbf{O} & \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}} - \mathbf{1} & \left(\mathbf{c_{i,j}} - \mathbf{c_j}\right)^{\mathbf{T}} \\ \mathbf{S_{i,j}} & \mathbf{c_{i,j}} - \mathbf{c_j} & -\Sigma_{\mathbf{j}} \end{pmatrix} \preccurlyeq \mathbf{0} \tag{9b}$$

where $\Sigma_{\mathbf{i}} = \mathbf{S}_{\mathbf{i}}^{\mathbf{2}}$ and $\Sigma_{\mathbf{j}} = \mathbf{S}_{\mathbf{j}}^{\mathbf{2}}$. The Lagrangian can be written as

$$\mathscr{L}\left(\mathbf{c_{i,j}}, \mathbf{S_{i,j}}, \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}, \mathbf{T_{i,j}}, \mathbf{M}^{(1)}_{\mathbf{i,j}}, \mathbf{M}^{(2)}_{\mathbf{i,j}}\right) :=$$

$$\left\langle \mathbb{I}, \mathbf{S_{i,j}} \right\rangle + \left\langle \mathbf{T_{i,j}}, \mathbf{S_{i,j}} \right\rangle - \left\langle \mathbf{M}^{(1)}_{\mathbf{i,j}}, \begin{pmatrix} -\boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}\mathbb{I} & \mathbf{O} & \mathbf{S_{i,j}} \\ \mathbf{O} & \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}} - \mathbf{1} & \left(\mathbf{c_{i,j}} - \mathbf{c_i}\right)^{\mathbf{T}} \\ \mathbf{S_{i,j}} & \mathbf{c_{i,j}} - \mathbf{c_i} & -\Sigma_{\mathbf{i}} \end{pmatrix} \right\rangle$$

$$- \left\langle \mathbf{M}^{(2)}_{\mathbf{i,j}}, \begin{pmatrix} -\boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}\mathbb{I} & \mathbf{O} & \mathbf{S_{i,j}} \\ \mathbf{O} & \boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}} - \mathbf{1} & \left(\mathbf{c_{i,j}} - \mathbf{c_j}\right)^{\mathbf{T}} \\ \mathbf{S_{i,j}} & \mathbf{c_{i,j}} - \mathbf{c_j} & -\Sigma_{\mathbf{j}} \end{pmatrix} \right\rangle,$$

with the dual problem being derived from

$$min_{\mathbf{M}^{(1)}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}, \mathbf{M}^{(2)}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}, \mathbf{T}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}} \left\{ max_{\mathbf{S}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}, \mathbf{c}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}} \mathscr{L}\left(\mathbf{c}_{\mathbf{i,j}}, \mathbf{S}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(1)}_{\mathbf{i,j}}, \boldsymbol{\lambda}^{(2)}_{\mathbf{i,j}}, \mathbf{T}_{\mathbf{i,j}}, \mathbf{M}^{(1)}_{\mathbf{i,j}}, \mathbf{M}^{(2)}_{\mathbf{i,j}}\right) \right\} .$$

Introducing the following notation for $\mathbf{M}^{(1)}_{\mathbf{i,j}}$ and $\mathbf{M}^{(2)}_{\mathbf{i,j}}$:

$$\mathbf{M}^{(1)}_{\mathbf{i,j}} = \begin{pmatrix} \mathbf{R}^{(1)}_{\mathbf{i,j}} & \mathbf{r}^{(1)}_{\mathbf{i,j}} & \mathbf{P}^{(1)}_{\mathbf{i,j}} \\ \left(\mathbf{r}^{(1)}_{\mathbf{i,j}}\right)^{\mathbf{T}} & \mathbf{p}^{(1)}_{\mathbf{i,j}} & \left(\mathbf{p}^{(1)}_{\mathbf{i,j}}\right)^{\mathbf{T}} \\ \mathbf{P}^{(1)}_{\mathbf{i,j}} & \mathbf{q}^{(1)}_{\mathbf{i,j}} & \mathbf{Q}^{(1)}_{\mathbf{i,j}} \end{pmatrix} \; and \; \mathbf{M}^{(2)}_{\mathbf{i,j}} = \begin{pmatrix} \mathbf{R}^{(2)}_{\mathbf{i,j}} & \mathbf{r}^{(2)}_{\mathbf{i,j}} & \mathbf{P}^{(2)}_{\mathbf{i,j}} \\ \left(\mathbf{r}^{(2)}_{\mathbf{i,j}}\right)^{\mathbf{T}} & \mathbf{p}^{(2)}_{\mathbf{i,j}} & \left(\mathbf{p}^{(2)}_{\mathbf{i,j}}\right)^{\mathbf{T}} \\ \mathbf{P}^{(2)}_{\mathbf{i,j}} & \mathbf{q}^{(2)}_{\mathbf{i,j}} & \mathbf{Q}^{(2)}_{\mathbf{i,j}} \end{pmatrix} \tag{10}$$

the dual is written explicitly as follows:

$$\begin{aligned} \widehat{\mathbf{O}}\left(\mathbf{c}_{\mathbf{i}}, \mathbf{c}_{\mathbf{j}}, \Sigma_{\mathbf{i}}, \Sigma_{\mathbf{j}}\right) := min_{\mathbf{M}^{(1)}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}, \mathbf{M}^{(2)}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}, \mathbf{T}_{\mathbf{i,j}} \succcurlyeq \mathbf{0}} \mathbf{p}^{(1)}_{\mathbf{i,j}} + \mathbf{p}^{(2)}_{\mathbf{i,j}} + \mathbf{2}\left(\mathbf{q}^{(1)}_{\mathbf{i,j}}\right)^{\mathbf{T}} \mathbf{c}_{\mathbf{i}} \\ + \mathbf{2}\left(\mathbf{q}^{(2)}_{\mathbf{i,j}}\right)^{\mathbf{T}} \mathbf{c}_{\mathbf{j}} + \left\langle \mathbf{Q}^{(1)}_{\mathbf{i,j}}, \Sigma_{\mathbf{i}} \right\rangle + \left\langle \mathbf{Q}^{(2)}_{\mathbf{i,j}}, \Sigma_{\mathbf{j}} \right\rangle \end{aligned} \tag{11a}$$

subject to

$$\mathbf{0} = \mathbb{I} + \mathbf{T}_{\mathbf{i,j}} - \mathbf{2P}^{(1)}_{\mathbf{i,j}} - \mathbf{2P}^{(2)}_{\mathbf{i,j}} \tag{11b}$$

$$\mathbf{0} = trace\left(\mathbf{R}^{(1)}_{\mathbf{i,j}}\right) - \mathbf{p}^{(1)}_{\mathbf{i,j}} \tag{11c}$$

$$\mathbf{0} = trace\left(\mathbf{R}^{(2)}_{\mathbf{i,j}}\right) - \mathbf{p}^{(2)}_{\mathbf{i,j}} \tag{11d}$$

$$\mathbf{0} = \mathbf{q}^{(1)}_{\mathbf{i,j}} + \mathbf{q}^{(2)}_{\mathbf{i,j}} \tag{11e}$$

We have assumed without loss of generality that $\mathbf{P}^{(1)}_{\mathbf{i,j}}$ and $\mathbf{P}^{(2)}_{\mathbf{i,j}}$ are in $\mathbf{S}\mathbb{R}^{\mathbf{n} \times \mathbf{n}}$; this follows from $\mathbf{S}_{\mathbf{i,j}} \in \mathbf{S}\mathbb{R}^{\mathbf{n} \times \mathbf{n}}$. A strong duality relationship holds between problems (9) and (11) because Slater's constraint qualification is satisfied for the second problem—it has a strictly feasible point. We construct this point by setting $\mathbf{T}_{\mathbf{i,j}} = \mathbb{I}$, and defining

$$\mathbf{M}^{(1)}_{\mathbf{i,j}} = \mathbf{M}^{(2)}_{\mathbf{i,j}} = \begin{pmatrix} \mathbb{I} & \mathbf{O} & \left(\frac{1}{2}\right)\mathbb{I} \\ \mathbf{0} & \mathbf{n} & \mathbf{0} \\ \left(\frac{1}{2}\right)\mathbb{I} & \mathbf{O} & \mathbb{I} \end{pmatrix}$$

It is easy to verify that these choices satisfy the linear constraints in (11), along with the (strict) interiority conditions $\mathbf{M}^{(1)}_{\mathbf{i,j}} \succ \mathbf{0}$, $\mathbf{M}^{(2)}_{\mathbf{i,j}} \succ \mathbf{0}$, $\mathbf{T}_{\mathbf{i,j}} \succ \mathbf{0}$.

We now turn to the problem of choosing ellipsoid positions and orientations. Using the notation defined in (9) and (11), C. Uhler and S. J. Wright formulated the min-max overlap problem as the following bi-level optimization problem:

$$min_{\boldsymbol{\xi},\left(\mathbf{c_{i,}}\ ,\mathbf{S_i},\varSigma_\mathbf{i}\right),\mathbf{i=1,2,\ldots,N}} \quad \boldsymbol{\xi} \tag{12a}$$

subject to

$$\boldsymbol{\xi} \geq \widehat{\mathbf{O}}\left(\mathbf{c_i}, \mathbf{c_j}, \varSigma_\mathbf{i}, \varSigma_\mathbf{j}\right), \quad \mathbf{1} \leq \mathbf{i} < j \leq N \tag{12b}$$

$$\mathscr{E}_\mathbf{i} \subset \mathscr{E}, \ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{12c}$$

$$\varSigma_\mathbf{i}{=}\mathbf{S_i^2}, \ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{12d}$$

$$semi-axes \ \ of \ \ \mathscr{E}_\mathbf{i} \ \ have \ \ lengths \ \mathbf{r_{i,1}}, \ldots, \mathbf{r_{i,n}}, \ \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{12e}$$

This problem is nonconvex.Finally, defining the containing ellipsoid to be

$$\mathscr{E}{=}\left\{\mathbf{x} \in \mathbb{R}^\mathbf{n}{:}(\mathbf{x}{-}\mathbf{c})^\mathbf{T} \varSigma^{-\mathbf{1}} \left(\mathbf{x}{-}\mathbf{c}\right) \leq \mathbf{1}\right\},$$

condition (12c) can be formulated as follows:

$$\begin{pmatrix} -\boldsymbol{\lambda}_\mathbf{i}\mathbb{I} & \mathbf{O} & \mathbf{S_i} \\ \mathbf{O} & \boldsymbol{\lambda}_\mathbf{i}{-}\mathbf{1} & (\mathbf{c_i}{-}\mathbf{c})^\mathbf{T} \\ \mathbf{S_i} & \mathbf{c_i}{-}\mathbf{c} & -\varSigma \end{pmatrix} \preccurlyeq \mathbf{0} \tag{13}$$

Given all these considerations, the relaxed version of (12) to be addressed in this section is expressed as follows:

$$min_{\boldsymbol{\xi},\left(\mathbf{c_{i,}}\ ,\mathbf{S_i},\varSigma_\mathbf{i}\right),\mathbf{i=1,2,\ldots,N}} \ \boldsymbol{\xi} \tag{14a}$$

subject to

$$\boldsymbol{\xi} \geq \widehat{\mathbf{O}}\left(\mathbf{c_i}, \mathbf{c_j}, \varSigma_\mathbf{i}, \varSigma_\mathbf{j}\right), \mathbf{1} \leq \mathbf{i} < j \leq N \tag{14b}$$

$$\begin{pmatrix} -\boldsymbol{\lambda}_\mathbf{i}\mathbb{I} & \mathbf{O} & \mathbf{S_i} \\ \mathbf{O} & \boldsymbol{\lambda}_\mathbf{i}{-}\mathbf{1} & (\mathbf{c_i}{-}\mathbf{c})^\mathbf{T} \\ \mathbf{S_i} & \mathbf{c_i}{-}\mathbf{c} & -\varSigma \end{pmatrix} \preccurlyeq \mathbf{0}{,}\mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{14c}$$

$$\begin{pmatrix} \varSigma_\mathbf{i} & \mathbf{S_i} \\ \mathbf{S_i} & \mathbb{I} \end{pmatrix} \succcurlyeq \mathbf{0}{,}\mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{14d}$$

$$trace\left(\mathbf{S_i}\right) = \mathbf{r_{i,1}} + \ldots + \mathbf{r_{i,n}}, \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{N} \tag{14e}$$

Note that when the ellipsoid $\mathscr{E}_\mathbf{i}$ is actually a circle, that is $\mathbf{r_{i,1}}=\ldots=\mathbf{r_{i,n}}$, we can fix $\mathbf{S_i}=\mathbf{r_{i,1}}\mathbb{I}$ and $\Sigma_\mathbf{i}=\mathbf{r_i^2}\mathbb{I}$ in (12), and eliminate these variables. Hence we can assume without loss of generality that $\mathbf{r_{i,1}}>\mathbf{r_{i,n}}$.

Properties of the Overlap Measure of an Ellipsoid Pair

To simplify the notation, we note that each dual overlap problem (10) has the general form

$$\mathbf{P}\left(\boldsymbol{\ell},\mathbf{C}\right):\mathbf{t}_{\boldsymbol{\ell}}^{*}\left(\mathbf{C}\right):=min_{\mathbf{M}_{\boldsymbol{\ell}}}\left\langle\mathbf{C},\mathbf{M}_{\boldsymbol{\ell}}\right\rangle \tag{15a}$$

subject to

$$\left\langle\mathbf{A}_{\boldsymbol{\ell},\mathbf{h}},\mathbf{M}_{\boldsymbol{\ell}}\right\rangle=\mathbf{b}_{\boldsymbol{\ell},\mathbf{h}},\mathbf{h}=\mathbf{1},\mathbf{2},\ldots,\mathbf{p}_{\boldsymbol{\ell}},\mathbf{M}_{\boldsymbol{\ell}}\succcurlyeq\mathbf{0} \tag{15b}$$

Here $\mathbf{C}$ captures the parameters that describe all the ellipsoids and $\mathbf{M}_{\boldsymbol{\ell}}$ is the dual variable for the overlap problem (11). We construct $\mathbf{C}$ as a block-diagonal matrix with $\mathbf{N}+\mathbf{1}$ diagonal blocks. First, there are $\mathbf{N}$ diagonal blocks of the form

$$\begin{pmatrix}\mathbf{O}&\mathbf{O}&\mathbf{O}\\ \mathbf{O}&\mathbf{1}&\mathbf{c}_\mathbf{i}^\mathbf{T}\\ \mathbf{O}&\mathbf{c}_\mathbf{i}&\Sigma_\mathbf{i}\end{pmatrix},\mathbf{i}=\mathbf{1},\mathbf{2},\ldots,\mathbf{N} \tag{16}$$

where each such block has dimension $\mathbf{7}\times\mathbf{7}$ and has the same partitioning scheme as the matrices $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{1})}$ and $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{2})}$ in (10). The remaining diagonal block in $\mathbf{C}$ is simply a $\mathbf{n}\times\mathbf{n}$ zero matrix that is used as the coefficient of the variable $\mathbf{T_{i,j}}$ in (11), for each pair $(\mathbf{i},\mathbf{j})$.

The variable $\mathbf{M}_{\boldsymbol{\ell}}$ in (15) has the same size and the same and block-diagonal structure as $\mathbf{C}$, where $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{1})}$ occupies the $\mathbf{i}$th block-diagonal location, $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{2})}$ occupies the $\mathbf{j}$th block-diagonal location, and $\mathbf{T_{i,j}}$ occupies the $(\mathbf{N}+\mathbf{1})$st block-diagonal location (which is the $\mathbf{n}\times\mathbf{n}$ submatrix that appears in the lower right corner of $\mathbf{M}_{\boldsymbol{\ell}}$). In the constraints (15b), the matrices $\mathbf{A}_{\boldsymbol{\ell},\mathbf{h}}$ are chosen to capture the constraints in (11) (all of which are equalities).

Structural constraints that enforce zeros in the locations of $\mathbf{M}_{\boldsymbol{\ell}}$ not occupied by the $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{1})}$, $\mathbf{M}_{\mathbf{i,j}}^{(\mathbf{2})}$, and $\mathbf{T_{i,j}}$ may be added to the formulation, but they are not necessary.

The primal form (8) of the overlap problem (15) has the form

$$max_{\zeta_{\boldsymbol{\ell}}=\left(\zeta_{\boldsymbol{\ell}}^{(1)},\ldots,\zeta_{\boldsymbol{\ell}}^{(\mathbf{p}_{\boldsymbol{\ell}})}\right)}\mathbf{b}_{\boldsymbol{\ell}}^{\mathbf{T}}\boldsymbol{\zeta}_{\boldsymbol{\ell}} \tag{17a}$$

subject to

$$\mathbf{C}-\sum_{\mathbf{h}=\mathbf{1}}^{\mathbf{p}_{\boldsymbol{\ell}}}\boldsymbol{\zeta}_{\boldsymbol{\ell}}^{(\mathbf{h})}\mathbf{A}_{\boldsymbol{\ell},\mathbf{h}}\succcurlyeq\mathbf{0} \tag{17b}$$

The program (11) and thus (15) always has a strictly feasible point. It follows that strong duality holds, that is, the optimal values of (15) and (17) are identical.

The Min-Max Overlap Problem and a Reference Problem

Using the notation of (15) and (17) to capture the relaxed min-maxoverlap problem (14), we can state this problem as follows:

$$\mathbf{P} : min_{\mathbf{C}\in\mathbf{K}}\, \mathbf{t}^*\,(\mathbf{C}) := max_{\ell=\mathbf{1,2,\ldots,m}}\mathbf{t}^*_{\ell}\,(\mathbf{C}) \tag{18}$$

Here each element $\ell \in \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{m}\}$ represents the overlap problem for a single pair of ellipsoids.

Note that $\mathbf{t}^*\,(\mathbf{C}) = -8$ if no pair of ellipsoids overlaps or touches. We now define a similar problem that depends on just a subset $\mathscr{F} \subset \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{m}\}$ of the overlaps. The objective of this "***reference problem***" is

$$\mathbf{t}^*_{\mathscr{F}}\,(\mathbf{C}) := max_{\ell\in\mathscr{F}}\mathbf{t}^*_{\ell}\,(\mathbf{C})\,, \tag{19}$$

where $\mathscr{F}$ is a subset of the strictly overlapping ellipsoid pairs, that is,

$$\mathscr{F} \subset \left\{\ell{=}\mathbf{1}, \mathbf{2}, \ldots, \mathbf{m}{:}\mathbf{t}^*_{\ell}\,(\mathbf{C}) > 0\right\}.$$

In the trust-region algorithm to be described below, the solutions $\mathbf{M}_{\ell}\,(\mathbf{C})$ of (15) for $\ell \in \mathscr{F}$ are used to construct a ***linearized subproblem*** whose solution is a step $\Delta\mathbf{C}$ in the parameter $\mathbf{C}$, assuming that the current $\mathbf{C}$ is feasible. The subproblem is as follows:

$$\ell\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho) : \quad \mathbf{r}\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho) := min_{\mathbf{r},\Delta\mathbf{C}}\,\mathbf{r} \tag{20a}$$

subject to

$$\mathbf{r} \geq \mathbf{t}^*_{\ell}\,(\mathbf{C}) + \langle \Delta\mathbf{C}, \mathbf{M}_{\ell}\,(\mathbf{C})\ \rangle\,, \ell \in \mathscr{F} \tag{20b}$$

$$\mathbf{C}{+}\Delta\mathbf{C} \in \mathbf{K}, \|\Delta\mathbf{C}\| \leq \rho \tag{20c}$$

Here $\rho > \mathbf{0}$ is a trust-region radius, and $\mathbf{M}_{\mathscr{F}}\,(\mathbf{C})$ denotes the set of matrices $\{\mathbf{M}_{\ell}\,(\mathbf{C}) : \ell \in \mathscr{F}\}$. The problem (20) is convex, and its feasible set is bounded, so it has an optimal value which we denote by $\Delta\mathbf{C}\,(\rho)$. Further, the KKT conditions are satisfied at this point.

For purposes of our main technical lemma, we define the "***predicted decrease***" from sub-problem $\ell\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho)$ as follows:

$$\Lambda\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho) : = \mathbf{t}^*_{\mathscr{F}}\,(\mathbf{C}) - \mathbf{r}\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho) \tag{21}$$

Note that since $\Delta\mathbf{C}{=}\mathbf{0}$ is feasible for (20), we have $\Lambda\,(\mathscr{F}, \mathbf{C}, \mathbf{M}_{\mathscr{F}}\,(\mathbf{C})\,, \rho) = \mathbf{0}$.

Trust Region Algorithm

We now define the algorithm for solving the problem **P** defined by (18).Note that in this general setting, $\mathbf{t}_{\ell}^{*}(\mathbf{C})$ defined by (15) is continuous on the set

$$\Psi_{\mathbf{j}} := \left\{\mathbf{C}{:}\mathbf{t}_{\ell}^{*}(\mathbf{C}) > -\infty\right\}$$

which is closed and convex. We make additional assumptions about the nature of the solutions to the parameterized primal–dual pair (15), (17), that do not hold in general, but which are satisfied for the application we consider here:

(i) $\mathbf{t}_{\ell}^{*}(\mathbf{C}) > -\infty \Rightarrow \mathbf{t}_{\ell}^{*}(\mathbf{C}) = \mathbf{0}$
(ii) *If* $\mathbf{t}_{\ell}^{*}(\mathbf{C}) > 0$, *the problem* (17) *has a strictly feasible point.*
We settle on the following requirement, which depends on parameters $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2 \in]\mathbf{0}, \mathbf{1}[$ with $\mathbf{0} < \boldsymbol{\eta}_1 < \boldsymbol{\eta}_2 < \mathbf{1}$:

Given $\mathbf{C}_{\mathbf{k}}$ for which $\mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}}) > 0$, we choose $\mathscr{F}_{\mathbf{k}}$ to satisfy:

$$\left\{\boldsymbol{\ell}{:}\mathbf{t}_{\ell}^{*}(\mathbf{C}_{\mathbf{k}}) \geq \boldsymbol{\eta}_2 \mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}}) \subset \mathscr{F}_{\mathbf{k}} \subset \left\{\boldsymbol{\ell}{:}\mathbf{t}_{\ell}^{*}(\mathbf{C}_{\mathbf{k}}) \geq \boldsymbol{\eta}_1 \mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}})\right\}\right\} \tag{22}$$

The Main Convergence Result for Algorithm 2 is given in the following.

Theorem 2 ([17]) *Suppose that Assumption 2 holds. Then either*

Algorithm 2 Packing Ellipsoids by Minimizing Overlap

```
Given $\mathbf{K} \subset \mathbb{SR}^{\mathbf{n}\times\mathbf{n}}$ compact; $\eta \in ]\mathbf{0}, \mathbf{1}[$; $\mathbf{c_1}$ and $\mathbf{c_2}$ with $\mathbf{0} < \mathbf{c_1} < \mathbf{c_2} < 1$;
$\Phi_1$ and $\Phi_2$ with $\mathbf{0} < \Phi_1 < 1 < \Phi_2$; and $\rho_{max} > \mathbf{0}$;
Choose $\mathbf{C_0} \in \mathbf{K}$, $\boldsymbol{\rho}_0 \in ]\mathbf{0}, \boldsymbol{\rho}_{max}]$;
for $\mathbf{k} = \mathbf{0}, \mathbf{1}, \mathbf{2}, \ldots$ do
    Define $\mathscr{F}_{\mathbf{k}}$ as in (22);
    Solve $\boldsymbol{\ell}\left(\mathscr{F}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}, \mathbf{M}_{\mathscr{F}_{\mathbf{k}}}(\mathbf{C}_{\mathbf{k}}), \boldsymbol{\rho}_{\mathbf{k}}\right)$ (20) to obtain $\Delta\mathbf{C}_{\mathbf{k}}$;
    Compute predicted decrease $\Lambda\left(\mathscr{F}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}, \mathbf{M}_{\mathscr{F}_{\mathbf{k}}}(\mathbf{C}_{\mathbf{k}}), \boldsymbol{\rho}_{\mathbf{k}}\right)$ from (21);
    if $\Lambda\left(\mathscr{F}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}, \mathbf{M}_{\mathscr{F}_{\mathbf{k}}}(\mathbf{C}_{\mathbf{k}}), \boldsymbol{\rho}_{\mathbf{k}}\right) = \mathbf{0}$ then
        stop and return $\mathbf{C}_{\mathbf{k}}$;
    end if
    if $\mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}} + \Delta\mathbf{C}_{\mathbf{k}}) \leq \mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}}) - \mathbf{c}_1\Lambda\left(\mathscr{F}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}, \mathbf{M}_{\mathscr{F}_{\mathbf{k}}}(\mathbf{C}_{\mathbf{k}}), \boldsymbol{\rho}_{\mathbf{k}}\right)$ then
        $\mathbf{C}_{\mathbf{k}+\mathbf{1}} \longleftarrow \mathbf{C}_{\mathbf{k}} + \Delta\mathbf{C}_{\mathbf{k}}$;
        if $\mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}} + \Delta\mathbf{C}_{\mathbf{k}}) \leq \mathbf{t}^{*}(\mathbf{C}_{\mathbf{k}}) - \mathbf{c}_2\Lambda\left(\mathscr{F}_{\mathbf{k}}, \mathbf{C}_{\mathbf{k}}, \mathbf{M}_{\mathscr{F}_{\mathbf{k}}}(\mathbf{C}_{\mathbf{k}}), \boldsymbol{\rho}_{\mathbf{k}}\right)$ then
            $\boldsymbol{\rho}_{\mathbf{k}+\mathbf{1}} \longleftarrow \mathbf{m}in\left\{\Phi_2\boldsymbol{\rho}_{\mathbf{k}}, \boldsymbol{\rho}_{max}\right\}$;
        end if
        if $\mathbf{t}^{*}\left(\mathbf{C}_{\mathbf{k}+\mathbf{1}}\right) \leq \mathbf{0}$ then
            stop and return $\mathbf{C}_{\mathbf{k}+\mathbf{1}}$;
        end if
    else
        $\mathbf{C}_{\mathbf{k}+\mathbf{1}} \longleftarrow \mathbf{C}_{\mathbf{k}}$
        $\boldsymbol{\rho}_{\mathbf{k}+\mathbf{1}} \longleftarrow \Phi_1\boldsymbol{\rho}_{\mathbf{k}}$;
    end if
end for
```

(a) Algorithm 2 terminates finitely at a point that is Clarke-stationary for problem $\mathbf{P}$ (18)*, or has a non-positive value of* $\mathbf{t}^*$*; or*

(b) it generates an infinite sequence of iterates $\{\mathbf{C_k}\}$ *for which accumulation points exist, and all accumulation points* $\overline{\mathbf{C}}$ *either are Clarke-stationary for* $\mathbf{P}$ *or have* $\mathbf{t}^*\left(\overline{\mathbf{C}}\right)=\mathbf{0}$.

References

1. M. Abramowitz, I.A. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Applied Mathematical Series, vol. 55 (National Bureau of Standards, Washington, DC, 1964)
2. A.V. Bondareno, D.P. Hardin, E.B. Saff, Minimal N-point diameters and f-best-packing constants in $\mathbb{R}^d$. Proc. AMS **142**(3), 981–988 (2014)
3. J. Borwein, A.S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics (Springer, New York, 2000)
4. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2003)
5. F.H. Clarke, *Optimization and Nonsmooth Analysis* (Wiley, New York, 1983)
6. R.L. Duncan, Hit probabilities for multiple weapons systems. SIAM Rev. **6**, 111–114 (1964)
7. A.R. Eckler, S.A. Burr, *Mathematical Models of Target Coverage and Missile Allocation*. MORS, Heritage Series (1972), (see also http://www.dtic.mil/dtic/tr/fulltext/u2/a953517.pdf)
8. J. Egeblad, Heuristics for multidimensional packing problems, Ph.D. thesis, Department of Computer Science, University of Copenhagen (2008), (see also http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.229.3169&rep=rep1&type=pdf)
9. D.C. Gilliland, Some bombing problems. Am. Math. Mon. **73**, 713–716 (1966)
10. F.E. Grubbs, Approximate circular and noncircular offset probabilities of hitting. Oper. Res. **12**, 51–62 (1964)
11. T.C. Hales, Cannonballs and honeycombs. Not. AMS **47**(4), 440–449 (2000)
12. Y. Li, H. Akeb, C.M. Li, Greedy algorithms for packing unequal circles. J. Oper. Res. Soc. **56**(5), 539–548 (2005)
13. G. Marsaglia, Some problems involving circular and spherical targets. Oper. Res. **13**, 18–27 (1965)
14. E.H. Neville, On the solution of numerical functional equations. Proc. Lond. Math. Soc. **14**(2), 308–326 (1915)
15. J.S. Przemieniecki, *Mathematical Methods in Defense Analyses*. Air Force Institute of Technology, Education Series, 3rd edn. (American Institute of Aeronautics and Astronautics, Reston, 2000). ISBN-13: 978–1563473975, ISBN-10: 1563473976
16. K. Stephenson, Circle packing: a mathematical tale. Not. AMS **50**(11), 1376–1388 (2003)
17. C. Uhler, S.J. Wright, Packing ellipsoids with overlap. SIAM Rev. **55**(4), 671–706 (2013)
18. A.J. Wilson, Volume-of-n-dimensional-ellipsoid. Sciencia Acta Xaveriana (SAX) **1**(1), 101–106 (2010), http://oaji.net/articles/2014/1420-1415594291.pdf

A Review of Several Optimization Problems Related to Security in Networked System

Bhaskar DasGupta and Venkatkumar Srinivasan

Abstract Security issues are becoming more and more important to activities of individuals, organizations, and the society in our modern networked computerized world. In this chapter we survey a few optimization frameworks for problems related to security of various networked system such as the internet or the power grid system.

Keywords Security • Networked system • Optimization framework

Introduction

Security issues are essential to activities of individuals, organizations, and the society as a whole in our modern networked computerized world in healthcare, power management, online purchase, banking, intra-business transactions, and many other similar activities in distributed-computing settings. A typical activity in such a networked system involves a set of (digital) transactions between various components ("agents") of the system to perform a specific task such as online purchase of an item or submitting an online application for a job, and requires interaction with various computer servers/databases and encryption services. Any compromise of these activities due to other malicious agents within the system or outside may lead to severe consequences such as disruption of critical infrastructure or national economy, and thus making sure that these activities are secure against such attacks is of paramount importance. The significance of maintaining security of networked systems has in fact led to organization of many security competitions, such as the *international Capture The Flag* [24] for researchers to discuss, discover, and validate new solutions for security issues.

In this chapter, we survey several optimization problems related to the security issues in distributed networked systems. We start by surveying some of the grand mathematical challenges in developing and analyzing security of real-world

B. DasGupta (✉) • V. Srinivasan
Department of Computer Science, University of Illinois at Chicago, Chicago, IL 50507, USA
e-mail: bdasgup@uic.edu; vsrini7@uic.edu

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_8

networked systems in section “Mathematical and Statistical Challenges in Networked Security”. Then, in the remaining sections we survey several optimization problems related to maintaining and evaluating securities of such systems.

Mathematical and Statistical Challenges in Networked Security

Some of the major mathematical and statistical challenges for security issues in networked systems are discussed in the two white papers [7, 15]. Based on these and other white papers, at least the following four possible challenge areas can be identified:

Data acquisition: The challenge here is to generate accurate trace and log data while maintaining their integrity throughout the lifetime of their intended use for scientific analysis and verification since lack of public data sets is a significant barrier in current research [19]. This challenge is also related to the so-called utility versus privacy trade-off issue [21] since making data publicly available may pose confidentiality and privacy issues.

Modelling networks: This challenge refers to the difficulties in developing mathematical network models that accurately model real-world networks and statistical methods for comparing networks (e.g., see [9]). For example, a typical question could be whether the distribution of degrees of nodes over the entire network is governed by power-laws or its variants?

Detection and response to security threats: This challenge refers to the difficulty in formulating and solving problems such as malicious code or behavior detection that provide long-term proactive approach to network security. Research methods for overcoming this challenge may involve techniques from diverse areas of mathematics or computer science such as dynamic data modelling methods, optimization methods, machine learning methods, and methods for uncertainty modelling via probabilistic models.

Modelling network dynamics: This challenge refers to developing appropriate mathematical models to understand the mechanism of spread of infections (i.e., time evolution of malicious attacks) in networks. Ideas from game theory or dynamical systems may be particularly useful in this context.

Application of Convex Optimization in Network Security

In this section, we review an application of convex or concave optimization methods by Vamvoudakis et al. [23] to model the complex behavior of a malicious attacker in a networked system. The model was incorporated in a cyber security advisory system to demonstrate its effectiveness. The optimization problem is formulated *from the point of view of a malicious attacker*, i.e., the goal is to find an optimal allocation of available resources for an attacker to maximize the potential damage.

We start by specifying a *model* of the damage caused by a (malicious) attacker of the given network. In the model, $t \in \{1, 2, \dots, T\}$ indicates the discrete time variable. Assume that there are a set of S services, indexed by $1, 2, \dots, S$, in our networked system that may be attacked for disruption. The following parameters are used in the model:

$u^s_{\mathrm{AR}_t} \geq 0$: A scalar quantifying the *amount of attack resources* (e.g., amount of money devoted to attack a particular resource) used by the attacker to attack service s at time t.

$x^s_{\mathrm{PD}_t} \geq 0$: A scalar denoting the *amount of potential damage* caused by attacks. In general $x^s_{\mathrm{PD}_t} = f^s_t \left(u^s_{\mathrm{AR}_t}\right)$ for some appropriate function $f^s_t : \mathbb{R}^+ \mapsto \mathbb{R}^+$.

g^s_t: $g^s_t \left(u^s_{\mathrm{AR}_t}\right)$ denotes the probability that the damage $f^s_t \left(u^s_{\mathrm{AR}_t}\right)$ is realized as a result of the attack $x^s_{\mathrm{PD}_t}$.

$y^s_{\mathrm{TD}_t}$: $y^s_{\mathrm{TD}_t} = g^s_t \left(u^s_{\mathrm{AR}_t}\right) f^s_t \left(u^s_{\mathrm{AR}_t}\right)$ is the *expected damage* caused by $u^s_{\mathrm{AR}_t}$.

y_{TD}: $y_{\mathrm{TD}} = \sum_{t=1}^{T} \sum_{s=1}^{S} y^s_{\mathrm{TD}_t}$ is the total expected damage.

U_{TR}: This is the total budget of attack resources available to the attacker.

In order to ensure that the resulting optimization problems are convex or concave, Vamvoudakis et al. [23] make the following assumptions that are justified for real applications of the model:

- f^s_t is a *linear* function, i.e.,

$$f^s_t \left(u^s_{\mathrm{AR}_t}\right) = a^s_t + b^s_t \, u^s_{\mathrm{AR}_t} \tag{1}$$

  for some constants $a^s_t, b^s_t \in \mathbb{R}^+$. The constant a^s_t models the extent of damage without any attack whereas the constant b^s_t models the extent of damage per unit of attack resources employed. The equation has the realistic implication that an increase in attack resources leads to an increase in the potential damage caused.
- g^s_t is a *linearly increasing* function projected to the interval [0, 1], i.e.,

$$g^s_t \left(u^s_{\mathrm{AR}_t}\right) = \begin{cases} 0, & \text{if } d^s_t \, u^s_{\mathrm{AR}_t} > c^s_t \\ c^s_t - d^s_t \, u^s_{\mathrm{AR}_t}, & \text{if } c^s_t - 1 \leq d^s_t \, u^s_{\mathrm{AR}_t} \leq c^s_t \\ 1, & \text{if } d^s_t \, u^s_{\mathrm{AR}_t} < c^s_t - 1 \end{cases} \tag{2}$$

  for some given constants $c^s_t, d^s_t \geq 0$. The constant c^s_t models the probability of damage realization without any attack whereas the constant d^s_t models the decrease of the probability of damage realization per unit of attack resources employed. Note that this choice of g^s_t models the realistic assumption that an increase in attack resources decreases the realization probability of the potential damage since a large-scale attack is much more likely to trigger defense mechanisms.

Now we can consider two different optimization problems for optimal allocation of available resources by an attacker to maximize the potential damage depending on the availability of relevant data.

Optimization Problem When All Relevant Damage Data Is Known When all the relevant damage data, i.e., all the numbers in $\{a_t^s, b_t^s, c_t^s, d_t^s \,|\, 1 \leq s \leq S, 1 \leq t \leq T\}$, are known *a-priori*, it is easy to see that the optimal attack resource allocation values (i.e., the $u_{\mathrm{AR}_t}^s$'s) that maximizes the total expected damage y_{TD} can be obtained by solving the following constrained optimization problem:

$$\begin{array}{lll} \textit{maximize} & y_{\mathrm{TD}} & \\ \textit{subject to} & \displaystyle\sum_{t=1}^{T}\sum_{s=1}^{S} u_{\mathrm{AR}_t}^s \leq U_{\mathrm{TR}} & \\ & u_{\mathrm{AR}_t}^s \geq 0, & 1 \leq s \leq S, 1 \leq t \leq T \end{array} \tag{3}$$

Although in general (3) may be difficult to solve, Vamvoudakis et al. [23] show that the special choices of f_t^s in (1) and g_t^s in (2) ensure that the above optimization problem is *equivalent* to solving the following *concave maximization* (or, equivalently convex minimization)[1] problem with linear constraints involving an addition set of z_t^s variables:

$$\begin{array}{lll} \textit{maximize} & \displaystyle\sum_{t=1}^{T}\sum_{s=1}^{S} \left(a_t^s + b_t^s\, u_{\mathrm{AR}_t}^s\right) \left(c_t^s - d_t^s\, u_{\mathrm{AR}_t}^s - z_t^s\right) & \\ \textit{subject to} & \displaystyle\sum_{t=1}^{T}\sum_{s=1}^{S} u_{\mathrm{AR}_t}^s \leq U_{\mathrm{TR}} & \\ & c_t^s - d_t^s\, u_{\mathrm{AR}_t}^s - z_t^s \leq 1, & 1 \leq s \leq S, 1 \leq t \leq T \\ & u_{\mathrm{AR}_t}^s \geq 0, & 1 \leq s \leq S, 1 \leq t \leq T \end{array} \tag{4}$$

and, moreover, if $0 \leq c_t^s \leq 1$ for all s and t, then one can set $z_t^s = 0$ for all s and t in the above concave optimization problem.

Optimization Problem When Not All Relevant Damage Data Is Known Often the parameter values in $\left\{a_t^s, b_t^s, c_t^s, d_t^s \,|\, 1 \leq s \leq S, 1 \leq t \leq T\right\}$ are *not* known *a-priori*. In that case, one needs to estimate these parameter values online based on past observations using some *machine learning* techniques such as the maximum-likelihood approach. Vamvoudakis et al. [23] propose the following approach for the parameter estimation problem:

[1] A function h of k variables is convex (resp. concave) if and only if, for all $x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k$ and for all $0 < \lambda < 1$, $h\big((1-\lambda)\,(x_1, x_2, \ldots, x_k) + \lambda\,(y_1, y_2, \ldots, y_k)\big) \geq (1-\lambda)\,h(x_1, x_2, \ldots, x_k) + \lambda\,h(y_1, y_2, \ldots, y_k)$ (resp. $h\big((1-\lambda)\,(x_1, x_2, \ldots, x_k) + \lambda\,(y_1, y_2, \ldots, y_k)\big) \leq (1-\lambda)\,h(x_1, x_2, \ldots, x_k) + \lambda\,h(y_1, y_2, \ldots, y_k)$). When the objective function and all the constraints are convex (resp. concave), we have a convex (resp. concave) optimization problem. The convexity or concavity property often makes an optimization problem easier to solve as opposed to the general case; see [3] for further details.

- Assume that these parameters are generated by a *linear* dynamical system over time t, i.e.,

$$a_t^s = C_a^s x_a^s(t) \ \text{ where } \ x_a^s(t) \ \text{ is generated by } \ x_a^s(t) = A_a^s x_a^s(t-1) + B_a^s w^s(t-1) \tag{5}$$

$$b_t^s = C_b^s x_b^s(t) \ \text{ where } \ x_b^s(t) \ \text{ is generated by } \ x_b^s(t) = A_b^s x_b^s(t-1) + B_b^s w^s(t-1) \tag{6}$$

$$c_t^s = C_c^s x_c^s(t) \ \text{ where } \ x_c^s(t) \ \text{ is generated by } \ x_c^s(t) = A_c^s x_c^s(t-1) + B_c^s w^s(t-1) \tag{7}$$

$$d_t^s = C_d^s x_d^s(t) \ \text{ where } \ x_d^s(t) \ \text{ is generated by } \ x_d^s(t) = A_d^s x_d^s(t-1) + B_d^s w^s(t-1) \tag{8}$$

  where $\left\{ A_j^s, B_j^s, C_j^s, D_j^s \mid 1 \leq s \leq S, j \in \{a, b, c, d\} \right\}$ are scalar parameters of the dynamics, and w_t^s are sequences generated by a random process with zero mean and variance z_t^s. Use historical data to estimate these dynamics using blackbox identification techniques.
- Now use online data to estimate the values of $\{ a_t^s, b_t^s, c_t^s, d_t^s \mid 1 \leq s \leq S, 1 \leq t \leq T\}$ based on past observations using a k-step ahead predictor in the following manner. Let $\left\{ a_t^s, b_t^s, c_t^s, d_t^s \mid 1 \leq s \leq S, 1 \leq t \leq k < T\right\}$ be the set of values observed (by the attacker) for these parameters up to sometime $k < T$ and the attacker needs to compute the "future" values of $u_{\mathrm{AR}_t}^s$'s for $k < t \leq T$. Then, one can do the following:
  - Estimate the values of $\left\{ a_t^s, b_t^s, c_t^s, d_t^s \mid 1 \leq s \leq S, k < t \leq T\right\}$ using (5)–(8). Let the estimated values for $a_t^s, b_t^s, c_t^s, d_t^s$ be denoted by $\widehat{a_t^s}, \widehat{b_t^s}, \widehat{c_t^s}, \widehat{d_t^s}$. Let $\widehat{f_t^s}$ and $\widehat{g_t^s}$ be the function values of f_t^s and g_t^s, respectively, for $k < t \leq T$ when the estimated values $\widehat{a_t^s}, \widehat{b_t^s}, \widehat{c_t^s}, \widehat{d_t^s}$ are used, i.e.,

$$\widehat{f_t^s}\left(u_{\mathrm{AR}_t}^s\right) = \widehat{a_t^s} + \widehat{b_t^s}\, u_{\mathrm{AR}_t}^s \text{ and } \widehat{g_t^s}\left(u_{\mathrm{AR}_t}^s\right) = \begin{cases} 0, & \text{if } \widehat{d_t^s}\, u_{\mathrm{AR}_t}^s > \widehat{c_t^s} \\ \widehat{c_t^s} - \widehat{d_t^s}\, u_{\mathrm{AR}_t}^s, & \text{if } \widehat{c_t^s} - 1 \leq \widehat{d_t^s}\, u_{\mathrm{AR}_t}^s \leq \widehat{c_t^s} \\ 1, & \text{if } \widehat{d_t^s}\, u_{\mathrm{AR}_t}^s < \widehat{c_t^s} - 1 \end{cases}.$$

  - Compute $u_{\mathrm{AR}_t}^s$ for $k < t \leq T$ by solving the following optimization problem:

$$\begin{array}{ll} \textit{maximize} & \displaystyle\sum_{t=1}^{k}\sum_{s=1}^{S} g_t^s\left(u_{\mathrm{AR}_t}^s\right) f_t^s\left(u_{\mathrm{AR}_t}^s\right) + \sum_{t=k+1}^{T}\sum_{s=1}^{S} \widehat{g_t^s}\left(u_{\mathrm{AR}_t}^s\right) \widehat{f_t^s}\left(u_{\mathrm{AR}_t}^s\right) \\ \textit{subject to} & \displaystyle\sum_{t=1}^{T}\sum_{s=1}^{S} u_{\mathrm{AR}_t}^s \leq U_{\mathrm{TR}} \\ & u_{\mathrm{AR}_t}^s \geq 0, \quad 1 \leq s \leq S,\ k \leq t \leq T \end{array}$$

  which is again a convex minimization problem similar to (4).

The k-step lookahead predictor can be used in an online fashion by the attacker for every successive value of k.

Application of Multi-Objective Distributed Constraint Optimization in Network Security

In the previous section we saw how to formulate and solve some problems related to the security of networked systems as a convex constraint optimization problem with a *single* objective function. In this section, we review the results of Okimoto et al. [18] that apply *multi-objective distributed* constraint optimization methods to formulate and solve problems related to security issues of networked systems. Okimoto et al. [18] do this by first formulating the security problem for networked system as a multi-objective distributed constraint optimization problem (MO-DCOP) using the formalization in [8], and then discussing some algorithmic approaches to solve such an optimization problem. Generally, multi-objective distributed constraint optimization methods are very suitable for formalizing applications related to multi-agent cooperation. An advantage of casting network security problems as an MO-DCOP is that multiple criteria (e.g., level of risk, loss of privacy, cost of operation) can be optimized *simultaneously* instead of separately; however, a disadvantage of this is that the resulting optimization problem may be computationally quite hard. The multi-objective distributed constraint optimization framework of [8] is an extension of *mono-objective* distributed constraint optimization framework in [16] for modelling applications related to multi-agent cooperation games.

The MO-DCOP proposed by Okimoto et al. [18] is described by the following parameters:

- A 5-tuple $\langle \mathcal{S}, \mathcal{X}, \mathcal{D}, \mathcal{C}, \mathcal{O} \rangle$ where
  - $\mathcal{S} = \{\,\mathsf{agent}_1, \mathsf{agent}_2, \dots, \mathsf{agent}_n\,\}$ is a set of n agents. An agent may be a human, a program, an organization, a country, *etc.*
  - $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ is a set of n variables, where x_i is owned by agent_i.
  - $\mathcal{D} = \{D_1, D_2 \dots, D_n\}$ is a set of n discrete domains, where D_i is the domain of values of variable x_i. The notation (x_i, d_i) will be used to denote an assignment of value $d_i \in D_i$ to variable x_i.
  - $\mathcal{O} = \left\{O^1, O^2, \dots, O^m\right\}$ is a set of m criteria that is to be optimized.
  - $\mathcal{C} = \left\{C^1, C^2, \dots, C^m\right\}$ is a set of m sets of constraints, where C^ℓ is the set of constraints corresponding to the ℓth criterion $\mathcal{O}^\ell$. A constraint relation $\left(\bowtie_{i,k}, \bowtie_{j,k}\right)^\ell \in C^\ell$ (for $k = 1, 2, \dots$) denotes a constraint of the type $\left\{(x_i, d_i), \left(x_j, d_j\right)\right\}$ involving the variables x_i and x_j, and is used to describe the condition of cooperation of agent_i and agent_j on the objective C^ℓ.
- $\left\{f^\ell_{i,j,k} \colon D_i \times D_j \mapsto \mathbb{R} \;\middle|\; 1 \le \ell \le m, 1 \le i,j \le n, k \in \mathbb{N}^+\right\}$ is a given set of cost functions such that $f^\ell_{i,j,k}\left(d_i, d_j\right)$ gives, for each objective C^ℓ and pairs x_i, x_j such that $\left(\bowtie_{i,k}, \bowtie_{j,k}\right)^\ell = \left\{(x_i, d_i), \left(x_j, d_j\right)\right\} \in C^\ell$, the *cost* for an assignment (decision) $\left\{(x_i, d_i), \left(x_j, d_j\right)\right\}$.

$$\mathcal{S} = \{\mathsf{agent}_1, \mathsf{agent}_2, \mathsf{agent}_3\}$$
$$\mathcal{X} = \{x_1, x_2, x_3\}$$
$$\mathcal{D} = \{D_1, D_2, D_3\},\ \forall\, i\colon\ D_i = \{\ \textit{scan},\ \ \textit{ignore}\ \}$$
$$\mathcal{O} = \{\ \textit{risk},\ \ \textit{resource budget}\ \}$$

$$\mathcal{C} = \left\{C^1, C^2\right\}\colon\ \begin{array}{l} C^1 = \left\{\left\{\overbrace{(x_1,\ \textit{scan}\ ),(x_2,\ \textit{ignore}\ )}^{\bowtie^1_{1,2,1}}\right\},\left\{\overbrace{(x_2,\ \textit{ignore}\ ),(x_3,\ \textit{ignore}\ )}^{\bowtie^1_{2,3,1}}\right\}\right\} \\ C^2 = \left\{\left\{\underbrace{(x_1,\ \textit{scan}\ ),(x_2,\ \textit{scan}\ )}_{\bowtie^2_{1,2,1}}\right\},\left\{\underbrace{(x_1,\ \textit{scan}\ ),(x_2,\ \textit{ignore}\ )}_{\bowtie^2_{1,2,2}}\right\}\right\} \end{array}$$

x_1	x_2	x_3	ℓ	k	$f^{\ell}_{i,j,k}\left(d_i, d_j\right)$
$d_1 = \textit{scan}$	$d_2 = \textit{ignore}$		1	1	6
	$d_2 = \textit{ignore}$	$d_3 = \textit{ignore}$	1	1	3
$d_1 = \textit{scan}$	$d_2 = \textit{scan}$		2	1	7
$d_1 = \textit{scan}$	$d_2 = \textit{ignore}$		2	2	4

Fig. 1 A toy example for the Mo-Dcop framework of Okimoto et al. [18]. The *shaded row* indicates that when *resource budget* is the optimization criterion the cost of agent_1 opting to *scan* and agent_2 opting to *ignore* is 4

For a set $\mathcal{A}$ of variable-value assignments and an objective $\mathcal{O}_\ell$, the cost incurred in optimizing this objective is then given by:

$$R^{\ell}(\mathcal{A}) = \sum_{k} \sum_{\substack{\left(\bowtie_{i,k},\bowtie_{j,k}\right)^{\ell} = \left\{(x_i,d_i),\left(x_j,d_j\right)\right\} \in C^{\ell} \\ \left\{(x_i,d_i),\left(x_j,d_j\right)\right\} \subseteq \mathcal{A}}} f^{\ell}_{i,j,k}\left(d_i, d_j\right)$$

and the solution corresponding to this variable-value assignment $\mathcal{A}$ over all objectives is then characterized by the cost vector

$$\Re(\mathcal{A}) = \left(R^1(\mathcal{A}), R^2(\mathcal{A}), \ldots, R^m(\mathcal{A})\right)$$

A toy example of the above framework is depicted in Fig. 1 for the case of three agents *not* all pairs of which cooperate with each other all the time.

Although ideally one would like to find a solution that optimizes *all* the objective functions *simultaneously*, such a solution may not even exist and thus one would resort to trade-offs among various objectives. One way to handle such a trade-off is by adopting the concept of *Pareto optimality* from game theory [10] to the above Mo-Dcop formulation in the following manner.

Definition 1 A cost vector $\mathfrak{R}(\mathcal{A}) = \big(R^1(\mathcal{A}), R^2(\mathcal{A}), \ldots, R^m(\mathcal{A})\big)$ is said to (strictly) dominate another cost vector $\mathfrak{R}(\mathcal{A}') = \big(R^1(\mathcal{A}'), R^2(\mathcal{A}'), \ldots, R^m(\mathcal{A}')\big)$, denoted by $\mathfrak{R}(\mathcal{A}) \prec \mathfrak{R}(\mathcal{A}')$, if and only if both the following conditions hold:

- $R^\ell(\mathcal{A}) \leq R^\ell(\mathcal{A}')$ for $1 \leq \ell \leq m$, and
- there exists at least one $\ell \in \{1, 2, \ldots, m\}$ such that $R^\ell(\mathcal{A}) < R^\ell(\mathcal{A}')$.

A cost vector $\mathfrak{R}(\mathcal{A})$ is then called Pareto optimal solution if and only if there does not exist another feasible cost vector $\mathcal{A}'$ such that $R(\mathcal{A}') \prec R(\mathcal{A})$.

Note that Pareto optimal solutions need *not* be unique. Okimoto et al. term a Pareto optimal solution as a *trade-off solution* in [18]. Algorithms for computing Pareto optimal solutions appear in the traditional computer science literature under names such as the *maximal vector computation* problem [11, 12]. A *pseudo-tree* based algorithm for solving multi-objective distributed constraint optimization problems appears in [14]. Okimoto et al. [18] extend the algorithmic approach in [14] by adding a pre-processing phase to design a new *branch-and-bound* search algorithm (BnB) for finding *all* trade-off solutions[2] using the branch-and-bound technique with a depth-first-search strategy. For evolutionary (genetic) algorithms to solve multi-objective distributed constraint optimization problems, see [4, 6]. One can also design approximation algorithms (heuristics) for solving multi-objective distributed constraint optimization problems, e.g., see the *bounded multi-objective max-sum algorithm* in [8].

Optimization Problems in Security for Power Networks

Maintaining a *secure* electric power distribution and transmission system against malicious attacks is an extremely important issue since almost any modern society relies critically on the proper operation of these systems. In this chapter we review some basic optimization problems related to this issue, and the application of ℓ_1-relaxation techniques of Sou et al. [22] in solving these optimization problems.

To begin with, a power network model is one with the following components:

- The network topology is specified by a directed graph $G = (V, E)$ with n nodes (buses) and m arcs (transmission lines). The corresponding (directed) edge-node incidence matrix of the graph is denoted by $A \in \{-1, 0, 1\}^{n \times m}$ where
$$A[u, e] = \begin{cases} -1, \text{ if } e = (u, v) \in E \\ \;\;\,1, \text{ if } e = (v, u) \in E \,. \\ \;\;\,0, \text{ otherwise} \end{cases}$$

[2] Okimoto et al. [18] claim that an advantage of finding all trade-off solutions is that agents can *dynamically* change decisions in case of emergencies. Unfortunately, the number of trade-off solutions may be exponential in the worst case.

- The physical property of the network is described by a *nonsingular diagonal* matrix $D \in \mathbb{R}^{m\times m}$ such that the reactance of the transmission line (arc) e is $1/D[e,e]$.
- The states of the nodes of the network are summarized by a *state vector* $\theta \in [0,2\pi)^{n-1}$, assuming constant bus voltages throughout but non-constant bus phase angles and using one arbitrary bus (node) as a reference.
- Assuming the DC power flow model and under malicious data attacks, the *measurement vector* z of the states of the buses that is obtained by a state estimator

$$z = H\theta + \widehat{z} \quad \text{with} \quad H = \begin{pmatrix} PD\mathcal{A}^{\mathrm{T}} \\ Q\mathcal{A}D\mathcal{A}^{\mathrm{T}} \end{pmatrix} \tag{9}$$

where

  - $\widehat{z} \in \mathbb{R}^{n-1}$ is the vector of malicious data attacks [13].
  - $\mathcal{A} \in \mathbb{R}^{(n-1)\times m}$ is obtained from A by removing the row corresponding to the reference bus (node).
  - P is a subset of rows of an identity matrix of appropriate dimension indicating flow measurements of which arcs (transmission lines) are actually taken.
  - Q is a subset of rows of an identity matrix of appropriate dimension indicating power injection measurements of which nodes (buses) are actually taken.

Typically, θ is estimated using the values in H and z. Assuming that the network is *observable* (in control-theoretic terms), it is known that an estimate $\widehat{\theta}$ of θ can be obtained using the following equation where W is a positive definite diagonal matrix [1, 17]:

$$\widehat{\theta} = \left(H^{\mathrm{T}}WH\right)^{-1} WH^{\mathrm{T}}z$$

To detect possible malicious attacks against the measurements via $\widehat{z}$, the commonly performed test [1, 17] is used: *if the norm* $\| z - H\widehat{\theta}\,\|$ *of the following residual quantity*

$$z - H\widehat{\theta} = \left(I - H\left(H^{\mathrm{T}}WH\right)^{-1} WH^{\mathrm{T}}\right)\widehat{z}$$

is large then trigger the alarm.

Although the above test works well if there is a single malicious attack on one data measurement, it may fail under *coordinated* malicious attacks on *multiple* data measurements. For such scenarios, a notion of *security index* was introduced in by Sandberg et al. in [20]. Intuitively, a small security index implies that the power network is more vulnerable to malicious attacks. Let H_ℓ denote the ℓth row of H and the notation $\|X\|_0$ for a vector X denote the cardinality (number of non-zero elements) of X. The security index for the power flow measurement of the kth transmission line (arc) for a given k is formulated as the optimal objective value of the following optimization problem:

$$\begin{array}{rl} \textit{minimize} & \|H\,\mathbf{x}\|_0 \\ \textit{subject to} & H_k\,\mathbf{x} = 1 \\ & \mathbf{x} \in \mathbb{R}^{n-1} \end{array} \tag{10}$$

The more general case when certain measurements are *protected* in the sense that they are too secure to be attacked can easily be handled by extending (10) in the following manner [2, 5, 13]. Let $\mathcal{I} \subset \{1, 2, \ldots, m\}$ denote the indices of those transmission lines whose power flow measurements are protected and let $H_{\mathcal{I}}$ be the submatrix of H with rows indexed by $\mathcal{I}$. Then, (10) can be generalized to the following cardinality minimization problem:

$$\begin{array}{rl} \textit{minimize} & \|H\,\mathbf{x}\|_0 \\ \textit{subject to} & H_k\,\mathbf{x} = 1 \\ & H_{\mathcal{I}}\,\mathbf{x} = 0 \\ & \mathbf{x} \in \mathbb{R}^{n-1} \end{array} \tag{11}$$

In general, there are no efficient algorithms for solving cardinality minimization problems and thus heuristics are often employed. Sou et al. in [22] provide an efficient application of ℓ_1-relaxation techniques to solve an important special case of (11) that assumes $H = PD\mathcal{A}^{\mathrm{T}}$ instead of the more general form shown in (9). They prove that this special case is in fact contained in the following type of optimization problem of a more general nature:

$$\begin{array}{rl} \textit{minimize} & \|C_{\{1,2,\ldots,m\}\setminus\mathcal{I}}\;\mathbf{x}\|_0 \\ \textit{subject to} & C_k\,\mathbf{x} = 1 \\ & C_{\mathcal{I}}\,\mathbf{x} = 0 \\ & \mathbf{x} \in \mathbb{R}^{n-1} \end{array} \tag{12}$$

where

- $C \in \mathbb{R}^{m\times(n-1)}$ is a given *totally unimodular* matrix, i.e., a matrix whose every square submatrix has a determinant of 0, 1, or -1,
- for any subset $Y \subset \{1, 2, \ldots, m\}$ A_Y is the submatrix of A with rows indexed by Y, and
- A_k is the kth row of A.

Then, a ℓ_1-relaxation of (12) can be obtained by replacing the objective function $\|C_{\{1,2,\ldots,m\}\setminus\mathcal{I}}\;\mathbf{x}\|_0$ by the objective function $\|C_{\{1,2,\ldots,m\}\setminus\mathcal{I}}\;\mathbf{x}\|_1$ that uses the ℓ_1 norm. This ℓ_1-relaxation can in turn be written down as a linear program and solved optimally. Sou et al. in [22] prove that an optimal solution of this linear program is in fact also an optimal solution of (12).

Conclusion

In this chapter we have surveyed a few optimization frameworks for problems related to security of networked system such as the internet or power grid system. There are other frameworks of modelling network security issues that we have not considered in this chapter, such as game-theoretic formulations or in the context of quantum computing. We believe that as networked systems of various nature become more common in everyday transactions, the corresponding security issues will give rise to more challenging optimization research questions.

Acknowledgements The authors were partially supported by NSF grant IIS-1160995.

References

1. A. Abur, A. Expósito, *Power System State Estimation* (Marcel Dekker, Inc., New York, 2004)
2. R. Bobba, K. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, T. Overbye, Detecting false data injection attacks on dc state estimation, in *1st Workshop on Secure Control Systems* (2010)
3. S. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
4. K. Bringmann, T. Friedrich, F. Neumann, M. Wagner, Approximation-guided evolutionary multi-objective optimization, *22nd International Joint Conference on Artificial Intelligence*, pp. 1198–1203 (2011)
5. G. Dán, H. Sandberg, Stealth attacks and protection schemes for state estimators in power systems, *1st IEEE International Conference on Smart Grid Communications*, pp. 214–219 (2010)
6. K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evolut. Comput. **6**(2), 182–197 (2002)
7. D.M. Dunlavy, B. Hendrickson, T.G. Kolda, Mathematical challenges in cybersecurity. Technical Report SAND2009-0805, Sandia National Laboratories (2009)
8. F.M.D. Fave, R. Stranders, A. Rogers, N.R. Jennings, Bounded decentralised coordination over multiple objectives, *10th International Conference on Autonomous Agents and Multiagent Systems*, pp. 371–378 (2011)
9. S. Floyd, V. Paxson, Difficulties in simulating the Internet. IEEE/ACM Trans. Netw. **9**(4), 392–403 (2001)
10. D. Fudenberg, J. Tirole, *Game Theory* (MIT Press, Cambridge, 1991)
11. P. Godfrey, R. Shipley, J. Gryz, Algorithms and analyses for maximal vector computation. VLDB J. **16**, 5–28 (2006)
12. H.T. Kung, F. Luccio, F.P. Preparata, On finding the maxima of a set of vectors. J. ACM **22**(4), 469–476 (1975)
13. Y. Liu, M. Reiter, P. Ning, False data injection attacks against state estimation in electric power grids, *16th ACM Conference on Computer and Communication Security*, pp. 21–32 (2009)
14. T. Matsui, M. Silaghi, K. Hirayama, M. Yokoo, H. Matsuo, Distributed search method with bounded cost vectors on multiple objective dcops, in *15th International Conference on Principles and Practice of Multi-Agent Systems*, pp. 137–152 (2012)
15. J. Meza, S. Campbell, D. Bailey, Mathematical and Statistical Opportunities in Cyber Security. Report LBNL-1667E, Lawrence Berkeley National Laboratory (2009)
16. P. Modi, W.-M. Shen, M. Tambe, M. Yokoo, ADOPT: asynchronous distributed constraint optimization with quality guarantees. Artif. Intell. **161**(1–2), 149–180 (2005)

17. A. Monticelli, *State Estimation in Electric Power Systems A Generalized Approach* (Kluwer Academic Publishers, Boston, 1999)
18. T. Okimoto, N. Ikegai, K. Inoue, H. Okada, T. Ribeiro, H. Maruyama, Cyber security problem based on multi-objective distributed constraint optimization technique, in *43rd Annual IEEE/IFIP Conference on Dependable Systems and Networks Workshop*, pp. 1–7 (2013)
19. V. Paxson, Strategies for sound internet measurement, in *4th ACM SIGCOMM Conference on Internet measurement*, pp. 263–271 (2004)
20. H. Sandberg, A. Teixeira, K.H. Johansson, On security indices for state estimators in power networks, in *1st Workshop on Secure Control Systems* (2010).
21. A.J. Slagell, K. Lakkaraju, K. Luo, Flaim: A multi-level anonymization framework for computer and network logs, in *20th USENIX Large Installation System Administration Conference*, pp. 63–77 (2006)
22. K.C. Sou, H. Sandberg, K.H. Johansson, On the exact solution to a smart grid cyber-security analysis problem. IEEE Trans. Smart Grid **4**(2), 856–865 (2013)
23. K.G. Vamvoudakis, J.P. Hespanha, R.A. Kemmerer, G. Vigna, Formulating cyber-security as convex optimization problems, in *Control of Cyber-Physical Systems*. Lecture Notes in Control and Information Sciences, vol. 449 (Springer, Berlin, 2013), pp. 85–100
24. G. Vigna, The 2011 UCSB iCTF: Description of the game (2011), http://ictf.cs.ucsb.edu/

On Some Information Geometric Approaches to Cyber Security

C.T.J. Dodson

Abstract Various contexts of relevance to cyber security involve the analysis of data that has a statistical character and in some cases the extraction of particular features from datasets of fitted distributions or empirical frequency distributions. Such statistics, for example, may be collected in the automated monitoring of IP-related data during accessing or attempted accessing of web-based resources, or may be triggered through an alert for suspected cyber attacks. Information geometry provides a Riemannian geometric framework in which to study smoothly parametrized families of probability density functions, thereby allowing the use of geometric tools to study statistical features of processes and possibly the representation of features that are associated with attacks. In particular, we can obtain mutual distances among members of the family from a collection of datasets, allowing, for example, measures of departures from Poisson random or uniformity, and discrimination between nearby distributions. Moreover, this allows the representation of large numbers of datasets in a way that respects any topological features in the frequency data and reveals subgroupings in the datasets using dimensionality reduction. Here some results are reported on statistical and information geometric studies concerning pseudorandom sequences, encryption-decryption timing analyses, comparisons of nearby signal distributions and departure from uniformity for evaluating obscuring techniques.

Keywords Cyber security • Empirical frequency distributions • Pseudorandom sequences • Encryption-decryption timing • Proximity to uniformity • Nearby signals discrimination • Information geometry • Gamma distributions • Gaussian distributions • Dimensionality reduction

MSC 53B20, 62M86

C.T.J. Dodson (✉)
School of Mathematics, University of Manchester, Manchester M13 9PL, UK
e-mail: ctdodson@manchester.ac.uk

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_9

Introduction

The British Columbia Institute of Technology (BCIT) maintained until 2006 an industrial cyber security incident database (ISID) [8], designed to track incidents of a cyber security nature that directly affected industrial control systems and processes. Byres and Lowe [8, 9] pointed out that from 1980 to 2000 the cyber threat was evenly split among internal, external and accidental cases. By 2001 this had changed to 70% external threat sources, 20% accidental, 5% internal and 5% other. Of these the internal security incidents arose from the following entry points: business network 43%, human machine interface (HMI) 29%, physical access to equipment 21% and laptop 7%. Externally the percentages included attacks from: remote internet 36%, remote dial-up 20%, remote unknown 12%, VPN connection 8%, remote wireless 8%, the remainder from remote trusted third party, remote Telco network, remote supervisory control and data acquisition (SCADA) network. The consequences of these attacks were a production loss of 41% and loss of ability to control or view the plant. Between 1995 and 2000 the number of security incidents averaged 2 per year but that had increased linearly to 10 per year by 2003. The current successor to ISID is the Repository of Industrial Security Incidents [54], a database of incidents of a cyber security nature that have (or could have) affected process control, industrial automation or supervisory control and data acquisition (SCADA) systems. For a current view of the problem of criminal use of encrypted messaging systems on smartphones, see the New York District Attorney's report to the 6th Annual Financial Crimes and Cybersecurity Symposium at the Federal Reserve Bank of New York on 15 November 2015 [61], with a large bibliography. This sets out the current capabilities of smartphones and tablets and makes a number of proposals.

The UK government Centre for the Protection of National Infrastructure (CPNI) [19] and the USA Homeland Security [60] provide up-to-date information and advice on cyber security. Wang and Lu [62] provided a comprehensive study of cyber security needs for the next generation power systems, particularly network vulnerabilities, attack countermeasures, secure communication protocols and architectures in the Smart Grid. The UK Information Assurance Advisory Council (IAAC) [58] provides a wide range of documentation, including the latest Korea-UK Initiatives in Cyber Security Research report [59]. The proceedings of the recent international conference at the University of Piraeus [44] provides a collection of more than 30 articles on cyber warfare and security and the book [50] contains 17 articles treating various aspects of cybersecurity. Via the assistance of the 2014 US AMS Network Science Mathematical Research Community, Burstein et al. [7] studied the problem of increasingly frequent events of Border Gateway Protocol route hijacking for traffic interception. They developed an optimal information monitoring strategy based on an abstract model for routing networks in which colluding sets of agent nodes conspire to divert traffic via them by sending false distance information to honest agent nodes.

In this paper we offer some geometrical methods for application in problems of cyber security which can be addressed through statistical analyses of data. Information geometry provides a Riemannian geometric framework in which to study smoothly parametrized families of probability density functions, thereby allowing the use of geometric tools to study statistical features of processes. Geometrical provision of this kind has proved an enormous advantage in theoretical physics and conversely, physical problems have stimulated many advances in differential geometry, global analysis and algebraic geometry. The geometrization of statistical theory [1–5, 22] has had similar success and its role in applications is now widespread and generating new developments of theory, algorithms and computational information geometry [48, 49]. We give a brief introduction to information geometry in sections "The Fisher Information Metric" and "Exponential Family of Distributions", which is sufficient for the understanding of techniques in the sequel. We outline the information geometry of univariate and multivariate Gaussians in section "Information Geometry of Gaussians", which we use in section "Dimensionality Reduction Methods". Situations in which such methods are relevant to cyber security include discrimination between nearby signal distributions, comparisons of real signal distributions with those obtained via random number generators in testing obscuring procedures, and in testing for anomalous behaviour, for example using departures from uniformity or independence.

One aspect of cyber security is concerned with the analysis of the stochastic process of attack events [21]. Such analyses can yield valuable data on the frequency distributions of attacks and these may be amenable to study using information geometric methods. In particular, spacings between events of interest may be representable via gamma distributions, since they span a range of behaviour from clustered through random (i.e. Poisson) to dispersed, Fig. 1; we discuss their information geometry in section "Information Geometry of the Gamma Manifold". Gamma distributions have the property that the standard deviation is proportional to the mean, characterized in Theorem 1 below, and they include a representation of Poisson processes through the 1-parameter family of exponential distributions; this is represented in Fig. 2. Their information geometry was used in a variety of applications [5, 24]. In a range of contexts in cryptology for encoding, decoding or for obscuring procedures, sequences of pseudorandom numbers are generated. Tests for randomness of such sequences have been studied extensively and the NIST Suite of tests [53] for cryptological purposes is widely employed. Information theoretic methods also are used, for example see Czregorzewski and Wieczorkowski [20] also Ryabko and Monarev [55] and the references therein for recent work. Covert timing channels operate by establishing an illegitimate communication channel between two processes and transmitting information via timing modulation, violating the underlying system's security policy. Recent studies have shown the vulnerability of popular computing environments, such as cloud, to these covert timing channels. Chen and Venkataramani [18] proposed an algorithm to detect the possible presence of covert timing channels on shared hardware that use contention-based patterns for communication. They obtained an event density histogram to

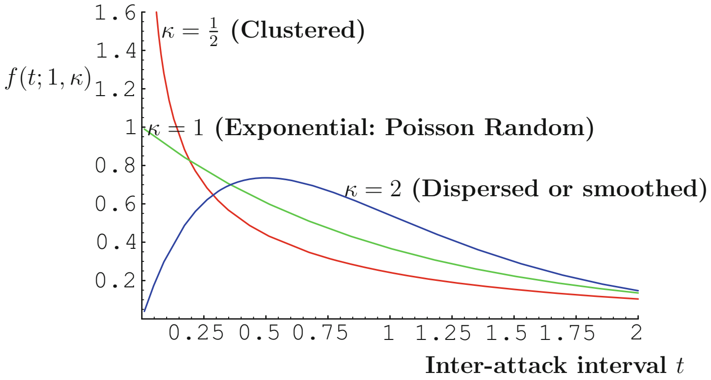


Fig. 1 Probability density functions, $f(t;\mu,\kappa)$, for gamma distributions of inter-attack intervals t with unit mean $\mu = 1$, and $\kappa = \frac{1}{2},\ 1,\ 2$. The case $\kappa = 1$ corresponds to an exponential distribution from an underlying Poisson process; $\kappa \neq 1$ represents some organization—clustering or dispersion

represent the probability distribution of event density and compared this to a Poisson process. We show in section "Statistics of Finite Random Spacing Sequences" how pseudorandom sequences may be tested using information geometry by using distances in the gamma manifold to compare maximum likelihood parameters for separation statistics of sequence elements.

In practical signal comparison situations [28], we obtain statistical data for an observable that is defined on some finite interval. We shall use as our model the family of log-gamma probability density functions, Fig. 3, defined for random variable $a \in (0,1]$ in section "Neighbourhoods of Uniformity in the Log-Gamma Manifold". The choice of log-gamma model is due to the fact that it contains a neighbourhood of the uniform distribution, and it has approximations to Gaussians truncated to domain $(0,1]$ and with arbitrarily small variance. The role of these functions in testing we discuss in section "Testing Nearby Signal Distributions and Drifts from Uniformity".

Encryption devices may be attacked by electromagnetic sensors that can extract information on the timing of processes for a chosen range of input data values. Given some knowledge of the software architecture, timings of operations typically relate to modular exponentiation steps, associated with the processing of the binary bits in the encryption key. This is discussed in section "Protecting Devices with Obscuring Techniques". In practice, clues to such timing information can be obtained from data on power consumption using electromagnetic sensors, possibly needing statistical processes to clean the data of noise. Kocher et al. [40] showed the effectiveness of Differential Power Analysis (DPA) in breaking encryption procedures using correlations between power consumption and data bit values during processing, claiming that most smart cards revealed their DES keys using fewer than 15 power traces. A practicable defence is to obscure the power usage data on timing information by spurious other processes. Then the effectiveness of such obscuring techniques can be evaluated using analyses of the distributions associated with time series from power usage. For example, a time series of power consumption using appropriately chosen thresholding and interval windows would yield a barchart and that would ideally be like that arising from Poisson processes, which for a given mean are maximally disorderd [36]. Information geometry can be used to measure differences from the Poisson model, equivalently from its associated exponential distribution—note that Grzegorzewski and Wieczorkowski [20] provided a detailed analysis of their entropy-based goodness-of-fit test for exponentiality.

Evaluation of cyber security may involve also identifying potentially anomalous behaviour in internet traffic on a network [47, 51], thus requiring extraction of appropriate features from a large data set of event frequency distributions. sometimes we can fit standard models to the empirical frequency distributions using maximum likelihood methods as illustrated for gamma distributions in section "Information Geometry of the Gamma Manifold". In the absence of a model family of distributions for which we have expressions for the information distances among the members, we can use the symmetrized Kullback–Leibler relative entropy expression, Eq. (40), to measure distance between empirical frequency distributions. Once we have extracted distance measures between all pairs of datasets we can use multi-dimensional scaling, or dimensionality reduction, to extract the three most significant features from the data set so that all samples can be displayed graphically in a 3-dimensional plot. The aim is to reveal groupings of data points that correspond to the prominent characteristics, the methodology is discussed in section "Dimensionality Reduction Methods".

Such a dimensionality reduction can reveal anomalous behaviour of a process by taking account of the true curved geometry of the data set, rather than displaying it as uncurved in a Euclidean geometry (cf. [12, Fig. 3.2]). The significance is that any non-obvious global topology of frequency connectivity in the data is revealed by the pattern of mutual separations in the embedding. An illustration using router traffic on the Abilene network showed how anomalous behaviour unseen by local methods could be picked up through dimensionality changes (cf. [12, Fig. 3.10]). Moreover, in document classification, the information metric approach outperformed standard Principal Component Analysis and Euclidean embeddings [13], and it outperformed traditional approaches to video indexing and retrieval with real world data [17]. In section "Dimensionality Reduction Methods" we outline how autocovariance extraction from time series data may be studied using information geometry and dimensionality reduction; we described an application to datasets of stochastic textures from 2-dimensional pixel arrays in [27].

We begin here by outlining the method to compute the Fisher information metric on a smoothly parametrized family of probability density functions, then illustrate it with explicit expressions for some important examples.

The Fisher Information Metric

Let $\Theta \subseteq R^n$ be the parameter space of an n-dimensional smooth family of probability density functions defined on some fixed d-dimensional event space $\Omega \subseteq R^d$, so we have a set

$$\{p_\theta | \theta \in \Theta\} \text{ with } p_\theta \geq 0 \text{ and } \int_\Omega p_\theta = 1 \quad \text{for all } \theta \in \Theta.$$

A fundamental property of a probability density function p_θ is its Shannon entropy, which is the negative of the expectation of its log-likelihood function, $l = \log p_\theta$, namely:

$$S(p_\theta) = -\int_\Omega p_\theta \log(p_\theta). \tag{1}$$

The derivatives of the log-likelihood function, $l = \log p_\theta$, yield a matrix function on Ω and the expectation of its entries is

$$g_{ij} = \int_\Omega p_\theta \left(\frac{\partial l}{\partial \theta^i} \frac{\partial l}{\partial \theta^j} \right) = -\int_\Omega p_\theta \left(\frac{\partial^2 l}{\partial \theta^i \partial \theta^j} \right), \tag{2}$$

for coordinates (θ^i) about $\theta \in \Theta \subseteq \mathbb{R}^n$.

This gives rise to a positive definite matrix depending only on the parameters, inducing a Riemannian metric structure g as a positive-definite symmetric quadratic form, on the space Θ of parameters (θ^i). From the construction of (2), a smooth invertible transformation of random variables, that is of the labelling of the points in the event space Ω while keeping the same parameters (θ^i), will leave the Riemannian metric unaltered. Formally, it induces a smooth diffeomorphism of manifolds that preserves the metric. This is a Riemannian isometry and the diffeomorphism is simply the identity map on parameters [5]. We shall see this explicitly below for the case of the log-gamma distribution section "Neighbourhoods of Uniformity in the Log-Gamma Manifold" and its associated Riemannian manifold.

The elements in the matrix (2) define the arc length function

$$ds^2 = \sum_{i,j} g_{ij}\, d\theta^i\, d\theta^j, \text{ often abbreviated to } ds^2 = g_{ij}\, d\theta^i\, d\theta^j \tag{3}$$

using the convention to sum over repeated indices.

The metric (2) is called the expected information metric or Fisher–Rao or Fisher metric for the manifold obtained from the family of probability density functions; the original ideas are due to Fisher [32] and Rao [52]. The second equality in Eq. (2) depends on certain regularity conditions [56] but when it holds it can be particularly convenient to use. Amari [1, 2, 4] and Amari and Nagaoka [3] provide accounts of the differential geometry that arises from the information metric. A wide range of applications is studied in [5, 24].

Exponential Family of Distributions

An n-dimensional parametric statistical model $\Theta \equiv \{p_\theta | \theta \in \Theta\}$ is said to be an exponential family or of exponential type, when the probability density functions in the family can be expressed in terms of functions $\{C, F_1, \ldots, F_n\}$ on Ω and a function φ on Θ as:

$$p(x; \theta) = e^{\{C(x) + \sum_i \theta_i F_i(x) - \varphi(\theta)\}}, \tag{4}$$

then we say that (θ_i) are its *natural parameters*, and φ is the potential function. From the normalization condition $\int p(x; \theta)\, dx = 1$ we obtain:

$$\varphi(\theta) = \log \int_\Omega e^{\{C(x) + \sum_i \theta_i F_i(x)\}}\, dx. \tag{5}$$

This potential function is therefore a distinguished function of the coordinates (θ_i) alone. Dodson and Matsuzoe [25] showed that use can be made of it for the presentation of the manifold as an immersion in $\mathbb{R}^{n+1}$ as follows:

$$\Phi : \mathbb{R}^n \to \mathbb{R}^{n+1} : (\theta_i) = \theta \mapsto (\theta, \varphi(\theta)). \tag{6}$$

With $\partial_i = \frac{\partial}{\partial \theta^i}$, we use from section "The Fisher Information Metric" the log-likelihood function $l(\theta, x) = \log(p_\theta(x))$ to obtain

$$\partial_i l(\theta, x) = F_i(x) - \partial_i \varphi(\theta)$$

and

$$\partial_i \partial_j l(\theta, x) = -\partial_i \partial_j \varphi(\theta).$$

The information metric g on the n-dimensional space of parameters $\Theta \subset \mathbb{R}^n$, equivalently on the set $S = \{p_\theta | \theta \in \Theta \subset \mathbb{R}^n\}$, has coordinates:

$$[g_{ij}] = -\int_\Omega [\partial_i \partial_j l(\theta, x)]\, p_\theta(x)\, dx = \partial_i \partial_j \varphi(\theta) = \varphi_{ij}(\theta). \tag{7}$$

Then, (S, g) is a Riemannian n-manifold. Distances between points in this manifold are computed as the geodesic length between the points, which is the infimum over all curves joining the points and in general difficult to obtain analytically [5, 26].

Information Geometry of Gaussians

The family of univariate normal or Gaussian density functions has event space $\Omega = \mathbb{R}$ and probability density functions given by

$$N \equiv \{N(\mu, \sigma^2)\} = \{n(x; \mu, \sigma) | \, \mu \in \mathbb{R}, \, \sigma \in \mathbb{R}^+\} \tag{8}$$

depending smoothly on the parameters mean μ and variance σ^2. So topologically, $N = \mathbb{R} \times \mathbb{R}^+$ is the upper half-plane, and the random variable is $x \in \Omega = \mathbb{R}$ with

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9}$$

The mean μ and standard deviation σ are commonly used as a local coordinate system (μ, σ). However, the univariate Gaussians (9) are of exponential type with natural coordinates $\theta_1 = \frac{\mu}{\sigma^2}$ and $\theta_2 = -\frac{1}{2\sigma^2}$. Then $(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$ is a natural coordinate system and

$$\varphi = -\frac{{\theta_1}^2}{4\,\theta_2} + \frac{1}{2}\,\log(-\frac{\pi}{\theta_2}) = \frac{\mu^2}{2\,\sigma^2} + \log(\sqrt{2\,\pi}\,\sigma) \tag{10}$$

is the corresponding potential function defined in section “Exponential Family of Distributions”.

Multivariate Gaussians for a given $k = 2, 3, \ldots\ldots, N$ are parametrized by a k-vector of means $\mu \in \mathbb{R}^k$ and a symmetric $k \times k$ covariance matrix $\Sigma \in \mathbb{R}^{(k^2+k)/2}$. There is no analytic expression for the information distance between two general k-variate Gaussians. What we have analytically are natural norms, on the space of means and on the space of covariances, giving the information distance $D(f^A, f^B)$ between two k-Gaussians $f^A(\mu^A, \Sigma^A)$ and $f^B(\mu^B, \Sigma^B)$ in two particular cases:

$\Sigma^A = \Sigma^B = \Sigma$: The common positive definite symmetric quadratic form Σ gives a norm on the difference vector of means:

$$D_\mu(f^A, f^B) = \sqrt{(\mu^A - \mu^B)^T \cdot \Sigma^{-1} \cdot (\mu^A - \mu^B)}. \tag{11}$$

$\mu^A = \mu^B = \mu$: A positive definite symmetric matrix constructed from the two covariance matrices Σ^A and Σ^B is

$$S^{AB} = {\Sigma^A}^{-1/2} \cdot \Sigma^B \cdot {\Sigma^A}^{-1/2}, \;\; \text{with } \{\lambda_j^{AB}\} = \mathrm{Eig}(S^{AB})$$

and it gives a norm on the space of differences between covariances [6] so we have

$$D_\Sigma(f^A, f^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{k} \log^2(\lambda_j^{AB})}. \tag{12}$$

In principle, (12) yields all of the true geodesic distances since the information metric is invariant under affine transformations of the mean [6, Appendix 1]; see also the article of Eriksen [30].

In practice, an approximate distance that is monotonically related to the true information distance may serve the purpose at hand, for example the symmetric sum:

$$D(f^A, f^B) \approx \frac{1}{2}\left(D_\mu(f^A, f^B) + D_\mu(f^B, f^A)\right) + D_\Sigma(f^A, f^B),$$

which adds to the value of (12) the average of (11) using both available covariances.

Information Geometry of the Gamma Manifold

The smooth family of gamma probability density functions is given by

$$f : [0, \infty) \to [0, \infty) : x \mapsto \frac{e^{-\frac{x\kappa}{\mu}} x^{\kappa-1} \left(\frac{\kappa}{\mu}\right)^{\kappa}}{\Gamma(\kappa)} \quad \mu, \kappa > 0. \tag{13}$$

Here μ is the mean, and the variance σ^2 is $\frac{\mu^2}{\kappa}$. So σ is proportional to the mean and the coefficient of variation, $\frac{1}{\sqrt{\kappa}}$, is unity in the case that (13) reduces to the exponential distribution. Thus, $\kappa = 1$ corresponds to an underlying Poisson random process complementary to the exponential distribution. When $\kappa < 1$ the random variable X represents spacings between events that are more clustered than for a Poisson process and when $\kappa > 1$ the spacings X are more evenly distributed than for Poisson. The case when $\mu = n$ is a positive integer and $\kappa = 2$ gives the Chi-Squared distribution with $n-1$ degrees of freedom; this is the distribution of $\frac{(n-1)s^2}{\sigma_G^2}$ for variances s^2 of samples of size n taken from a Gaussian population with variance σ_G^2. Illustrations of some gamma probability density functions are given in Fig. 1. The gamma distribution is of exponential type, as we see by making the substitution of parameters $(\mu, \kappa) \mapsto (\nu = \frac{\kappa}{\mu}, \kappa)$. Then the probability density functions have the form

$$p(x; \nu, \kappa) = \nu^\kappa \frac{x^{\kappa-1} e^{-x\nu}}{\Gamma(\kappa)}. \tag{14}$$

In this case (ν, κ) is a natural coordinate system of the 1-connection and

$$\varphi(\theta) = \log \Gamma(\kappa) - \kappa \log \nu \tag{15}$$

is the corresponding potential function defined in section "Exponential Family of Distributions". An embedding of the gamma manifold in $\mathbb{R}^3$ using (15) in (6)

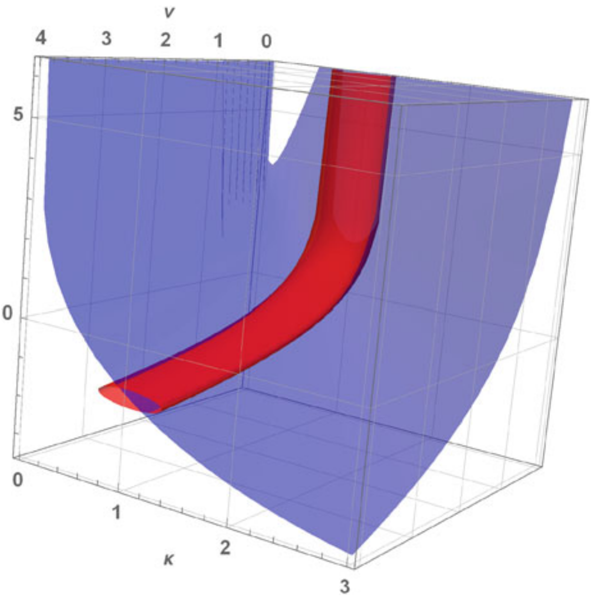


Fig. 2 An embedding of the gamma manifold, as a surface in $\mathbb{R}^3$ using (15) in (6) including also a tubular neighbourhood of the exponential distributions, which all lie on the curve $\kappa = 1$

is shown in Fig. 2, including also a tubular neighbourhood of the exponential distributions, which all lie on the curve $\kappa = 1$.

The Fisher metric is given by the Hessian of φ, that is, with respect to natural coordinates:

$$\left[g_{ij}\right](\nu,\kappa) = \left[\frac{\partial^2\varphi(\theta)}{\partial\theta_i\partial\theta_j}\right] = \begin{bmatrix} \frac{\kappa}{\nu^2} & -\frac{1}{\nu} \\ -\frac{1}{\nu} & \psi''(\kappa) \end{bmatrix} = \begin{bmatrix} \frac{\kappa}{\nu^2} & -\frac{1}{\nu} \\ -\frac{1}{\nu} & \frac{d^2}{d\kappa^2}\log(\Gamma) \end{bmatrix}. \tag{16}$$

In terms of the original parameters (μ,κ) in (13) the metric turns out to be diagonal:

$$\left[g_{ij}\right](\mu,\kappa) = = \begin{bmatrix} \frac{\kappa}{\mu^2} & 0 \\ 0 & \frac{d^2}{d\kappa^2}\log(\Gamma) - \frac{1}{\kappa} \end{bmatrix}. \tag{17}$$

So the coordinates (μ,κ) yield an orthogonal basis of tangent vectors, which is useful in calculations because then the arc length function is simply

$$ds^2 = \frac{\kappa}{\mu^2}\,d\gamma^2 + \left(\left(\frac{\Gamma'(\kappa)}{\Gamma(\kappa)}\right)' - \frac{1}{\kappa}\right)d\kappa^2.$$

We note the following important uniqueness property:

Theorem 1 (Cf [24, 37, 43]) *For independent positive random variables with a common probability density function f, having independence of the sample mean and the sample coefficient of variation is equivalent to f being the gamma distribution.*

A proof of Theorem 1 was given in [37] but in fact the result seems to have been known much earlier and in [24] we gave a proof partly based on the 1954 article by Laha [43], of interest for the methodology using Laplace Transforms, which are related to moment generating functions in statistics [31]. Other useful information geometric results involve very commonly occurring distributions and have been applied in a number of areas [5, 24]. Some may have application in measuring and representing statistical processes relevant to cyber security when there is a need to study the neighbourhood of a target distribution:

Theorem 2 *Every neighbourhood of a Poisson process contains a neighbourhood of processes subordinate to gamma probability density functions.*

Theorem 3 *Every neighbourhood of a uniform process contains a neighbourhood of processes subordinate to log-gamma probability density functions.*

Theorem 4 *Every neighbourhood of an independent pair of identical Poisson processes contains a neighbourhood of bivariate processes subordinate to Freund bivariate exponential probability density functions.*

Theorem 5 *The 5-dimensional space of bivariate Gaussians admits a 2-dimensional subspace through which can be provided a neighbourhood of independence for bivariate Gaussian processes.*

Theorem 6 *Via the Central Limit Theorem, by continuity, the tubular neighbourhoods of the curve of zero covariance for bivariate Gaussian processes will contain all limiting bivariate processes sufficiently close to the independence case for all processes with marginals that converge in probability density function to Gaussians.*

The characterizing property in Theorem 1 is one of the main reasons for the large number of applications of gamma distributions: many families of near-random natural processes have standard deviation approximately proportional to the mean [5], increasing together with time or changing ambient conditions, and this is easily tested for in practice. To fit a gamma distribution to data we can obtain the maximum likelihood parameter values $\hat{\mu}, \hat{\kappa}$, as follows.

Given a set of identically distributed, independent data values $X_1, X_2, \ldots, X_n$, the ‘maximum likelihood’ or ‘maximum entropy’ parameter values $\hat{\mu}, \hat{\kappa}$ for fitting the gamma distribution (13) are computed in terms of the mean and mean logarithm of the X_i by maximizing the likelihood function

$$L_f(\mu, \kappa) = \prod_{i=1}^{n} f(X_i; \mu, \kappa).$$

By taking the logarithm and setting the gradient to zero we obtain

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{18}$$

$$\begin{aligned}\log\hat{\kappa} - \frac{\Gamma'(\hat{\kappa})}{\Gamma(\hat{\kappa})} &= \log\bar{X} - \frac{1}{n}\sum_{i=1}^{n}\log X_i \\ &= \log\bar{X} - \overline{\log X}. \end{aligned} \tag{19}$$

Neighbourhoods of Uniformity in the Log-Gamma Manifold

The log-gamma density (21) actually arises from the gamma density

$$f(x,\mu,\kappa) = \frac{x^{\kappa-1}\,\mu^{\kappa}}{\Gamma(\kappa)}\,e^{-x\mu}. \tag{20}$$

via the change of variable from $x \in \mathbb{R}^{+}$ to $a \in (0,1]$ via $x = -\log a$.

The smooth family of log-gamma distributions has probability density functions of form

$$P(a,\mu,\kappa) = \frac{a^{\mu-1}\mu^{\kappa}\left|\log\left(\frac{1}{a}\right)\right|^{\kappa-1}}{\Gamma(\kappa)} \tag{21}$$

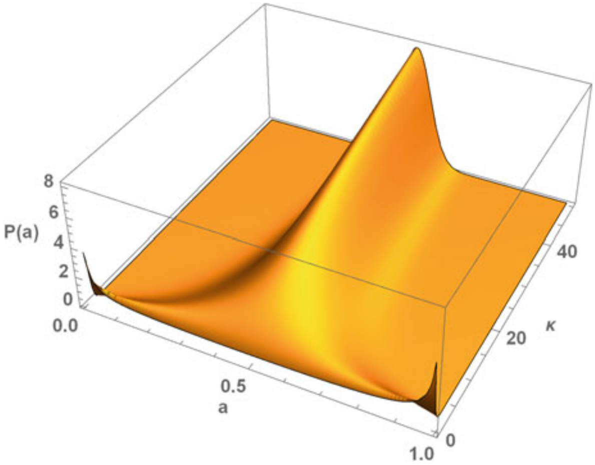


Fig. 3 The log-gamma family of probability densities (21) with central mean $\overline{a} = \frac{1}{2}$ as a surface. The surface tends to the delta function as $\kappa \to \infty$ and coincides with the constant 1 at $\kappa = 1$

for random variable $a \in (0, 1]$ and parameters $\mu, \kappa > 0$, see Fig. 10. The mean and variance are given by

$$\overline{a} = E(a) = \left(\frac{\mu}{1+\mu}\right)^{\kappa} \tag{22}$$

$$\sigma^2(a) = \left(\frac{\mu}{\mu+2}\right)^{\kappa} - \left(\frac{\mu}{1+\mu}\right)^{2\kappa}. \tag{23}$$

In this family the locus of those with central mean $E(a) = \frac{1}{2}$ satisfies

$$\mu(2^{\frac{1}{\kappa}} - 1) = 1 \tag{24}$$

shown in Fig. 3; the uniform density is the special case with $\kappa = \mu = 1$. Figure 4 shows a spherical neighbourhood in Euclidean space centred on the point at the uniform density on $(0, 1]$ in the curved surface representing all log-gamma distributions; the uniform density is represented by the point $\nu = 1, \kappa = 1$. This provides a method to measure departures from uniformity. The important information geometric property is that the Riemannian manifold of gamma distributions is isometric to that of log-gamma distributions, this is discussed with applications in [5].

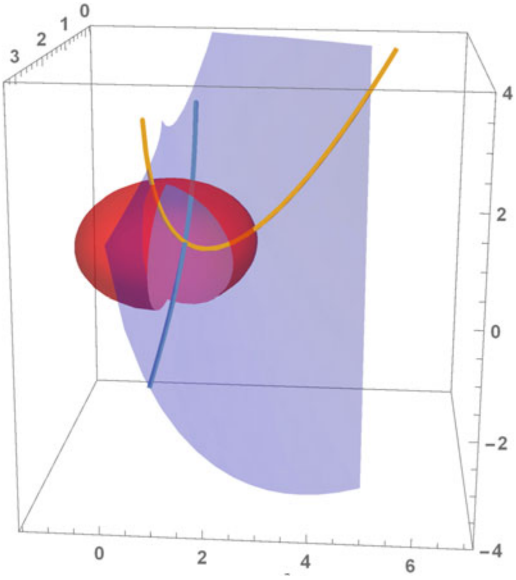


Fig. 4 An affine immersion in Euclidean space $\mathbb{R}^3$ of the curved surface of log-gamma probability densities on (0, 1]. The two curves in the surface represent the log-gamma distributions with $\nu = 1$ and $\kappa = 1$, and the spherical neighbourhood in $\mathbb{R}^3$ is centred on their intersection, which point represents the uniform distribution

Neighbourhoods of Randomness in the Gamma Manifold

Cyber software sometimes uses pseudorandom number generators to provide seed numbers for algorithms and sequences for testing procedures or for comparison with application sequences in attempts to obscure internal processes, which otherwise might reveal clues to the timing of underlying operations through power traces. Cryptological attacks on encryption/decryption devices may be defended against by obscuring algorithms that overlay randomizing procedures; then there is a need to compare nearby signal distributions and again the information metric can help. The Poisson distribution of events on a line is such that the probability of an event in an interval depends only on the size of the interval, not on the position of the interval in the line. Then the distribution of lengths of intervals between successive events is easily shown to be exponential, and that distribution has maximal entropy within the family of gamma distributions, so it is maximally haphazard and involves fewest assumptions.

Tests for randomness of number sequences have been studied extensively and the NIST Suite of tests [53] for cryptological purposes is widely employed. Information theoretic methods also are used, for example see [20] also [55] and the references therein for recent work. In [24] we added to the latter by outlining how finite length pseudorandom sequences may be tested quickly and easily using information geometry by computing distances in the gamma manifold to compare maximum likelihood parameters for separation statistics of sequence elements. A Poisson process defines a unique exponential distribution, the exponential distributions with different means are special cases of gamma distributions and the information geometry of the gamma family determines a metric structure for neighbourhoods, cf. Fig. 2, of the 1-parameter curve of exponential distributions in the Riemannian manifold of gamma distributions [5].

Mathematica [63] simulations were made of Poisson line processes using random number sequences of length $n = 100{,}000$ for which spacing statistics were computed [24]. Figure 5 shows maximum likelihood gamma parameter κ values from the simulations. The surface height in Fig. 6 represents upper bounds on information geometric distances from $(\mu, \kappa) = (511, 1)$ in the gamma manifold. This employs the approximate geodesic mesh function we described in Arwini and Dodson [5].

$$Distance[(511,1),(\mu,\kappa)] \leq \left| \frac{d^2 \log \Gamma}{d\kappa^2}(\kappa) - \frac{d^2 \log \Gamma}{d\kappa^2}(1) \right| + \left| \log \frac{511}{\mu} \right|. \tag{25}$$

The points shown in Fig. 6 are maximum likelihood gamma parameters from the *Mathematica* simulations of Poisson random processes of 100,000 events with expected separation $\mu = 511$. In the data from 500 such simulations the ranges of maximum likelihood gamma parameters were

$$419 \leq \mu \leq 643 \text{ and } 0.62 \leq \kappa \leq 1.56.$$

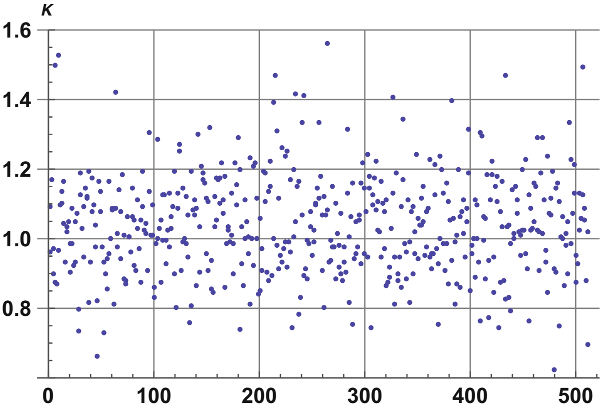


Fig. 5 The problem of finite length pseudorandom sequences. Maximum likelihood gamma parameter κ fitted to separation statistics for simulations of Poisson line processes of length 100,000 with expected parameter $\kappa = 1$. These simulations used the pseudorandom number generator in Mathematica [63]. In the limit, as the sequence length tends to infinity we expect the gamma parameter μ to tend to 1

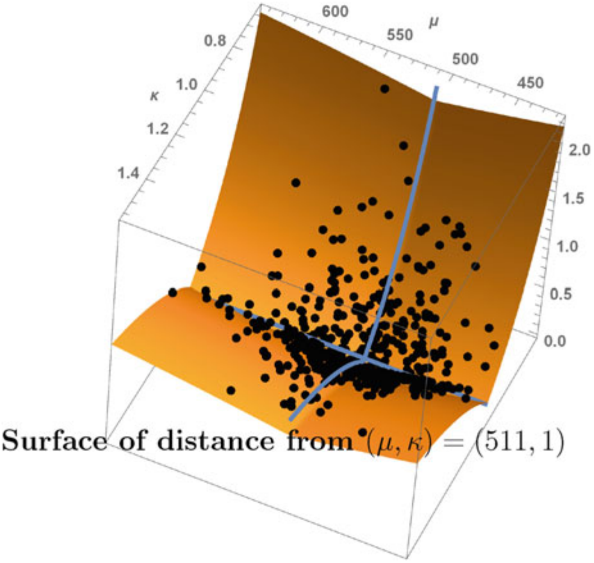


Fig. 6 Distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from the target $(\mu, \kappa) = (511, 1)$ from Eq. (25). Also shown are data points of maximum likelihood gamma parameters for interval lengths between events from simulations of Poisson random processes of 100,000, for events with expected separation $\mu = 511$. In the limit as the sequence length tends to infinity we expect the gamma parameter κ to tend to 1

So, even with strings of 100,000 nominally random numbers, there is considerable variability from the expected distribution of spacings in the sorted strings. Clearly as the sequence length tends to infinity we expect the gamma parameter κ to tend to 1. However, finite sequences must be used in real applications and then provision of a metric structure allows us, for example, to compare real sequence generating procedures against an ideal Poisson random model in the space of gamma distributions. Representations like that in Fig. 6 could be used to trigger an alert or action when an automated updating sequence of events abnormally strays over a chosen threshold of distance from a target.

Statistics of Finite Random Spacing Sequences

In the perfectly *random* case of haphazard allocation of events along a time line, the result is an exponential distribution of inter-event intervals when the line is infinite. However, for finite length processes it is a little more involved and we need to analyse this first in order to provide our reference structure.

Think of a sequence of different events among which we have distinguished one, represented by the letter X, while all others are represented by ?. In the cyber security context, X is the event of an attack. The relative abundance of X is given by the probability p that an arbitrarily chosen location in the sequence has an occurrence of X. Then $1-p$ is the probability that the location contains a different event from X. If the locations of X are chosen with uniform probability subject to the constraint that the net density of X in the chain is p, then either X happens or it does not; we have a binomial process.

It follows that, in a sequence of n events, the mean or expected number of occurrences of X is np and its variance is $np(1-p)$, but it is not immediately clear what will be the distribution of lengths of intervals between consecutive occurrences of X. Evidently the distribution of such lengths r, measured in units of one location length also is controlled by the underlying binomial distribution.

We are interested in the probability of finding in a sequence of n events a subsequence of form

$$\underbrace{\cdots ?X\overbrace{?\cdots?}X?\cdots},$$

where the overbrace $\overbrace{\qquad}$ encompasses precisely r events that are not X (i.e. not cyber attacks) and the underbrace $\underbrace{\qquad}$ encompasses precisely n events, the whole sequence.

Derivation of the Distributions

In a sequence of n locations filled by events we consider the probability of finding a subsequence containing two X's separated by exactly r non-X ?'s, that is the occurrence of an inter-X space length r. In our case the random variable r is the interval between successive cyber attacks.

The probability distribution function $\mathbb{P}(r,p,n)$ for inter-X space length r reduces to the first expression below (26), which is a geometric distribution and simplifies to (27)

$$\mathbb{P}(r,p,n) = \frac{\left(p^2(1-p)^r(n-r-2)\right)}{\sum_{r=0}^{n-2}(p^2(1-p)^r(n-r-2))}, \tag{26}$$

$$= \frac{(1-p)^{1+r}\,p^2\,(n-r-2)}{-1+(1-p)^n+p\,(n+p-n\,p)}, \tag{27}$$

$$\text{for } r = 0, 1, \ldots, (n-2).$$

The mean $\bar{r}$ and standard deviation σ_r of the distribution (27) are given for $r = 0, 1, \ldots, (n-2)$, by

$$\bar{r} = \sum_{r=0}^{n-2} r\,\mathbb{P}(r,p,n)$$

$$= \frac{((1-p)^n\,(2+(-3+n)\,p)) + (-1+p)^2\,(-2+(-1+n)\,p)}{p\,((1-p)^n+p\,(n+p-n\,p)-1)} \tag{28}$$

$$\sigma_r = \sqrt{\left(\sum_{r=0}^{n-2} r^2\,\mathbb{P}(r,p,n)\right) - \bar{r}^2} \tag{29}$$

$$= \sqrt{\frac{(p-1)\left(-2\,(1-p)^{2\,n} - (1-p)^n\,N(r,p,n) - (p-1)^2\,(2+(n-1)\,p\,(n\,p-4))\right)}{p^2\,((1-p)^n+p\,(n+p-n\,p)-1)^2}},$$

where we make the abbreviation

$$N(r,p,n) = \left(4\,n\,p - 4 + (n-6)\,(n-1)\,p^2 + (n-2)^2\,(n-1)\,p^3\right) \tag{30}$$

$$\text{for } r = 0, 1, \ldots, (n-2).$$

The coefficient of variation is given by

$$cv_r = \frac{\sigma_r}{\bar{r}} = \frac{p\,((1-p)^n - 1 + p\,(n+p-n\,p))\,L(r,p,n)}{(1-p)^n\,(2+(n-3)\,p) + (p-1)^2\,((-1+n)\,p-2)}$$

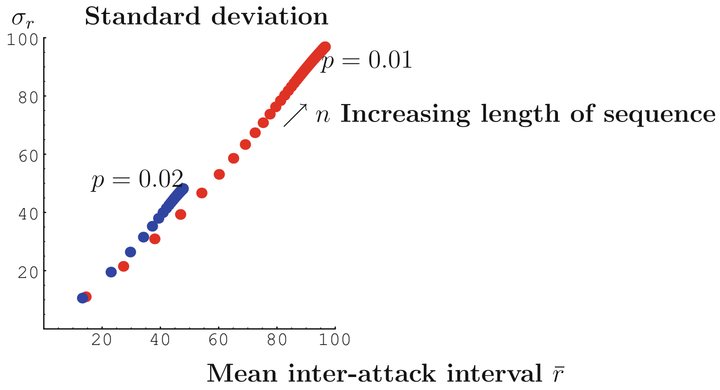


Fig. 7 Effect of sequence length n in random event sequences of length from $n = 50$ to $n = 4000$ in steps of 50. Plot of standard deviation σ_r against mean $\bar{r}$ for inter-attack interval distributions (27). The mean probability for the occurrence of an attack is $p = 0.01$ (*right*, *red*) and $p = 0.02$ (*left*, *blue*). The standard deviation is roughly equal to the mean; mean and standard deviation increase monotonically with increasing n

where we make the abbreviations

$$L(r,p,n) = \sqrt{\frac{(1-p)\left(2\,(1-p)^{2n} + (1-p)^n\, M(r,p,n) + (p-1)^2\ (2+(n-1)\ p\ (n\,p-4))\right)}{p^2\ ((1-p)^n + p\ (n+p-n\,p-1))^2}}$$

$$M(r,p,n) = \left(4(n\,p-1) + (n-6)\ (n-1)\ p^2 + (n-2)^2\ (n-1)\ p^3\right).$$

The two main variables are: the number n of events in the sequence, and the abundance probability p of occurrence of attacks. Their effects on the statistics of the distribution of inter-attack intervals are illustrated in Figs. 7 and 8, respectively. Figure 9 plots the maximum likelihood gamma parameter κ against the mean inter-attack interval $\bar{r}$.

Testing Nearby Signal Distributions and Drifts from Uniformity

In practical signal comparison situations [28], we obtain statistical data for an observable that is defined on some finite interval. We shall use as our model the family (21) of log-gamma probability density functions defined for random variable $a \in (0,1]$. The choice of log-gamma model is due to the fact that it contains a neighbourhood of the uniform distribution, illustrated in Fig. 4, namely for parameter values near $(\mu,\kappa) = (1,1)$ in (21). Also, for parameter values $\kappa >> 1$ it has approximations to Gaussians truncated to domain $(0,1]$ and with

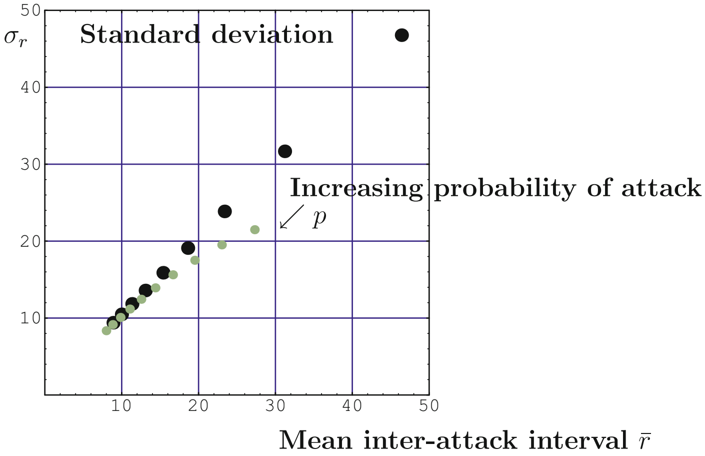


Fig. 8 Effect of attack probability p, over the range $0.01 \leq p \leq 0.1$ in steps of 0.01. Plot of standard deviation σ_r against mean $\bar{r}$ for inter-attack interval in random sequences of length $n = 200$ events (*light green*) and length $n = 500$ events (*dark green*), with probability $0.01 \leq p \leq 0.1$ for occurrence of attack. The standard deviation is for many practical purposes proportional to the mean; mean and standard deviation decrease monotonically with increasing p

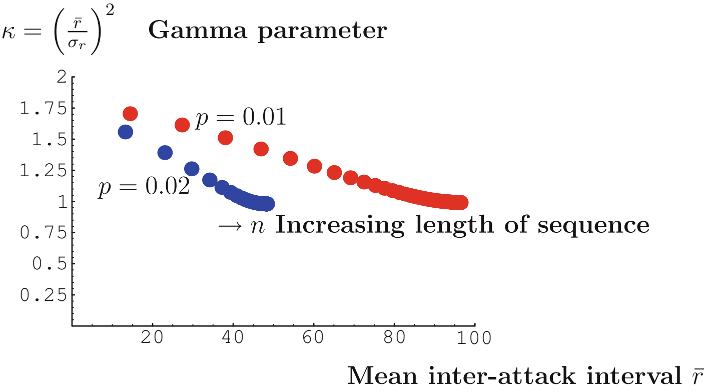


Fig. 9 Effect of sequence length n in random event sequences of length from $n = 50$ to $n = 4000$ in steps of 50. Plot of gamma parameter κ from against mean $\bar{r}$ for inter-attack interval distributions (27). The mean probability for the occurrence of an attack is $p = 0.01$ (*right*, *red*) and $p = 0.02$ (*left*, *blue*), corresponding to the cases in Fig. 7. We expect that, as $n \rightarrow \infty$, so $\kappa \rightarrow 1$, the random case

arbitrarily small variance. Figure 10 illustrates symmetric such cases with mean value $E(a) = \frac{1}{2}$. From the information metric cf. (21) on the space of these probability density functions we can obtain information distances between nearby distributions as follows. Suppose that we record data on amplitude $a \in (0, 1]$ for two cases with parameters (μ_0, κ_0) and $(\mu_0 + \Delta\mu, \kappa_0 + \Delta\kappa)$ for small $\Delta\mu, \Delta\kappa$. Then

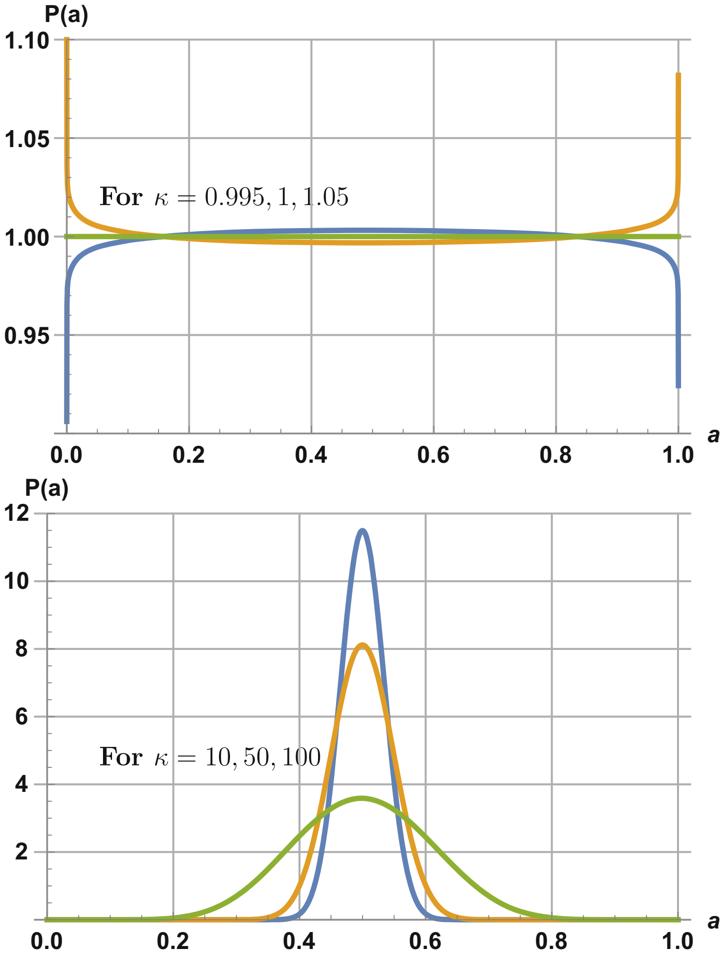


Fig. 10 Examples from the log-gamma family of probability densities with central mean $E(a) = \frac{1}{2}$. *Upper*: near to uniform, $\kappa = 0.995, 1, 1.05$. *Lower*: approximations to truncated Gaussians, $\kappa = 10, 50, 100$

the information distance Δs between these distributions is approximated from

$$\text{In } (\nu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \frac{\kappa_0}{\nu_0^2}\,\Delta\nu^2 - \frac{2}{\nu_0}\,\Delta\nu\Delta\kappa + \frac{d^2 \log \Gamma}{d\kappa^2}(\kappa_0)\;\Delta\kappa^2 \quad (31)$$

$$\text{In } (\mu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \frac{\kappa_0}{\mu_0^2}\,\Delta\mu^2 + \left(\frac{d^2 \log \Gamma}{d\kappa^2}(\kappa_0) - \frac{1}{\kappa_0}\right)\Delta\kappa^2 \quad (32)$$

Table 1 Numerical values of $\frac{d^2 \log \Gamma}{d\kappa^2}(\kappa_0) - \frac{1}{\kappa_0}$ for $\kappa = 1, 2, \ldots, 10$ to illustrate the relative effects of $\Delta\mu$ and $\Delta\kappa$ on Δs in (32)

κ_0	$\frac{d^2 \log \Gamma}{d\kappa^2}(\kappa_0) - \frac{1}{\kappa_0}$
1	0.644934
2	0.144934
3	0.0616007
4	0.033823
5	0.021323
6	0.0146563
7	0.010688
8	0.00813701
9	0.0064009
10	0.00516634

Note that, as κ_0 increases from 1, the factor $\frac{d^2 \log \Gamma}{d\kappa^2}(\kappa_0) - \frac{1}{\kappa_0}$ decreases monotonically from $\frac{\pi^2}{6} - 1$. So, in the information metric, the difference $\Delta\mu$ has increasing prominence over $\Delta\kappa$ as we see in Table 1.

Two particular cases are of interest:

Near to the uniform distribution: Here we have $(\mu_0 = 1, \kappa_0 = 1)$ and Δs^2 reduces to

$$\text{In } (\nu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \Delta\nu^2 - 2\Delta\nu\Delta\kappa + 1.645\Delta\kappa^2 \tag{33}$$

$$\text{In } (\mu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \Delta\mu^2 + 0.645\Delta\kappa^2 \tag{34}$$

Two nearby unimodular distributions: Here we have $\kappa_0 >> 1$ and Δs^2 reduces to

$$\text{In } (\nu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \frac{\kappa_0}{\nu_0}^2 \Delta\nu^2 - \frac{2}{\nu}\Delta\nu\Delta\kappa \tag{35}$$

$$\text{In } (\mu, \kappa) \text{ coordinates} \quad \Delta s^2 \approx \frac{\kappa_0}{\mu_0}^2 \Delta\mu^2. \tag{36}$$

For automated security monitoring of sample distributions, these distance values can be used for creating alerts or action when certain chosen threshold deviations arise from target uniform or chosen truncated Gaussian distributions.

Protecting Devices with Obscuring Techniques

Public key encryption, such as RSA, employs modular arithmetic with a very large modulus. It is necessary to compute

$$R \equiv y^e \ (mod\, m) \ \text{ or } \ R \equiv y^d \ (mod\, m) \tag{37}$$

for, respectively, encrypting or decrypting a message y. The modulus m is chosen to be the product of two large prime numbers p, q, which are kept secret; then choose d, e such that

$$ed \equiv 1 \ (mod\,(p-1)(q-1)). \tag{38}$$

The modulus m and the encryption key e are made public; the decryption key d is secret.

Encoding and decoding computations both involve repeated numerical exponentiation procedures. Kocher et al. [40] showed the effectiveness of Differential Power Analysis (DPA) in breaking encryption procedures using correlations between power consumption and data bit values during processing, claiming that most smart cards reveal their DES keys using fewer than 15 power traces.

Chari et al. [16] provided a probabilistic encoding (secret sharing) scheme for effectively secure computation. They obtained lower bounds on the number of power traces needed to distinguish distributions statistically, under certain assumptions about Gaussian noise functions. DPA attacks depend on the assumption that power consumption in a given clock cycle will have a distribution depending on the initial state; the attacker needs to distinguish between different ‘nearby’ distributions in the presence of noise. Zero-Knowledge proofs allow verification of secret-based actions without revealing the secrets. Goldreich et al. [34] discussed the class of promise problems in which interaction may give additional information in the context of Statistical Zero-Knowledge (SZK). They invoked two types of difference between distributions: the ‘statistical difference’ (SZK) and the ‘entropy difference’ of two random variables. In this context, typically, one of the distributions is the uniform distribution. Thus, in the contexts of DPA and SZK tests, it is necessary to compare two nearby distributions on bounded domains, other situations may need similar comparisons.

Accordingly, some knowledge of the design of an implementation and information on the timing or power consumption during computational stages could yield clues to the decryption key d. Canvel and Dodson [10, 11] showed how timing analyses of the modular exponentiation algorithm quickly reveal the private key, regardless of its length. An obscuring procedure could mask the timing information but that may not be straightforward for some small memory devices. It is important to be able to assess departures from Poisson randomness of underlying or overlaid procedures that are inherent in devices and here we outline some information geometric methods to add to the standard tests [53].

In cryptographic attacks, differential Power Analysis (DPA) methods and Statistical Zero-Knowledge (SZK) proofs depend on discrimination between noisy samples drawn from pairs of closely similar distributions. In many cases the distributions resemble truncated Gaussians; sometimes one distribution is uniform. A log-gamma family of probability density functions provides a 2-dimensional metric space of distributions on $(0, 1]$, ranging from the uniform distribution to symmetric unimodular distributions of arbitrarily small variance. Illustrative calculations are provided here; more discussion is given in [5]. An attack can make use of the time taken for the computations for a chosen set of y values, given some knowledge of the design of the device being used. So, access to timing data of operations on a submission of a sequence of chosen messages $y_i, i = 1, 2, \ldots$ could reveal clues to the succession of bits in the encryption key e. We investigated the square and multiply implementation to perform the exponentiations [11], using Head's algorithm, and we obtained $x^n (\text{mod } m)$ where $n = d_k d_k - 1 \ldots d_1 d_0$ is the exponent in binary form as follows [33]:

Modular exponentiation algorithm for $x^n (\text{mod } m)$

```
While $n > 0$
  Let $d = n - 2[n/2]$ If $d = 1$ then
    $r = xr(\text{mod } m)$ (using Head's algorithm)
  End If
  $x = x^2(\text{mod } m)$ (using Head's algorithm)
  $n = (n - d)/2$
End While
```

Using the R.D. Oliviera library `timer.h`, Canvel [10] programmed in C++ to obtain the timings of the portions of the code for each exponent bit in the square and multiply algorithm. The full code is available in the thesis [10] and results were reported for p and q of lengths from 50 to 290 bits, and e from 20 to 100 bits, both of these ranges in steps of 5 bits. By way of illustration, Fig. 11 shows a typical barchart of timing values, for a 20 bit encryption key e and 195 bit p and q. We can see that the encryption key e can be read from right to left off the barchart of timings.

In practice, clues to such timing information can be obtained from data on power consumption using electromagnetic sensors, possibly needing statistical processes to clean the data of noise. A practicable defence is to obscure the power usage data on timing information by spurious other processes. Then the effectiveness of such obscuring techniques can be evaluated using analyses of the distributions associated with time series from power usage. For example, a time series of power consumption using appropriately chosen thresholding and interval windows would yield a barchart and that would ideally be like that arising from Poisson processes, which for a given mean are maximally disordered [36]. Information geometry can be used to measure differences from the Poisson model, equivalently from its associated exponential distribution. Figure 11 shows the actual timing data of binary digits in the encryption key, illustrating precisely the opposite of maximal disorder: a binary barchart revealing the digit sequence.

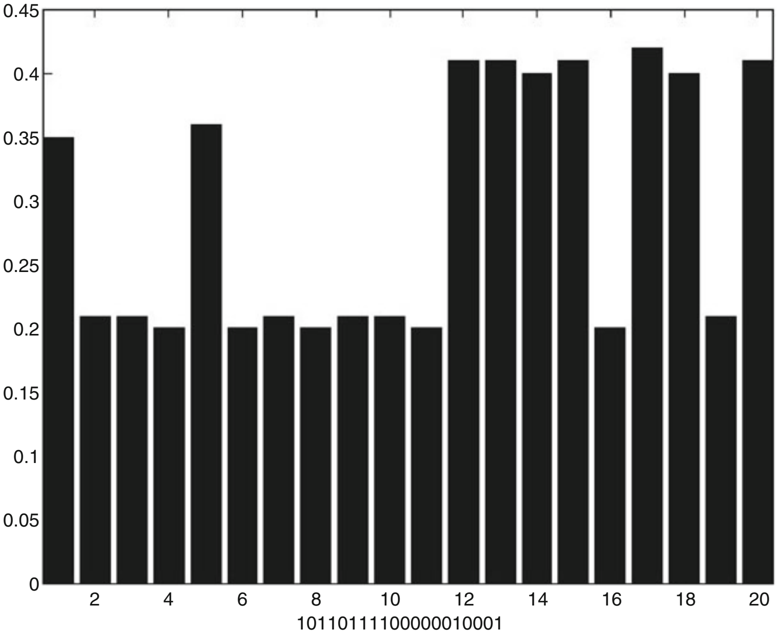


Fig. 11 Example for timings using an encrypting key e of 20 bits with p and q of size 195 bits. The encryption key bits 10110111100000010001 can be read from *right* to *left*

Dimensionality Reduction Methods

A general class of methods used to represent high-dimensional datasets is called multidimensional scaling or dimensionality reduction. In many real world problems we encounter high dimensionality in large data sets and often do not know the optimal net probability density function family for the features represented in the data. A fundamental problem in the identification of probability densities from large multidimensional data sets, that of efficient dimensionality reduction, was addressed by Carter and his co-workers [12–15]. They used information geometry to obtain nearest neighbour distances by means of geodesic estimates subordinate to a Fisher information metric.

This method takes account of the curved geometry of the data set, rather than displaying it as uncurved in a Euclidean geometry cf. [12, Fig. 3.2]. The significance is that the non-obvious global topology of frequency connectivity in the data is revealed by the geodesics. An illustration using router traffic on the Abilene network showed how anomalous behaviour unseen by local methods could be picked up through dimensionality changes cf. [12, Fig. 3.10]. Moreover, in document classification, the information metric approach outperformed standard Principal Component Analysis and Euclidean embeddings [13], and it outperformed traditional approaches to video indexing and retrieval with real world data [17].

Such information geometric methods could extend to anomaly detection using large sample size data sets derived from an underlying probability distribution in which the parameterization is unknown. A comparison of relevant information theoretic measures that are important for anomaly detection can be found in Lee and Xiang [45]. Raginsky et al. [51] provided a *filtering* and *hedging* joint approach to the detection of anomalies in sequentially observed noisy data, by comparing the current belief against a time-varying and data-adaptive threshold. The threshold is adjusted based on the available feedback from an end user. The thesis of Liu [46] addressed the problem of intrusion detection for wireless networks and developed a hybrid anomaly intrusion detection approach, based on two data mining techniques, association-rule mining and cross-feature mining.

The methods described by Carter et al. [12, 13] reduce the dimensionality of data sets and hence identify clustering of sets with similar features through 3-dimensional rendering of the resultant plots. In cyber security we anticipate a large data set $X_1, X_2, .., X_N$ of distributions which represent to differing degrees features relating to potential cyber attacks. Such data may be collected automatically and routinely, with the objective of identifying anomalous behaviour associated with attempted cyber attacks. The analytic procedure consists of a series of computational steps:

1. Compute mutual 'information distances' $D(i,j)$ among the members of the dataset of distributions $X_1, X_2, .., X_N$.
2. The array of $N \times N$ differences $D(i,j)$ is a symmetric positive definite matrix with diagonal zero. Centralize this by subtracting row and column means and then adding back the grand mean to give $CD(i,j)$.
3. The centralized matrix $CD(i,j)$ is again symmetric positive definite with diagonal zero. Its N eigenvalues $ECD(i)$ are necessarily real, and there are N corresponding N-dimensional eigenvectors $VCD(i)$.
4. Let A be the 3×3 diagonal matrix of the first three eigenvalues of largest absolute magnitude and let B be the $3 \times N$ matrix of the corresponding eigenvectors. The matrix product $A \cdot B$ yields a $3 \times N$ matrix and its transpose is an $N \times 3$ matrix T, which gives us N coordinate values (x_i, y_i, z_i) to embed the N samples in $\mathbb{R}^3$.

In the case when no obvious model family of distributions is available, Step 1 above could be effected by means of the widely used Kullback–Leibler measure of relative entropy [41, 42] which in symmetrized form for two probability density functions f_1, f_2 defined on $\Omega \subseteq \mathbb{R}^n$ is

$$D_{KL}(f_1, f_2) = \frac{1}{2}\left(\int_\Omega f_1 \, \log\left(\frac{f_1}{f_2}\right) + \int_\Omega f_2 \, \log\left(\frac{f_2}{f_1}\right)\right) \tag{39}$$

and for numerical, normalized frequency distributions F, G with bin indexing J we could use the discrete form

$$D_{KL}(F,G) = \frac{1}{2}\left(\sum_{j\in J} F_j \log\left(\frac{F_j}{G_j}\right) + \sum_{j\in J} G_j \log\left(\frac{G_j}{F_j}\right)\right). \tag{40}$$

See Johnson and Sinanovic [39] for more symmetrizing options with the Kullback–Leibler relative entropy. We note that for such an empirical frequency distribution F the entropy is

$$S(F) = \sum_{j\in J} F_j. \tag{41}$$

The kind of data sets collected by security monitoring software is usually customized to the context and network connectivity to be protected. Gu et al. [35] used maximum entropy (41) and relative entropy (i.e. Kullback–Leibler (40)) to compare a pre-trained baseline distribution for internet traffic with the current traffic being monitored; their results indicated some success in anomaly detection, including synchronizing (SYN) attacks and port scanning reconnaissance attacks. Lee and Xiang [45] discuss intrusion detection systems, such as SunSHIELD [57] and *tcdump* [38], which collect system and network activity data and analyse it to determine whether an attack is in progress.

Carter's thesis [12] included an illustration using router traffic on the Abilene network that showed how anomalous behaviour unseen by local methods could be picked up through dimensionality changes cf. [12, Fig. 3.10]. The data was from the Abilene Network of 11 routers which comprise the core of the 'edu' network; the number of packets on each of these routers was taken every 5 min through 1–2 January 2005, which yielded an 11-dimensional dataset of 576 samples.

Some illustrations of applications of dimensionality reduction to stochastic textures in one and two dimensional processes can be found in [27]. An example of a 1-dimensional stochastic texture is a grey-level barcode for a genome, obtained by mapping the 20 amino acids onto grey levels. The neighbourhood structure of grey-level values yielded autocovariances and hence multivariate Gaussian distributions; then the dimensionality reduction of a large dataset of these could be represented on a plot in $\mathbb{R}^3$ which illustrated the main features discriminating among yeast, human, and randomly generated genomes, cf. [27, Figs. 16 and 17]. Two-dimensional stochastic textures can be represented as surface topographies and these were studied in [29].

Autocovariance extractions can be made from administratively monitored time series for a local network of nominally secure access points to a server providing sensitive data. Then for the case of k variables arising from value x_1 at a point t_1 and the value of the succession of averages of its successive neighbours at distances giving $\bar{x}_2$ at $t_{\pm 2}$ and so on to $\bar{x}_k$ at $t_{\pm k}$ in the series, we can compute analytically the information distance $D(\Sigma^A, \Sigma^B)$ between pairs of $k\times k$ covariance matrices Σ^A, Σ^B from section "Information Geometry of Gaussians"

$$D(\Sigma^A, \Sigma^B) = \sqrt{\frac{1}{2}\sum_{j=1}^{k} \log^2(\lambda_j^{AB})} \tag{42}$$

$$\text{where } \{\lambda_j^{AB}\} = \mathrm{Eig}(S^{AB}) \text{ and } S^{AB} = \Sigma^{A^{-1/2}} \cdot \Sigma^B \cdot \Sigma^{A^{-1/2}}. \tag{43}$$

Multivariate monitoring data in n-dimensional pixel form can be similarly treated through its autocovariance matrices, the only difference being that the neighbours for a given pixel $(x_i) \in \mathbb{R}^n$ form annular regions of radius $1, 2, \ldots, k$ around the pixel [29].

In the case that the data is more complex, for example mixtures of multivariate Gaussians, analytic information distances are not available but approximations have been obtained, for example in [23] where also the effects of different weighting sequences were investigated. In many applications, a true information metric is not essential since relative discriminations that are monotonically related to it may be adequate for purposes of comparison between datasets.

Discussion

The framework of Riemannian geometry has been enormously valuable in the development over the past century and a half of physical models for real processes in time and space, as well as in the abstract high-dimensional and infinite-dimensional spaces that are important for physical field theories. Conversely, these developments in physical models have stimulated developments in geometry, global analysis and functional analysis, to extend mathematical structures to contexts that were not previously envisaged. The more recent development of cheap computer power and rapid processing has made possible very effective developments in computational geometry for numerical solution of complex problems.

Analytic computing packages like *Mathematica* [63] enable rapid handling of complex operations in large numbers of variables to obtain analytic solutions to mathematical problems that were previously inaccessible. Moreover, this analytic work is easily performed by anyone with experience in writing mathematical expressions and it is easy to display results and effects of variables graphically, including with animation. The development during the last 70 years of information geometry has provided all the tools of Riemannian geometry, for use on smoothly parametrized families of probability density functions, and modern computational information geometry makes possible the analysis of complex statistical models. Moreover, it provides their representation through natural embeddings such as Figs. 2, 4, 6, which are more easily interpreted than tables of data.

Encoding and decoding algorithms for public key encryption typically involve repeated numerical exponentiation procedures. Correlations, between power consumption and data bit values, during processing by a device when supplied with chosen inputs can reveal clues to the encryption key, if a sensor can be placed to pick up the power consumption traces, section “Protecting Devices with Obscuring Techniques”. This stems from the fact that the processing algorithms use different times to run for steps involving a ‘1’ bit from a ‘0’ bit in the key, as we illustrated using modular exponentiation with Head’s algorithm in Fig. 11 above, from our report of timing studies in [11]. One practicable defence is to obscure the power usage data on timing information by spurious other processes. Then the effectiveness of such obscuring techniques can be evaluated using analyses of the distributions associated with time series from power usage. For example, a time series of power consumption using appropriately chosen thresholds and interval windows would yield a barchart that would ideally be like that arising from Poisson random processes, hiding any influence from internal process timings. In evaluating such obscuring techniques, information geometry can measure differences from the Poisson ideal.

The monitoring of networks to identify breaches of cyber security may often involve the automated collection of very large volumes of possibly high dimensional time series data, which may be representable through empirical frequency distributions tailored to the process to be protected. In some cases these distributions may be interpreted as perturbations of, for example, Poisson random processes or uniform processes, which could be an ideal target for the data in the absence of attacks or anomalous events. We have seen in section “Information Geometry of the Gamma Manifold” that the family of gamma distributions and its logarithmic version have relevance to this context. These have simple information geometry, which we have illustrated in some example calculations, section “Testing Nearby Signal Distributions and Drifts from Uniformity”, concerning the measurement of departures from uniformity and from a Poisson random process. There also we showed how to provide a distance between two nearby truncated Gaussian-like distributions, Fig. 10. For automated security monitoring of sample distributions, these distance values can be used for creating alerts or initiating action when certain chosen threshold deviations arise from target uniform or chosen truncated Gaussian distributions. In the absence of a model family of distributions we can use the symmetrized Kullback–Leibler relative entropy expression, Eq. (40), to measure distance between empirical frequency distributions.

In other cases, for n-dimensional data values $(x_i) \in \mathbb{R}^n$ for any $n = 1, 2, \ldots,$ we can compute $(k \times k)$ autocovariance matrices from values at a point in $\mathbb{R}^n$ and the averages of its jth-neighbours for a set of $j = 1, 2, \ldots, k$. These are amenable to an interpretation through the information geometry of k-variate Gaussian covariances, which also have relatively simple information geometry as we showed in section “Information Geometry of Gaussians” where we have provided explicit analytic expressions, Eq. (42) for distances between two $(k \times k)$ covariance matrices.

For the situations when large numbers of empirical or fitted distributions have to be handled, we have described explicitly in section “Dimensionality Reduction Methods” the steps to follow in the method of dimensionality reduction. The methodology represents all of the datasets on one plot in $\mathbb{R}^3$, which reveals the topology of the dataset and any prominent features, such as clusters and natural subgroups within the data, regardless of the number of datasets. This procedure has been shown to outperform the more common Principal Component Analysis [13] in feature identification.

References

1. S.-I. Amari, Theory of information spaces—a geometrical foundation of the analysis of communication systems. Res. Assoc. Appl. Geom. Memoirs **4**, 171–216 (1968)
2. S.-I. Amari, *Differential Geometrical Methods in Statistics*. Springer Lecture Notes in Statistics, vol. 28 (Springer, Berlin, 1985)
3. S.-I. Amari, H. Nagaoka, *Methods of Information Geometry.* American Mathematical Society (Oxford University Press, Oxford, 2000)
4. S.-I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen, C.R. Rao, *Differential Geometry in Statistical Inference*. Lecture Notes Monograph Series, vol. 10 (Institute of Mathematical Statistics, Hayward California, 1987)
5. K. Arwini, C.T.J. Dodson, *Information Geometry Near Randomness and Near Independence.* Lecture Notes in Mathematics (Springer, Berlin, 2008)
6. C. Atkinson, A.F.S. Mitchell, Rao's distance measure. Sankhya Indian J. Stat. A **48**(3), 345–365 (1981)
7. D. Burstein, F. Kenter, J. Kun, F. Shi, Information monitoring in routing networks (2015), 12 pp. http://arxiv.org/pdf/1507.05206.pdf
8. E. Byrse, D. Leversage, The industrial security incident database (2006). http://www.securitymetrics.org/attachments/Metricon-1-Leversage-Rump.pdf
9. E. Byrse, J. Lowe, The myths and facts behind cyber security risks for industrial control systems. VDE 2004 Congress, VDE, Berlin, Oct 2004
10. B. Canvel, Timing tags for exponentiations for RSA. MSc Thesis, Department of Mathematics, University of Manchester Institute of Science and Technology, Manchester (1999)
11. B. Canvel, C.T.J. Dodson, Public key cryptosystem timing analysis, in *CRYPTO 2000*, Rump Session Santa Barbara, 20–24 Aug 2000. http://www.maths.manchester.ac.uk/~kd/PREPRINTS/rsatim.pdf, 27 Aug 2000
12. K.M. Carter, Dimensionality reduction on statistical manifolds. PhD thesis, University of Michigan (2009). http://tbayes.eecs.umich.edu/kmcarter/thesis
13. K.M. Carter, R. Raich, A.O. Hero III, FINE: information embedding for document classification, in *Proceedings of 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Mar 2008. http://tbayes.eecs.umich.edu/kmcarter/fine_doc
14. K.M. Carter, R. Raich, W.G. Finn, A.O. Hero III, Fisher information nonparametric embedding. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 2093–2098 (2009). http://arxiv.org/abs/0802.2050v1
15. K.M. Carter, R. Raich,W.G. Finn, A.O. Hero III, Information-geometric dimensionality reduction. IEEE Signal Process. Mag. **99**, 89–99 (2011). http://web.eecs.umich.edu/~hero/Preprints/carter_spsmag_igdr_rev3.pdf
16. S. Chari, C.S. Jutla, J.R. Rao, P. Rohatgi, Towards sound approaches to counteract power-analysis attacks, in *Advances in Cryptology-CRYPTO '99*, ed. by M. Wiener. Lecture Notes in Computer Science, vol. 1666 (Springer, Berlin, 1999), pp. 398–412
17. X. Chen, A. Hero, Fisher information embedding for video indexing and retrieval, in *SPIE Electronic Imaging Conference*, San Jose (2011). nts/ChenEI11.web.eecs.umich.edu/~hero/Prepripdf
18. J. Chen, G. Venkataramani, An algorithm for detecting contention-based covert timing channels on shared hardware, in *Proceedings of HASP '14 Third Workshop on Hardware and Architectural Support for Security and Privacy*. ACM Digital Library dl.acm.org http://www.seas.gwu.edu/~guruv/hasp14.pdf

19. CPNI: UK Centre for the Protection of National Infrastructure (2015), http://www.cpni.gov.uk/advice/cyber/
20. P. Crzegorzewski, R. Wieczorkowski, Entropy-based goodness-of-fit test for exponentiality. Commun. Stat. Theory Methods **28**(5), 1183–1202 (1999)
21. N.J. Daras, Stochastic analysis of cyber attacks, in *Applications of Mathematics and Informatics in Science and Engineering*, ed. by N.J. Daras. Springer Optimization and its Applications, vol. 91 (Springer, Berlin, 2014), pp. 105–129
22. C.T.J. Dodson (ed.), *Proceedings of Workshop on Geometrization of Statistical Theory*, Lancaster 28–31 Oct 1987 (ULDM Publications, University of Lancaster, Lancaster, 1987)
23. C.T.J. Dodson, Information distance estimation between mixtures of multivariate Gaussians, in *Presentation at Workshop on Computational Information Geometry for Image and Signal Processing*, ICMS Edinburgh, 21–25 Sept 2015
24. C.T.J. Dodson, Some illustrations of information geometry in biology and physics, in *Handbook of Research on Computational Science and Engineering: Theory and Practice*, ed. by J. Leng, W. Sharrock, IGI-Global, Hershey, PA, 2012, pp. 287–315. http://www.igi-global.com/book/handbook-research-computational-science-engineering/51940
25. C.T.J. Dodson, H. Matsuzoe, An affine embedding of the gamma manifold. InterStat **2002**(2), 1–6 (2002)
26. C.T.J. Dodson, T. Poston, *Tensor Geometry*, 2nd edn. Graduate Texts in Mathematics, vol. 130 (Springer, New York, 1991)
27. C.T.J. Dodson, W.W. Sampson, Dimensionality reduction for classification of stochastic texture images, in *Geometric Theory of Information*, ed. by F. Nielsen (Springer, Heidelberg, 2014), pp. 1013–1015
28. C.T.J. Dodson, S.M. Thompson. A metric space of test distributions for DPA and SZK proofs. Poster Session, *Eurocrypt 2000*, Bruges, 14–19 May 2000. http://www.maths.manchester.ac.uk/~kd/PREPRINTS/mstd.pdf
29. C.T.J. Dodson, M. Mettänen, W.W. Sampson, Dimensionality reduction for characterization of surface topographies, in *Presentation at Workshop on Computational Information Geometry for Image and Signal Processing*, ICMS Edinburgh, 21–25 Sept 2015
30. P.S. Eriksen, Geodesics connected with the Fisher metric on the multivariate normal manifold, in ed. by C.T.J. Dodson *Proceedings of the GST Workshop*, Lancaster 1987, pp. 225–229. http://trove.nla.gov.au/version/21925860
31. W. Feller, *An Introduction to Probability Theory and its Applications*, 2nd edn. vol. II (Wiley, New York, 1971)
32. R.A. Fisher, Theory of statistical estimation. Proc. Camb. Philos. Soc. **122**, 700–725 (1925)
33. P. Ginlin, *Primes and Programming: An Introduction to Number Theory with Computing* (Cambridge University Press, Cambridge, 1993)
34. O. Goldreich, A. Sahai, S. Vadham, Can statistical zero-knowledge be made non-interactive? Or, on the relationship of SZK and NISZK, in *Advances in Cryptology-CRYPTO '99*, ed. by M. Wiener. Lecture Notes in Computer Science, vol. 1666 (Springer, Berlin, 1999), pp. 467–484
35. Y. Gu, A. McCallum, D. Towsley, Detecting anomalies in network traffic using maximum entropy estimation, in *Proceedings of Internet Measurement Conference 2005*, pp 345–350. More details are in the Technical Report from the Department of Computer Science, UMASS, Amherst 2005
36. F.A. Haight, *Handbook of the Poisson Distribution* (Wiley, New York, 1967)
37. T.-Y. Hwang, C.-Y. Hu, On a characterization of the gamma distribution: the independence of the sample mean and the sample coefficient of variation. Ann. Inst. Stat. Math. **51**(4), 749–753 (1999)
38. V. Jacobson, C. Leres, S. McCanne: *tcdump* via anonymous ftp.ee.lbl.gov, June 1989
39. D.H. Johnson, S. Sinanovic, Symmetrizing the Kullback–Leibler Distance. *Rice University doc* (2001), https://scholarship.rice.edu/handle/1911/19969
40. P. Kocher, J. Jaffe, B. Jun, Differential power analysis, in *Advances in Cryptology-CRYPTO '99*, ed. by M. Wiener. Lecture Notes in Computer Science, vol. 1666 (Springer, Berlin, 1999), pp. 388–397
41. S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959)

42. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
43. R.G. Laha, On a characterization of the gamma distribution. Ann. Math. Stat. **25**, 784–787 (1954)
44. A. Liaropoulos, G. Tsihrintzis (eds.), *Proceedings of* 13*th European Conference on Cyber Warfare and Security*, University of Piraeus, 3–4 July 2014
45. W. Lee, D. Xiang, Information-theoretic measures for anomaly detection, in *Proceedings of IEEE Symposium Security and Privacy* (2001), pp. 130–143. doi:10.1109/SECPRI.2001.924294
46. Y. Liu, Intrusion detection for wireless networks. PhD thesis, Stevens Institute of Technology. ACM Digital Library dl.acm.org (2006)
47. L.A. Maglaras, J. Jiang, T.J. Cruz, Integrated OCSVM mechanism for intrusion detection in SCASA systems. Electron. Lett. **50**(1), 1935–1936 (2014). Cf also: combining ensemble methods and social network metrics for improving accuracy of OCSVM on intrusion detection in SCADA systems. http://arxiv.org/pdf/1507.02825.pdf, 25 pp.
48. F. Nielsen, F. Barbaresco (eds.), *Geometric Science of Information*. GSI2013, vol. 8085. Lecture Notes in Computer Science (Springer, Heidelberg, 2013)
49. F. Nielsen (ed.), *Geometric Theory of Information* (Springer, Heidelberg, 2014)
50. R.E. Pino (ed.), *Network Science and Cybersecurity* (Springer, New York, 2014)
51. M. Raginsky, R. Willett, C. Horn, J. Silva, R. Marcia, Sequential anomaly detection in the presence of noise and limited feedback. ArXiv arXiv:0911.2904v4 (2012) 1–19
52. C.R. Rao, Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
53. A. Rushkin, J. Soto et al., *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. National Institute of Standards & Technology*, Gaithersburg, MD (2001)
54. RISI online incident database (2015), http://www.risidata.com/Database
55. B.Y. Ryabko, V.A. Monarev, Using information theory approach to randomness testing. J. Stat. Plan. Inf. **133**(1), 95–110 (2005)
56. S.D. Silvey, *Statistical Inference* (Chapman and Hall, Cambridge, 1975)
57. SunSoft *SunSHIELD Basic Security Module Guide* (Soft, Mountain View, CA, 1995). https://docs.oracle.com/cd/E19457-01/801-6636/801-6636.pdf
58. The Information Assurance Advisory Council (IAAC) (2017), http://www.iaac.org.uk/
59. P. Trim, H.Y. Youm (eds.), *Korea-UK Initiatives in Cyber Security Research: Government, University and Industry Collaboration*, Report Submitted to the Korean Government and the UK Government Mar (2015) http://www.iaac.org.uk/media/1356/cyber-security-report-trim-and-youm-march2015.pdf
60. USA homeland security: cybersecurity (2017), https://www.dhs.gov/topic/cybersecurity
61. C.R. Vance Jr., Smartphone encryption and public safety, in 6^{th} *Annual Financial Crimes and Cybersecurity Symposium*. Federal Reserve Bank of New York 15 Nov (2015), 42 pp. http://manhattanda.org/sites/default/files/11.18.15
62. W. Wang, Z. Lu, Cyber security in the smart grid: survey and challenges. Comput. Netw. **57**(5), 1344–1371 (2013). http://www.sciencedirect.com/science/article/pii/S1389128613000042
63. S. Wolfram, *The Mathematica Book*, 3rd edn. (Cambridge University Press, Cambridge, 1996)

A Survey of Recent Inequalities for Relative Operator Entropy

Silvestru Sever Dragomir

Abstract The concepts of relative operator entropy and operator entropy play an important role in different subjects, such as statistical mechanics, information theory, dynamical systems and ergodic theory, biology, economics, human and social sciences. They are closely related to the problem of the quantification of entanglement, the distinguishability of quantum states and to thermodynamical ideas. In this paper we survey some recent inequalities obtained by the author for the relative operator entropy $S(\cdot|\cdot)$, for positive invertible operators A and B in general, and, in particular when they satisfy the boundedness condition $mA \leq B \leq MA$ for some m, M with $0 < m < M$. Natural applications for the operator entropy $\eta(\cdot)$ are provided. In the end, some trace inequalities for trace class operators A and B that satisfy the normality condition $\operatorname{tr}(A) = \operatorname{tr}(B) = 1$ are also given.

Keywords Relative operator entropy • Operator entropy • Young's inequality • Convex functions • Operator inequalities • Means

Introduction

The concept of entropy was introduced in thermodynamics by Clausius in 1865, and some of the main steps towards the consolidation of the concept were taken by Boltzmann and Gibbs. Many generalizations and reformulations of this notion have been proposed, with motivations and applications in different subjects, such as statistical mechanics, information theory, dynamical systems and ergodic theory, biology, economics, human and social sciences. In quantum mechanics, pure states of physical systems are described by vectors in a Hilbert space, while mixed states are described by positive semi-definite matrices with trace one. Such matrices are called density matrices. In these contexts one mathematical function emerges as a

S.S. Dragomir (✉)
Mathematics, College of Engineering & Science, Victoria University, Melbourne, Australia

DST-NRF Centre of Excellence in the Mathematical and Statistical Sciences, School of Computer Science and Applied Mathematics, University of Witwatersrand, Johannesburg, South Africa
e-mail: sever.dragomir@vu.edu.au

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_10

central quantity, namely the relative operator entropy. It has a number of remarkable properties [6, 7, 12, 19] and is closely related to the problem of the quantification of entanglement, the distinguishability of quantum states and to thermodynamical ideas. Any new inequality relating the relative entropy to other entropic quantities is therefore expected to lead to potentially important new insights into any of these topics and is potentially an important contribution.

To be more precise, recall that Kamei and Fujii [6, 7] defined the *relative operator entropy* $S(A|B)$, for positive invertible operators A and B, by

$$S(A|B) := A^{\frac{1}{2}}\left(\ln A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\right)A^{\frac{1}{2}} \tag{1}$$

which is a relative version of the operator entropy considered by Nakamura–Umegaki [21].

In general, we can define for positive operators $A,\ B$

$$S(A|B) := s - \lim_{\varepsilon\to 0+} S(A+\varepsilon I|B)$$

if it exists, here I is the identity operator.

For the entropy function $\eta(t) = -t\ln t$, the operator entropy has the following expression:

$$\eta(A) = -A\ln A = S(A|I) \geq 0$$

for positive contraction A. This shows that the relative operator entropy (1) is a relative version of the operator entropy.

Following [12, pp. 149–155] we recall some important properties of relative operator entropy for A and B positive invertible operators:

(i) We have the equalities

$$\begin{aligned} S(A|B) &= -A^{1/2}\left(\ln A^{1/2}B^{-1}A^{1/2}\right)A^{1/2} \\ &= B^{1/2}\eta\left(B^{-1/2}AB^{-1/2}\right)B^{1/2}; \end{aligned} \tag{2}$$

(ii) We have the inequalities

$$S(A|B) \leq A(\ln\|B\| - \ln A) \text{ and } S(A|B) \leq B - A;$$

(iii) For any C, D positive invertible operators we have that

$$S(A+B|C+D) \geq S(A|C) + S(B|D);$$

(iv) If $B \leq C$, then

$$S(A|B) \leq S(A|C);$$

(v) If $B_n \downarrow B$, then

$$S(A|B_n) \downarrow S(A|B);$$

(vi) For $\alpha > 0$ we have

$$S(\alpha A|\alpha B) = \alpha S(A|B);$$

(vii) For every operator T we have

$$T^* S(A|B) T \le S\left(T^* AT|T^* BT\right).$$

The relative operator entropy is jointly concave, namely for any positive invertible operators $A, B, \ C, \ D$ we have

$$S(tA + (1-t)B|tC + (1-t)D) \ge tS(A|C) + (1-t)S(B|D)$$

for any $t \in [0, 1]$.

For recent results related on operator entropy see also [8, 10, 13, 16, 17, 20, 22].

The famous *Young inequality* for scalars says that if $a, b > 0$ and $\nu \in [0, 1]$, then

$$a^{1-\nu} b^{\nu} \le (1-\nu) a + \nu b \tag{3}$$

with equality if and only if $a = b$. The inequality (3) is also called *ν-weighted arithmetic-geometric mean inequality*.

We recall that *Specht's ratio* is defined by [23]

$$S(h) := \begin{cases} \dfrac{h^{\frac{1}{h-1}}}{e \ln\left(h^{\frac{1}{h-1}}\right)} \text{ if } h \in (0,1) \cup (1,\infty) \\ \\ 1 \text{ if } h = 1. \end{cases} \tag{4}$$

It is well known that $\lim_{h \to 1} S(h) = 1, \ S(h) = S\left(\frac{1}{h}\right) > 1$ for $h > 0, \ h \ne 1$. The function is decreasing on $(0, 1)$ and increasing on $(1, \infty)$.

The following inequality provides a refinement and a multiplicative reverse for Young's inequality

$$S\left(\left(\frac{a}{b}\right)^r\right) a^{1-\nu} b^{\nu} \le (1-\nu) a + \nu b \le S\left(\frac{a}{b}\right) a^{1-\nu} b^{\nu}, \tag{5}$$

where $a, b > 0, \ \nu \in [0, 1], \ r = \min\{1-\nu, \nu\}$.

The second inequality in (5) is due to Tominaga [24] while the first one is due to Furuichi [9].

We consider the *Kantorovich's constant* defined by

$$K(h) := \frac{(h+1)^2}{4h}, h > 0. \tag{6}$$

The function K is decreasing on $(0,1)$ and increasing on $[1,\infty)$, $K(h) \geq 1$ for any $h > 0$ and $K(h) = K\left(\frac{1}{h}\right)$ for any $h > 0$.

The following multiplicative refinement and reverse of Young inequality in terms of Kantorovich's constant holds

$$K^r\left(\frac{a}{b}\right) a^{1-\nu} b^\nu \leq (1-\nu) a + \nu b \leq K^R\left(\frac{a}{b}\right) a^{1-\nu} b^\nu \tag{7}$$

where $a, b > 0$, $\nu \in [0,1]$, $r = \min\{1-\nu, \nu\}$ and $R = \max\{1-\nu, \nu\}$.

The first inequality in (7) was obtained by Zou et al. in [26] while the second by Liao et al. [18].

Kittaneh and Manasrah [14, 15] provided a refinement and an additive reverse for Young inequality as follows:

$$r\left(\sqrt{a} - \sqrt{b}\right)^2 \leq (1-\nu) a + \nu b - a^{1-\nu} b^\nu \leq R\left(\sqrt{a} - \sqrt{b}\right)^2 \tag{8}$$

where $a, b > 0$, $\nu \in [0,1]$, $r = \min\{1-\nu, \nu\}$ and $R = \max\{1-\nu, \nu\}$. The case $\nu = \frac{1}{2}$ reduces (8) to an identity.

In the recent paper [2] we obtained the following reverses of Young's inequality as well:

$$0 \leq (1-\nu) a + \nu b - a^{1-\nu} b^\nu \leq \nu(1-\nu)(a-b)(\ln a - \ln b) \tag{9}$$

and

$$1 \leq \frac{(1-\nu) a + \nu b}{a^{1-\nu} b^\nu} \leq \exp\left[4\nu(1-\nu)\left(K\left(\frac{a}{b}\right) - 1\right)\right], \tag{10}$$

where $a,\ b > 0$, $\nu \in [0,1]$.

In [3] we obtained the following inequalities that improve the corresponding results of Furuichi and Minculete from [11]

$$\begin{aligned} &\frac{1}{2}\nu(1-\nu)(\ln a - \ln b)^2 \min\{a,b\} \qquad (11)\\ &\leq (1-\nu) a + \nu b - a^{1-\nu} b^\nu \\ &\leq \frac{1}{2}\nu(1-\nu)(\ln a - \ln b)^2 \max\{a,b\} \end{aligned}$$

and

$$\begin{aligned}&\exp\left[\frac{1}{2}\nu(1-\nu)\left(1-\frac{\min\{a,b\}}{\max\{a,b\}}\right)^2\right]\\ &\leq \frac{(1-\nu)a+\nu b}{a^{1-\nu}b^{\nu}}\\ &\leq \exp\left[\frac{1}{2}\nu(1-\nu)\left(\frac{\max\{a,b\}}{\min\{a,b\}}-1\right)^2\right]\end{aligned} \tag{12}$$

for any $a,\ b>0$ and $\nu\in[0,1]$.

In this paper we survey some recent inequalities obtained by the author for the relative operator entropy $S(\cdot|\cdot)$, for positive invertible operators A and B in general, and, in particular when they satisfy the boundedness condition

$$mA\leq B\leq MA \tag{13}$$

for some $m,\ M$ with $0<m<M$. Natural applications for the operator entropy $\eta(\cdot)$ are provided. In the end, some trace inequalities for trace class operators A and B that satisfy the normality condition $\operatorname{tr}(A)=\operatorname{tr}(B)=1$ are also given.

Inequalities via Uhlmann's Representation

In [25], A. Uhlmann has shown that the relative operator entropy $S(A|B)$ can be represented as the strong limit

$$S(A|B)=s-\lim_{t\to 0}\frac{A\sharp_t B-A}{t} \tag{14}$$

where

$$A\sharp_\nu B:=A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^\nu A^{1/2},\ \nu\in[0,1]$$

is the *weighted geometric mean* of positive invertible operators A and B. For $\nu=\frac{1}{2}$ we denote $A\sharp B$.

We have:

Theorem 1 (Dragomir, 2015 [4]) *Let $A,\ B$ be two positive invertible operators, then we have*

$$S(A|B)\leq 2(A\sharp B-A)\leq B-A. \tag{15}$$

Proof From the inequality (8) for $\nu \in \left(0, \frac{1}{2}\right)$ and $a,\ b > 0$ we have

$$\nu \left(\sqrt{a} - \sqrt{b}\right)^2 \leq (1-\nu)\, a + \nu b - a^{1-\nu} b^{\nu}$$

that is equivalent to

$$a - 2\sqrt{ab} + b \leq b - a + \frac{1}{\nu}\left(a - a^{1-\nu} b^{\nu}\right)$$

and to

$$\frac{1}{\nu}\left(a^{1-\nu} b^{\nu} - a\right) \leq 2\left(\sqrt{ab} - a\right). \tag{16}$$

If we take in (16) $a = 1$, then we get

$$\frac{1}{\nu}\left(b^{\nu} - 1\right) \leq 2\left(b^{1/2} - 1\right), \tag{17}$$

for any $\nu \in \left(0, \frac{1}{2}\right)$ and $a,\ b > 0$.

If we use the continuous functional calculus, then we have for any positive operator X that

$$\frac{1}{\nu}\left(X^{\nu} - 1\right) \leq 2\left(X^{1/2} - 1\right), \tag{18}$$

for any $\nu \in \left(0, \frac{1}{2}\right)$.

If we take in (18) $X = A^{-1/2}BA^{-1/2}$, then we get

$$\frac{1}{\nu}\left(\left(A^{-1/2}BA^{-1/2}\right)^{\nu} - 1\right) \leq 2\left(\left(A^{-1/2}BA^{-1/2}\right)^{1/2} - 1\right), \tag{19}$$

for any $\nu \in \left(0, \frac{1}{2}\right)$.

Multiplying both sides of (19) by $A^{1/2}$ we get

$$\begin{aligned} &\frac{1}{\nu}\left(A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{\nu} A^{1/2} - A\right) \\ &\leq 2\left(A^{1/2}\left(A^{-1/2}BA^{-1/2}\right)^{1/2} A^{1/2} - A\right), \end{aligned} \tag{20}$$

for any $\nu \in \left(0, \frac{1}{2}\right)$.

By taking the strong limit over $\nu \to 0+$ in (20) and by using the representation (14) we obtain the first inequality in (15).

By the operator geometric mean-arithmetic mean inequality $A \sharp B \leq \frac{1}{2}(A + B)$ we deduce the second part of (15).

Remark 1 The inequality (15) is an improvement of the result from (ii) in the introduction.

Corollary 1 *For any positive invertible operator C we have*

$$\eta(C) \le 2\left(C^{1/2} - C\right) \le I - C.$$

Theorem 2 (Dragomir, 2015 [4]) *Let A, B be two positive invertible operators, then we have*

$$\begin{aligned}&(0 \le)\, \frac{1}{2} A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2 \left(\frac{1}{2}I - \left|A^{-1/2}BA^{-1/2} - \frac{1}{2}I\right|\right) A^{1/2} \\ &\le B - A - S(A|B) \\ &\le \frac{1}{2} A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2 \left(\frac{1}{2}I + \left|A^{-1/2}BA^{-1/2} - \frac{1}{2}I\right|\right) A^{1/2}.\end{aligned} \tag{21}$$

Proof From (9) we have

$$\begin{aligned}&\frac{1}{2}\nu(1-\nu)(\ln a - \ln b)^2 \min\{a, b\} \\ &\le (1-\nu)a + \nu b - a^{1-\nu}b^{\nu} \\ &\le \frac{1}{2}\nu(1-\nu)(\ln a - \ln b)^2 \max\{a, b\}\end{aligned}$$

for any for $\nu \in (0, 1)$ and $a,\ b > 0$.

This is equivalent to

$$\begin{aligned}&\frac{1}{2}(1-\nu)(\ln a - \ln b)^2 \min\{a, b\} \\ &\le b - a + \frac{1}{\nu}\left(a - a^{1-\nu}b^{\nu}\right) \\ &\le \frac{1}{2}(1-\nu)(\ln a - \ln b)^2 \max\{a, b\}\end{aligned} \tag{22}$$

for any for $\nu \in (0, 1)$ and $a,\ b > 0$.

If we replace in (22) $a = 1$ and $b = x$, then we get

$$\begin{aligned}&(0 \le)\, \frac{1}{2}(1-\nu)(\ln x)^2 \min\{1, x\} \\ &\le x - 1 + \frac{1}{\nu}(1 - x^{\nu}) \\ &\le \frac{1}{2}(1-\nu)(\ln x)^2 \max\{1, x\}\end{aligned}$$

for any for $\nu \in (0, 1)$ and $x > 0$.

If we use the continuous functional calculus, then we have for any positive operator X that

$$\begin{aligned}&\frac{1}{2}(1-\nu)(\ln X)^2\left(\frac{1}{2}I-\left|X-\frac{1}{2}I\right|\right)\\&\leq X-1+\frac{1}{\nu}(1-X^\nu)\\&\leq\frac{1}{2}(1-\nu)(\ln X)^2\left(\frac{1}{2}I+\left|X-\frac{1}{2}I\right|\right)\end{aligned}\tag{23}$$

for any for $\nu\in(0,1)$.

If we take in (23) $X=A^{-1/2}BA^{-1/2}$, then we get

$$\begin{aligned}&\frac{1}{2}(1-\nu)\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I-\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)\\&\leq A^{-1/2}BA^{-1/2}-1+\frac{1}{\nu}\left(1-\left(A^{-1/2}BA^{-1/2}\right)^\nu\right)\\&\leq\frac{1}{2}(1-\nu)\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I+\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)\end{aligned}\tag{24}$$

for any for $\nu\in(0,1)$.

Multiplying both sides of (24) by $A^{1/2}$ we get

$$\begin{aligned}&\frac{1}{2}(1-\nu)A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I-\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)A^{1/2}\\&\leq B-A+\frac{1}{\nu}A^{1/2}\left(I-\left(A^{-1/2}BA^{-1/2}\right)^\nu\right)A^{1/2}\\&\leq\frac{1}{2}(1-\nu)A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I+\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)A^{1/2}\end{aligned}\tag{25}$$

for any for $\nu\in(0,1)$.

This is an inequality of interest in itself.

Now, if we let $\nu\to0+$ in (25), then we get

$$\begin{aligned}&(0\leq)\,\frac{1}{2}A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I-\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)A^{1/2}\\&\leq B-A-S(A|B)\\&\leq\frac{1}{2}A^{1/2}\left(\ln A^{-1/2}BA^{-1/2}\right)^2\left(\frac{1}{2}I+\left|A^{-1/2}BA^{-1/2}-\frac{1}{2}I\right|\right)A^{1/2},\end{aligned}$$

which proves the desired result (21).

Corollary 2 *For any positive invertible operator C we have*

$$(0 \leq) \frac{1}{2} (\ln C)^2 \left(\frac{1}{2} I - \left| C - \frac{1}{2} I \right| \right) \tag{26}$$
$$\leq I - C - \eta (C) \leq \frac{1}{2} (\ln C)^2 \left(\frac{1}{2} I + \left| C - \frac{1}{2} I \right| \right).$$

Trapezoid Error Estimates

As shown below, by making use of the geometric mean-arithmetic mean inequality, one can prove that

$$\frac{\ln m}{M - m} (MA - B) + \frac{\ln M}{M - m} (B - mA) \leq S (A|B) \tag{27}$$

for positive invertible operators A and B that satisfy the condition (13).

Therefore, it is a natural question to ask how far the right term is from the left term in (27).

In the following, we provide some upper and positive lower bounds for the difference

$$S (A|B) - \frac{\ln m}{M - m} (MA - B) - \frac{\ln M}{M - m} (B - mA)$$

under the above assumptions.

Theorem 3 (Dragomir, 2015 [4]) *Let A, B be two positive invertible operators such that the condition (13) is valid, then we have*

$$0 \leq A^{\frac{1}{2}} \Upsilon_{m,M} \left(A^{-1/2} B A^{-1/2} \right) A^{\frac{1}{2}} \tag{28}$$
$$\leq S (A|B) - \frac{\ln m}{M - m} (MA - B) - \frac{\ln M}{M - m} (B - mA)$$
$$\leq \ln S \left(\frac{M}{m} \right) A$$

where

$$\Upsilon_{m,M} (x) := \ln S \left(\left(\frac{M}{m} \right)^{\frac{1}{2} - \frac{1}{M-m} \left| x - \frac{m+M}{2} \right|} \right) \geq 0 \tag{29}$$

for $x \in [m, M]$.

Proof From (5) we have

$$S\left(\left(\frac{M}{m}\right)^{\min\{\nu,1-\nu\}}\right)m^{1-\nu}M^{\nu}\leq(1-\nu)\,m+\nu M \tag{30}$$
$$\leq S\left(\frac{M}{m}\right)m^{1-\nu}M^{\nu},$$

for any $\nu\in[0,1]$.

If we take in (30) $\nu=\frac{x-m}{M-m}\in[0,1]$ with $x\in[m,M]$, then we get

$$S\left(\left(\frac{M}{m}\right)^{\min\left\{\frac{x-m}{M-m},\frac{M-x}{M-m}\right\}}\right)m^{\frac{M-x}{M-m}}M^{\frac{x-m}{M-m}}\leq x$$
$$\leq S\left(\frac{M}{m}\right)m^{\frac{M-x}{M-m}}M^{\frac{x-m}{M-m}},$$

and by taking the logarithm we obtain

$$\ln S\left(\left(\frac{M}{m}\right)^{\min\left\{\frac{x-m}{M-m},\frac{M-x}{M-m}\right\}}\right) \tag{31}$$
$$\leq\ln x-\frac{M-x}{M-m}\ln m-\frac{x-m}{M-m}\ln M\leq\ln S\left(\frac{M}{m}\right).$$

Since

$$\min\left\{\frac{x-m}{M-m},\frac{M-x}{M-m}\right\}=\frac{1}{2}-\left|\frac{x-\frac{m+M}{2}}{M-m}\right|$$

for any $x\in[m,M]$, then by (31) we get

$$\Upsilon_{m,M}(x)\leq\ln x-\frac{M-x}{M-m}\ln m-\frac{x-m}{M-m}\ln M\leq\ln S\left(\frac{M}{m}\right) \tag{32}$$

for any $x\in[m,M]$, where $\Upsilon_{m,M}$ is (the continuous function) defined by (29).

Using the continuous functional calculus we have from (62) that

$$\Upsilon_{m,M}(X)\leq\ln X-\frac{\ln m}{M-m}(MI-X)-\frac{\ln M}{M-m}(X-mI) \tag{33}$$
$$\leq\ln S\left(\frac{M}{m}\right)I$$

for any self-adjoint operator X with the property that $mI\leq X\leq MI$.

Multiplying both sides of (13) by $A^{-1/2}$ we get

$$mI \leq A^{-1/2}BA^{-1/2} \leq MI$$

and by replacing X by $A^{-1/2}BA^{-1/2}$ in (33) we get

$$\begin{aligned} &\varUpsilon_{m,M}\left(A^{-1/2}BA^{-1/2}\right) \\ &\leq \ln A^{-1/2}BA^{-1/2} \\ &-\frac{\ln m}{M-m}\left(MI - A^{-1/2}BA^{-1/2}\right) - \frac{\ln M}{M-m}\left(A^{-1/2}BA^{-1/2} - mI\right) \\ &\leq \ln S\left(\frac{M}{m}\right) I. \end{aligned} \tag{34}$$

Multiplying both sides of (34) by $A^{1/2}$ we get the desired result (28).

Corollary 3 *Assume that $pI \leq C \leq PI$ for some p, P with $0 < p < P$. Then we have for operator entropy $\eta(C) = -C \ln C$ that*

$$\begin{aligned} 0 \leq C\varPsi_{p,P}\left(C^{-1}\right) \leq \eta(C) + \frac{P\ln P}{P-p}(C - pI) + \frac{p\ln p}{P-p}(PI - C) \\ \leq \ln S\left(\frac{P}{p}\right) C \end{aligned} \tag{35}$$

where

$$\varPsi_{p,P}(x) = \ln S\left(\left(\frac{P}{p}\right)^{\frac{1}{2} - \frac{pP}{P-p}\left|x - \frac{p+P}{2pP}\right|}\right)$$

where $x \in \left[\frac{1}{P}, \frac{1}{p}\right]$.

Proof We have

$$\frac{1}{P}C \leq I \leq \frac{1}{p}C.$$

If we take $B = I,\ A = C,\ m = \frac{1}{P}$ and $M = \frac{1}{p}$ in Theorem 3, then we get

$$\begin{aligned} &C^{\frac{1}{2}}\varUpsilon_{\frac{1}{P},\frac{1}{p}}\left(C^{-1}\right)C^{\frac{1}{2}} \\ &\leq S(C|I) - \frac{\ln\frac{1}{P}}{\frac{1}{p}-\frac{1}{P}}\left(\frac{1}{p}C - I\right) - \frac{\ln\frac{1}{p}}{\frac{1}{p}-\frac{1}{P}}\left(I - \frac{1}{P}C\right) \leq \ln S\left(\frac{P}{p}\right) C, \end{aligned}$$

namely

$$\begin{aligned}C^{\frac{1}{2}}\Psi_{p,P}\left(C^{-1}\right)C^{\frac{1}{2}} &= C\Psi_{p,P}\left(C^{-1}\right)\\ &\leq S\left(C|I\right)+\frac{P\ln P}{P-p}\left(C-pI\right)+\frac{p\ln p}{P-p}\left(PI-C\right)\\ &\leq \ln S\left(\frac{P}{p}\right)C,\end{aligned}$$

where

$$\Upsilon_{\frac{1}{P},\frac{1}{p}}\left(x\right)=\Psi_{p,P}\left(x\right)=\ln S\left(\left(\frac{P}{p}\right)^{\frac{1}{2}-\frac{pP}{P-p}\left|x-\frac{p+P}{2pP}\right|}\right),$$

with $x\in\left[\frac{1}{P},\frac{1}{p}\right]$.

We also have:

Theorem 4 (Dragomir, 2015 [4]) *With the assumptions of Theorem 3 we have*

$$\begin{aligned}0 &\leq \left(\frac{1}{2}A-\frac{1}{M-m}A^{1/2}\left|A^{-1/2}\left(B-\frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2}\right)\\ &\times K\left(\frac{M}{m}\right)\\ &\leq S\left(A|B\right)-\frac{\ln m}{M-m}\left(MA-B\right)-\frac{\ln M}{M-m}\left(B-mA\right)\\ &\leq \left(\frac{1}{2}A+\frac{1}{M-m}A^{1/2}\left|A^{-1/2}\left(B-\frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2}\right)\\ &\times K\left(\frac{M}{m}\right).\end{aligned} \tag{36}$$

Proof Using the inequality (7) we have

$$\begin{aligned}K^{\min\{\nu,1-\nu\}}\left(\frac{M}{m}\right)m^{1-\nu}M^{\nu} &\leq \left(1-\nu\right)m+\nu M\\ &\leq K^{\max\{\nu,1-\nu\}}\left(\frac{M}{m}\right)m^{1-\nu}M^{\nu}\end{aligned} \tag{37}$$

for any $\nu\in\left[0,1\right]$.

If we take in (37) $\nu = \frac{x-m}{M-m} \in [0,1]$ with $x \in [m, M]$, then we get

$$K^{\min\left\{\frac{x-m}{M-m}, \frac{M-x}{M-m}\right\}}\left(\frac{M}{m}\right) m^{\frac{M-x}{M-m}} M^{\frac{x-m}{M-m}}$$
$$\leq x \leq K^{\max\left\{\frac{x-m}{M-m}, \frac{M-x}{M-m}\right\}}\left(\frac{M}{m}\right) m^{\frac{M-x}{M-m}} M^{\frac{x-m}{M-m}},$$

which is equivalent to

$$(0 \leq) \min\left\{\frac{x-m}{M-m}, \frac{M-x}{M-m}\right\} K\left(\frac{M}{m}\right)$$
$$\leq \ln x - \frac{M-x}{M-m}\ln m - \frac{x-m}{M-m}\ln M$$
$$\leq \max\left\{\frac{x-m}{M-m}, \frac{M-x}{M-m}\right\} K\left(\frac{M}{m}\right)$$

or to

$$(0 \leq) \left(\frac{1}{2} - \frac{1}{M-m}\left|x - \frac{m+M}{2}\right|\right) K\left(\frac{M}{m}\right)$$
$$\leq \ln x - \frac{M-x}{M-m}\ln m - \frac{x-m}{M-m}\ln M$$
$$\leq \left(\frac{1}{2} + \frac{1}{M-m}\left|x - \frac{m+M}{2}\right|\right) K\left(\frac{M}{m}\right).$$

By making use of a similar argument to the one in the proof of Theorem 3 we get the desired result (36).

Remark 2 If A and B commute, then

$$A^{1/2}\left|A^{-1/2}\left(B - \frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2} = \left|B - \frac{m+M}{2}A\right|,$$
$$S(A|B) = A(\ln B - \ln A)$$

and by (36) we have

$$0 \leq \left(\frac{1}{2}A - \frac{1}{M-m}\left|B - \frac{m+M}{2}A\right|\right) K\left(\frac{M}{m}\right) \tag{38}$$
$$\leq A(\ln B - \ln A) - \frac{\ln m}{M-m}(MA - B) - \frac{\ln M}{M-m}(B - mA)$$
$$\leq \left(\frac{1}{2}A + \frac{1}{M-m}\left|B - \frac{m+M}{2}A\right|\right) K\left(\frac{M}{m}\right).$$

Corollary 4 *With the assumptions of Corollary 3 we have*

$$\begin{aligned}&\left(\frac{1}{2}C-\frac{pP}{P-p}\left|I-\frac{p+P}{2pP}C\right|\right)K\left(\frac{P}{p}\right)\\&\le\eta\left(C\right)+\frac{P\ln P}{P-p}\left(C-pI\right)+\frac{p\ln p}{P-p}\left(PI-C\right)\\&\le\left(\frac{1}{2}C+\frac{pP}{P-p}\left|I-\frac{p+P}{2pP}C\right|\right)K\left(\frac{P}{p}\right).\end{aligned}\tag{39}$$

Proof Follows by Theorem 4 on choosing $B=I,\ A=C,\ m=\frac{1}{P}$ and $M=\frac{1}{p}$ and taking into account that, by the continuous functional calculus for C, we have

$$C^{1/2}\left|C^{-1/2}\left(I-\frac{p+P}{2pP}C\right)C^{-1/2}\right|C^{1/2}=\left|I-\frac{p+P}{2pP}C\right|.$$

Theorem 5 (Dragomir, 2015 [4]) *With the assumptions of Theorem 3 we have*

$$\begin{aligned}&(0\le)\,S\left(A|B\right)-\frac{\ln m}{M-m}\left(MA-B\right)-\frac{\ln M}{M-m}\left(B-mA\right)\\&\le\frac{4}{\left(M-m\right)^2}\left(K\left(\frac{M}{m}\right)-1\right)\left(B-mA\right)A^{-1}\left(MA-B\right)\\&\le\left(K\left(\frac{M}{m}\right)-1\right)A.\end{aligned}\tag{40}$$

Proof From the inequality (10) we have

$$(1\le)\,\frac{\left(1-\nu\right)m+\nu M}{m^{1-\nu}M^{\nu}}\le\exp\left[4\nu\left(1-\nu\right)\left(K\left(\frac{M}{m}\right)-1\right)\right],\tag{41}$$

for any $\nu\in[0,1]$.

If we take in (41) $\nu=\frac{x-m}{M-m}\in[0,1]$ with $x\in[m,M]$ then we get

$$(1\le)\,\frac{x}{m^{\frac{M-x}{M-m}}M^{\frac{x-m}{M-m}}}\le\exp\left[\frac{4\left(x-m\right)\left(M-x\right)}{\left(M-m\right)^2}\left(K\left(\frac{M}{m}\right)-1\right)\right].\tag{42}$$

Taking the logarithm in (42) we get

$$\begin{aligned}&(0\le)\ln x-\frac{M-x}{M-m}\ln m-\frac{x-m}{M-m}\ln M\\&\le\frac{4\left(x-m\right)\left(M-x\right)}{\left(M-m\right)^2}\left(K\left(\frac{M}{m}\right)-1\right)\end{aligned}$$

for any $x\in[m,M]$.

Making use of a similar argument to the one from the proof of Theorem 3 we get

$$\begin{aligned}&S\left(A|B\right)-\frac{\ln m}{M-m}\left(MA-B\right)-\frac{\ln M}{M-m}\left(B-mA\right)\\&\leq\frac{4}{\left(M-m\right)^{2}}\left(K\left(\frac{M}{m}\right)-1\right)\\&\times A^{1/2}\left(A^{-1/2}BA^{-1/2}-m\right)\left(M-A^{-1/2}BA^{-1/2}\right)A^{1/2}\end{aligned}$$

and since

$$\begin{aligned}&A^{1/2}\left(A^{-1/2}BA^{-1/2}-m\right)\left(M-A^{-1/2}BA^{-1/2}\right)A^{1/2}\\&=A^{1/2}\left(A^{-1/2}\left(B-mA\right)A^{-1/2}\right)\left(A^{-1/2}\left(MA-B\right)A^{-1/2}\right)A^{1/2}\\&=\left(B-mA\right)A^{-1}\left(MA-B\right),\end{aligned}$$

we obtain the first part of (40).

The second part follows by the inequality

$$\frac{4\left(x-m\right)\left(M-x\right)}{\left(M-m\right)^{2}}\leq 1$$

for any $x\in[m,M]$.

Corollary 5 *With the assumptions of Corollary 3 we have*

$$\begin{aligned}&(0\leq)\,\eta\left(C\right)+\frac{P\ln P}{P-p}\left(C-pI\right)+\frac{p\ln p}{P-p}\left(PI-C\right)\\&\leq\frac{4pP}{\left(P-p\right)^{2}}\left(K\left(\frac{P}{p}\right)-1\right)\left(IP-C\right)C^{-1}\left(C-Ip\right)\\&\leq\left(K\left(\frac{P}{p}\right)-1\right)C.\end{aligned}\tag{43}$$

Finally, we have:

Theorem 6 (Dragomir, 2015 [4]) *With the assumptions of Theorem 3 we have*

$$\begin{aligned}&\frac{1}{2M^{2}}\left(B-mA\right)A^{-1}\left(MA-B\right)\\&\leq S\left(A|B\right)-\frac{\ln m}{M-m}\left(MA-B\right)-\frac{\ln M}{M-m}\left(B-mA\right)\\&\leq\frac{1}{2m^{2}}\left(B-mA\right)A^{-1}\left(MA-B\right)\end{aligned}\tag{44}$$

Proof From the inequality (12) we have

$$\begin{aligned}&\exp\left[\frac{1}{2}\nu\left(1-\nu\right)\left(1-\frac{m}{M}\right)^2\right]\\&\leq\frac{\left(1-\nu\right)m+\nu M}{m^{1-\nu}M^{\nu}}\\&\leq\exp\left[\frac{1}{2}\nu\left(1-\nu\right)\left(\frac{M}{m}-1\right)^2\right]\end{aligned}\tag{45}$$

for any $\nu\in[0,1]$.

If we take in (41) $\nu=\frac{x-m}{M-m}\in[0,1]$ with $x\in[m,M]$, then we get

$$\begin{aligned}&\exp\left[\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{\left(M-m\right)^2}\left(1-\frac{m}{M}\right)^2\right]\\&\leq\frac{x}{m^{\frac{M-x}{M-m}}M^{\frac{x-m}{M-m}}}\\&\leq\exp\left[\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{\left(M-m\right)^2}\left(\frac{M}{m}-1\right)^2\right]\end{aligned}$$

that is equivalent to

$$\begin{aligned}\exp\left[\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{M^2}\right]&\leq\frac{x}{m^{\frac{M-x}{M-m}}M^{\frac{x-m}{M-m}}}\\&\leq\exp\left[\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{m^2}\right].\end{aligned}$$

On taking the logarithm, we get

$$\begin{aligned}\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{M^2}&\leq\ln x-\frac{M-x}{M-m}\ln m-\frac{x-m}{M-m}\ln M\\&\leq\frac{1}{2}\frac{\left(x-m\right)\left(M-x\right)}{m^2},\end{aligned}\tag{46}$$

for any $x\in[m,M]$.

Making use of a similar argument to the one from the proofs of Theorems 3 and 5 we get the desired result (44).

Corollary 6 *With the assumptions of Corollary 3 we have*

$$\frac{1}{2}\frac{p}{P}\left(IP-C\right)C^{-1}\left(C-Ip\right)\tag{47}$$

$$\leq \eta(C) + \frac{P\ln P}{P-p}(C-pI) + \frac{p\ln p}{P-p}(PI-C)$$
$$\leq \frac{1}{2}\frac{P}{p}(IP-C)\,C^{-1}(C-Ip).$$

Absolute Value Upper and Lower Bounds

Observe that, if we replace in (2) B with A, then we get

$$\begin{aligned} S(B|A) &= A^{1/2}\eta\left(A^{-1/2}BA^{-1/2}\right)A^{1/2} \\ &= A^{1/2}\left(-A^{-1/2}BA^{-1/2}\ln\left(A^{-1/2}BA^{-1/2}\right)\right)A^{1/2}, \end{aligned}$$

therefore we have

$$A^{1/2}\left(A^{-1/2}BA^{-1/2}\ln\left(A^{-1/2}BA^{-1/2}\right)\right)A^{1/2} = -S(B|A)$$

for positive invertible operators A and B.

It is well known that, in general $S(A|B)$ is not equal to $S(B|A)$.

Motivated by the above results, we establish in this paper some bounds for the quantity $S(B|A)$ under the same assumptions (13) for the operators A and B. For this purpose, we use some scalar inequalities for convex functions from [1, 2] and [3]. Applications for the operator entropy $\eta(C) = -C\ln C = S(C|I)$ under the natural assumption $pI \leq C \leq PI$ for some constants $p,\ P$ with $0 < p < P$, are also provided.

We have:

Theorem 7 (Dragomir, 2015 [5]) *Let $A,\ B$ be two positive invertible operators such that the condition (13) is valid, then we have*

$$\begin{aligned} &2\left(\frac{1}{2}A - \frac{1}{M-m}A^{1/2}\left|A^{-1/2}\left(B-\frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2}\right) \\ &\times K(m,M) \\ &\leq \frac{m\ln m}{M-m}(MA-B) + \frac{M\ln M}{M-m}(B-mA) + S(B|A) \\ &\leq 2\left(\frac{1}{2}A + \frac{1}{M-m}A^{1/2}\left|A^{-1/2}\left(B-\frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2}\right) \\ &\times K(m,M), \end{aligned} \tag{48}$$

where

$$K(m,M):=\left[\frac{m\ln m+M\ln M}{2}-\left(\frac{m+M}{2}\right)\ln\left(\frac{m+M}{2}\right)\right]$$
$$=\ln\left(\frac{G\left(m^m,M^M\right)}{[A(m,M)]^{A(m,M)}}\right)$$

and $G(a,b):=\sqrt{ab}$ *is the geometric mean while* $A(a,b):=\frac{a+b}{2}$ *is the arithmetic mean of positive numbers* $a,\ b$.

Proof Recall the following result obtained by the author in 2006 [1] that provides a refinement and a reverse for the weighted Jensen's discrete inequality:

$$n\min_{j\in\{1,2,\dots,n\}}\left\{p_j\right\}\left[\frac{1}{n}\sum_{j=1}^{n}\Phi\left(x_j\right)-\Phi\left(\frac{1}{n}\sum_{j=1}^{n}x_j\right)\right] \tag{49}$$
$$\le\frac{1}{P_n}\sum_{j=1}^{n}p_j\Phi\left(x_j\right)-\Phi\left(\frac{1}{P_n}\sum_{j=1}^{n}p_jx_j\right)$$
$$\le n\max_{j\in\{1,2,\dots,n\}}\left\{p_j\right\}\left[\frac{1}{n}\sum_{j=1}^{n}\Phi\left(x_j\right)-\Phi\left(\frac{1}{n}\sum_{j=1}^{n}x_j\right)\right],$$

where $\Phi:C\to\mathbb{R}$ is a convex function defined on convex subset C of the linear space X, $\left\{x_j\right\}_{j\in\{1,2,\dots,n\}}$are vectors in C and $\left\{p_j\right\}_{j\in\{1,2,\dots,n\}}$are nonnegative numbers with $P_n=\sum_{j=1}^{n}p_j>0$. For $n=2$, we deduce from (49) that

$$2r\left[\frac{\Phi(x)+\Phi(y)}{2}-\Phi\left(\frac{x+y}{2}\right)\right] \tag{50}$$
$$\le\nu\Phi(x)+(1-\nu)\Phi(y)-\Phi[\nu x+(1-\nu)y]$$
$$\le 2R\left[\frac{\Phi(x)+\Phi(y)}{2}-\Phi\left(\frac{x+y}{2}\right)\right]$$

for any $x,\ y\in\mathbb{R}$ and $\nu\in[0,1]$, where $r=\min\{\nu,1-\nu\}$ and $R=\max\{\nu,1-\nu\}$.

Now, if we take in (50) the convex function $\Phi(t)=t\ln t,\ t>0$, then we get

$$2r\left[\frac{x\ln x+y\ln y}{2}-\left(\frac{x+y}{2}\right)\ln\left(\frac{x+y}{2}\right)\right] \tag{51}$$
$$\le\nu x\ln x+(1-\nu)y\ln y-[\nu x+(1-\nu)y]\ln[\nu x+(1-\nu)y]$$
$$\le 2R\left[\frac{x\ln x+y\ln y}{2}-\left(\frac{x+y}{2}\right)\ln\left(\frac{x+y}{2}\right)\right]$$

for any $x,\ y>0$ and $\nu\in[0,1]$.

This is an inequality of interest in itself as well.

Now, if we take in (51) $x = m,\ y = M$ and $\nu = \frac{M-u}{M-m} \in [0,1]$ with $u \in [m,M]$ then we get

$$\begin{aligned}
&2\min\left\{\frac{M-u}{M-m},\frac{u-m}{M-m}\right\} \\
&\times\left[\frac{m\ln m + M\ln M}{2} - \left(\frac{m+M}{2}\right)\ln\left(\frac{m+M}{2}\right)\right] \\
&\le \frac{M-u}{M-m}m\ln m + \frac{u-m}{M-m}M\ln M - u\ln u \\
&\le 2\max\left\{\frac{M-u}{M-m},\frac{u-m}{M-m}\right\} \\
&\times\left[\frac{m\ln m + M\ln M}{2} - \left(\frac{m+M}{2}\right)\ln\left(\frac{m+M}{2}\right)\right].
\end{aligned} \tag{52}$$

Since

$$\min\left\{\frac{M-u}{M-m},\frac{u-m}{M-m}\right\} = \frac{1}{2} - \left|\frac{u-\frac{m+M}{2}}{M-m}\right|$$

and

$$\max\left\{\frac{M-u}{M-m},\frac{u-m}{M-m}\right\} = \frac{1}{2} + \left|\frac{u-\frac{m+M}{2}}{M-m}\right|,$$

then from (52) we have

$$\begin{aligned}
&2\left(\frac{1}{2} - \frac{1}{M-m}\left|u - \frac{m+M}{2}\right|\right)K(m,M) \\
&\le \frac{M-u}{M-m}m\ln m + \frac{u-m}{M-m}M\ln M - u\ln u \\
&\le 2\left(\frac{1}{2} + \frac{1}{M-m}\left|u - \frac{m+M}{2}\right|\right)K(m,M)
\end{aligned} \tag{53}$$

for any $u \in [m,M]$.

Using the continuous functional calculus we have from (53) that

$$2\left(\frac{1}{2}I - \frac{1}{M-m}\left|X - \frac{m+M}{2}I\right|\right)K(m,M) \tag{54}$$

$$\leq m\ln m\frac{MI-X}{M-m}+M\ln M\frac{X-mI}{M-m}-X\ln X$$
$$\leq 2\left(\frac{1}{2}I+\frac{1}{M-m}\left|X-\frac{m+M}{2}I\right|\right)K(m,M)$$

for any self-adjoint operator X with the property that $mI\leq X\leq MI$.

Multiplying both sides of (13) by $A^{-1/2}$ we get

$$mI\leq A^{-1/2}BA^{-1/2}\leq MI$$

and by replacing X by $A^{-1/2}BA^{-1/2}$ in (54) we obtain

$$\begin{aligned}&2\left(\frac{1}{2}I-\frac{1}{M-m}\left|A^{-1/2}BA^{-1/2}-\frac{m+M}{2}I\right|\right)K(m,M)\\&\leq m\ln m\frac{MI-A^{-1/2}BA^{-1/2}}{M-m}+M\ln M\frac{A^{-1/2}BA^{-1/2}-mI}{M-m}\\&-A^{-1/2}BA^{-1/2}\ln(A^{-1/2}BA^{-1/2})\\&\leq 2\left(\frac{1}{2}I+\frac{1}{M-m}\left|A^{-1/2}BA^{-1/2}-\frac{m+M}{2}I\right|\right)K(m,M).\end{aligned}\tag{55}$$

Multiplying both sides of (55) by $A^{1/2}$ we get the desired result (48).

Remark 3 If A and B commute, then

$$A^{1/2}\left|A^{-1/2}\left(B-\frac{m+M}{2}A\right)A^{-1/2}\right|A^{1/2}=\left|B-\frac{m+M}{2}A\right|,$$

$$S(B|A)=B(\ln A-\ln B)$$

and by (48) we have

$$\begin{aligned}&(0\leq)\,2\left(\frac{1}{2}A-\frac{1}{M-m}\left|B-\frac{m+M}{2}A\right|\right)K(m,M)\\&\leq\frac{m\ln m}{M-m}(MA-B)+\frac{M\ln M}{M-m}(B-mA)+B(\ln A-\ln B)\\&\leq 2\left(\frac{1}{2}A+\frac{1}{M-m}\left|B-\frac{m+M}{2}A\right|\right)K(m,M).\end{aligned}\tag{56}$$

The above result can be applied for the operator entropy

$$\eta(C)=-C\ln C=S(C|I)$$

as follows:

Corollary 7 *Assume that $pI \leq C \leq PI$ for some p, P with $0 < p < P$. Then we have that*

$$\begin{aligned}(0 \leq)\, 2\left(\frac{1}{2}I - \frac{1}{P-p}\left|C - \frac{p+P}{2}I\right|\right) K(p,P) \\ \leq \frac{p\ln p}{P-p}(PI - C) + \frac{P\ln P}{P-p}(C - pI) + \eta(C) \\ \leq 2\left(\frac{1}{2}I + \frac{1}{P-p}\left|C - \frac{p+P}{2}I\right|\right) K(p,P).\end{aligned} \tag{57}$$

An Upper Bound in Terms of Logarithm

We have the following inequality of interest for convex functions, see, for instance, [2]:

Lemma 1 *Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a convex function on the interval I, $a, b \in \mathring{I}$, the interior of I, with $a < b$ and $\nu \in [0,1]$. Then*

$$\begin{aligned}\nu(1-\nu)(b-a)\left[f'_+((1-\nu)a + \nu b) - f'_-((1-\nu)a + \nu b)\right] \\ \leq (1-\nu)f(a) + \nu f(b) - f((1-\nu)a + \nu b) \\ \leq \nu(1-\nu)(b-a)\left[f'_-(b) - f'_+(a)\right].\end{aligned} \tag{58}$$

In particular, we have

$$\begin{aligned}\frac{1}{4}(b-a)\left[f'_+\left(\frac{a+b}{2}\right) - f'_-\left(\frac{a+b}{2}\right)\right] \\ \leq \frac{f(a)+f(b)}{2} - f\left(\frac{a+b}{2}\right) \\ \leq \frac{1}{4}(b-a)\left[f'_-(b) - f'_+(a)\right].\end{aligned} \tag{59}$$

The constant $\frac{1}{4}$ is best possible in both inequalities from (59).

Proof The case $\nu = 0$ or $\nu = 1$ reduces to equality in (58).

Since f is convex on I it follows that the function is differentiable on $\mathring{I}$ except a countably number of points, the lateral derivatives $f'_\pm$ exist in each point of $\mathring{I}$, they are increasing on $\mathring{I}$ and $f'_- \leq f'_+$ on $\mathring{I}$.

For any $x, y \in \mathring{I}$ we have for the Lebesgue integral

$$f(x) = f(y) + \int_y^x f'(s)\,ds = f(y) + (x-y)\int_0^1 f'((1-t)y + tx)\,dt. \tag{60}$$

Assume that $a < b$ and $\nu \in (0,1)$. By (60) we have

$$f((1-\nu)a+\nu b) \qquad (61)$$
$$= f(a) + \nu(b-a)\int_0^1 f'((1-t)a + t((1-\nu)a+\nu b))\,dt$$

and

$$f((1-\nu)a+\nu b) \qquad (62)$$
$$= f(b) - (1-\nu)(b-a)\int_0^1 f'((1-t)b + t((1-\nu)a+\nu b))\,dt.$$

If we multiply (61) by $1-\nu$, (61) by ν and add the obtained equalities, then we get

$$\begin{aligned} f((1-\nu)a+\nu b) = (1-\nu)f(a) + \nu f(b) &\\ + (1-\nu)\nu(b-a)\int_0^1 f'((1-t)a + t((1-\nu)a+\nu b))\,dt &\\ - (1-\nu)\nu(b-a)\int_0^1 f'((1-t)b + t((1-\nu)a+\nu b))\,dt, & \end{aligned}$$

which is equivalent to

$$\begin{aligned} &(1-\nu)f(a) + \nu f(b) - f((1-\nu)a+\nu b) \qquad (63)\\ &= (1-\nu)\nu(b-a)\\ &\times \int_0^1 \left[f'((1-t)b + t((1-\nu)a+\nu b)) - f'((1-t)a + t((1-\nu)a+\nu b))\right]dt. \end{aligned}$$

That is an equality of interest in itself.

Since $a < b$ and $\nu \in (0,1)$, then $(1-\nu)a+\nu b \in (a,b)$ and

$$(1-t)a + t((1-\nu)a+\nu b) \in [a, (1-\nu)a+\nu b]$$

while

$$(1-t)b + t((1-\nu)a+\nu b) \in [(1-\nu)a+\nu b, b]$$

for any $t \in [0,1]$.

By the monotonicity of the derivative we have

$$f'_+((1-\nu)a+\nu b) \le f'((1-t)b + t((1-\nu)a+\nu b)) \le f'_-(b) \qquad (64)$$

and

$$f'_+(a) \le f'((1-t)a + t((1-\nu)a + \nu b)) \le f'_-((1-\nu)a + \nu b) \tag{65}$$

for any $t \in [0,1]$.

By integrating the inequalities (64) and (65) we get

$$f'_+((1-\nu)a + \nu b) \le \int_0^1 f'((1-t)b + t((1-\nu)a + \nu b))\,dt \le f'_-(b)$$

and

$$f'_+(a) \le \int_0^1 f'((1-t)a + t((1-\nu)a + \nu b))\,dt \le f'_-((1-\nu)a + \nu b),$$

which implies that

$$\begin{aligned}
&f'_+((1-\nu)a + \nu b) - f'_-((1-\nu)a + \nu b)\\
&\le \int_0^1 f'((1-t)b + t((1-\nu)a + \nu b))\,dt\\
&- \int_0^1 f'((1-t)a + t((1-\nu)a + \nu b))\,dt\\
&\le f'_-(b) - f'_+(a).
\end{aligned}$$

Making use of the equality (63) we obtain the desired result (58).

If we consider the convex function $f : [a,b] \to \mathbb{R}, f(x) = \left|x - \frac{a+b}{2}\right|$, then we have $f'_+\left(\frac{a+b}{2}\right) = 1$, $f'_-\left(\frac{a+b}{2}\right) = -1$ and by replacing in (59) we get in all terms the same quantity $\frac{1}{2}(b-a)$ which show that the constant $\frac{1}{4}$ is best possible in both inequalities from (59).

We can state the following result:

Theorem 8 (Dragomir, 2015 [5]) *Let A, B be two positive invertible operators such that the condition (13) is valid, then we have*

$$\begin{aligned}
&(0 \le) \frac{m \ln m}{M-m}(MA - B) + \frac{M \ln M}{M-m}(B - mA) + S(B|A) \\
&\le \frac{\ln M - \ln m}{M-m}(B - mA)A^{-1}(MA - B)\\
&\le \frac{1}{4}(M-m)(\ln M - \ln m)A.
\end{aligned} \tag{66}$$

Proof If we consider the differentiable convex function $f(t) = t \ln t,\ t > 0$, then $f'(t) = \ln t + 1$ and by (58) we have

$$0 \le (1-\nu)\, a \ln a + \nu b \ln b - ((1-\nu)\, a + \nu b) \ln ((1-\nu)\, a + \nu b) \le \nu (1-\nu)(b-a)(\ln b - \ln a) \tag{67}$$

for any $a, b > 0$ and $\nu \in [0, 1]$.

On applying the inequality (67) on the interval $[m, M]$ and for $\nu = \frac{x-m}{M-m} \in [0, 1]$ with $x \in [m, M]$ then we get

$$\begin{aligned} 0 &\le m \ln m \frac{M-x}{M-m} + M \ln M \frac{x-m}{M-m} - x \ln x \\ &\le \frac{(x-m)(M-x)}{M-m} (\ln M - \ln m) \\ &\le \frac{1}{4} (M-m)(\ln M - \ln m). \end{aligned} \tag{68}$$

Using the continuous functional calculus we have from (68) that

$$\begin{aligned} 0 &\le m \ln m \frac{MI - X}{M-m} + M \ln M \frac{X - mI}{M-m} - X \ln X \\ &\le (\ln M - \ln m) \frac{(X - mI)(M - XI)}{M-m} \\ &\le \frac{1}{4} (M-m)(\ln M - \ln m) I \end{aligned} \tag{69}$$

for any self-adjoint operator X with the property that $mI \le X \le MI$.

By replacing X by $A^{-1/2}BA^{-1/2}$ in (65) we get

$$\begin{aligned} 0 &\le m \ln m \frac{MI - A^{-1/2}BA^{-1/2}}{M-m} + M \ln M \frac{A^{-1/2}BA^{-1/2} - mI}{M-m} \\ &- A^{-1/2}BA^{-1/2} \ln(A^{-1/2}BA^{-1/2}) \\ &\le (\ln M - \ln m) \frac{\left(A^{-1/2}BA^{-1/2} - mI\right)\left(MI - A^{-1/2}BA^{-1/2}\right)}{M-m} \\ &\le \frac{1}{4} (M-m)(\ln M - \ln m) I. \end{aligned} \tag{70}$$

Multiplying both sides of (70) by $A^{1/2}$ we get the desired result (66).

Corollary 8 *Assume that $pI \le C \le PI$ for some $p,\ P$ with $0 < p < P$. Then we have that*

$$\begin{aligned}
&(0 \leq) \frac{p\ln p}{P-p}\left(PI-C\right) + \frac{P\ln P}{P-p}\left(C-pI\right) + \eta\left(C\right) \\
&\leq \frac{\ln P-\ln p}{P-p}\left(C-pI\right)\left(PI-C\right) \leq \frac{1}{4}\left(P-p\right)\left(\ln P-\ln p\right).
\end{aligned} \tag{71}$$

Further Lower and Upper Bounds

We have the following result, see for instance [3]:

Lemma 2 *Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a twice differentiable function on the interval $\mathring{I}$, the interior of I. If there exists the constants $d,\ D$ such that*

$$d \leq f''\left(t\right) \leq D \text{ for any } t \in \mathring{I}, \tag{72}$$

then

$$\begin{aligned}
\frac{1}{2}\nu\left(1-\nu\right)d\left(b-a\right)^2 &\leq (1-\nu)f\left(a\right) + \nu f\left(b\right) - f\left((1-\nu)\,a + \nu b\right) \\
&\leq \frac{1}{2}\nu\left(1-\nu\right)D\left(b-a\right)^2
\end{aligned} \tag{73}$$

for any $a,\ b \in \mathring{I}$ and $\nu \in [0,1]$.

In particular, we have

$$\frac{1}{8}\left(b-a\right)^2 d \leq \frac{f\left(a\right)+f\left(b\right)}{2} - f\left(\frac{a+b}{2}\right) \leq \frac{1}{8}\left(b-a\right)^2 D, \tag{74}$$

for any $a,\ b \in \mathring{I}$.

The constant $\frac{1}{8}$ is best possible in both inequalities in (74).

Proof We consider the auxiliary function $f_D : I \subset \mathbb{R} \to \mathbb{R}$ defined by $f_D\left(x\right) = \frac{1}{2}Dx^2 - f\left(x\right)$. The function f_D is differentiable on $\mathring{I}$ and $f_D''\left(x\right) = D - f''\left(x\right) \geq 0$, showing that f_D is a convex function on $\mathring{I}$.

By the convexity of f_D we have for any $a,\ b \in \mathring{I}$ and $\nu \in [0,1]$ that

$$\begin{aligned}
0 &\leq (1-\nu)f_D\left(a\right) + \nu f_D\left(b\right) - f_D\left((1-\nu)\,a + \nu b\right) \\
&= (1-\nu)\left(\frac{1}{2}Da^2 - f\left(a\right)\right) + \nu\left(\frac{1}{2}Db^2 - f\left(b\right)\right) \\
&- \left(\frac{1}{2}D\left((1-\nu)\,a + \nu b\right)^2 - f_D\left((1-\nu)\,a + \nu b\right)\right)
\end{aligned}$$

$$= \frac{1}{2}D\left[(1-\nu)a^2 + \nu b^2 - ((1-\nu)a + \nu b)^2\right]$$
$$-(1-\nu)f(a) - \nu f(b) + f_D((1-\nu)a + \nu b)$$
$$= \frac{1}{2}\nu(1-\nu)D(b-a)^2 - (1-\nu)f(a) - \nu f(b) + f_D((1-\nu)a + \nu b),$$

which implies the second inequality in (73).

The first inequality follows in a similar way by considering the auxiliary function $f_d : I \subset \mathbb{R} \to \mathbb{R}$ defined by $f_D(x) = f(x) - \frac{1}{2}dx^2$ that is twice differentiable and convex on $\mathring{I}$.

If we take $f(x) = x^2$, then (13) holds with equality for $d = D = 2$ and (59) reduces to an equality as well.

If $D > 0$, the second inequality in (73) is better than the corresponding inequality obtained by Furuichi and Minculete in [11] by applying Lagrange's theorem two times. They had instead of $\frac{1}{2}$ the constant 1. Our method also allowed to obtain, for $d > 0$, a lower bound that cannot be established by Lagrange's theorem method employed in [11].

We can state the following result:

Theorem 9 (Dragomir, 2015 [5]) *Let A, B be two positive invertible operators such that the condition (13) is valid, then we have*

$$(0 \le)\, \frac{1}{2M}(B - mA)A^{-1}(MA - B) \tag{75}$$
$$\le \frac{m\ln m}{M-m}(MA - B) + \frac{M\ln M}{M-m}(B - mA) + S(B|A)$$
$$\le \frac{1}{2m}(B - mA)A^{-1}(MA - B).$$

Proof If we consider the convex function $f(t) = t\ln t,\ t > 0$, then $f''(t) = \frac{1}{t}$ and by (73) we have

$$\frac{1}{2}\nu(1-\nu)\frac{1}{\max\{a,b\}}(b-a)^2 \tag{76}$$
$$\le (1-\nu)a\ln a + \nu b\ln b - ((1-\nu)a + \nu b)\ln((1-\nu)a + \nu b)$$
$$\le \frac{1}{2}\nu(1-\nu)\frac{1}{\min\{a,b\}}(b-a)^2$$

for any $a,\ b > 0$ and $\nu \in [0,1]$.

On applying the inequality (76) on the interval $[m, M]$ and for $\nu = \frac{x-m}{M-m} \in [0,1]$ with $x \in [m, M]$ then we get

$$\frac{1}{2M}(x-m)(M-x) \le \frac{M-x}{M-m} m \ln m + \frac{x-m}{M-m} M \ln M - x \ln x \tag{77}$$
$$\le \frac{1}{2m}(x-m)(M-x).$$

Using the continuous functional calculus we have from (77) that

$$\frac{1}{2M}(X-mI)(M-XI) \le \frac{MI-X}{M-m} m \ln m + \frac{X-mI}{M-m} M \ln M - X \ln X \tag{78}$$
$$\le \frac{1}{2m}(X-mI)(M-XI)$$

for any self-adjoint operator X with the property that $mI \le X \le MI$.

Now, on using a similar argument to the one in the proof of Theorem 8 we deduce the desired result (75).

Finally, we have

Corollary 9 *Assume that $pI \le C \le PI$ for some p, P with $0 < p < P$. Then we have the inequalities*

$$(0 \le)\, \frac{1}{2P}(C-pI)(PI-C) \tag{79}$$
$$\le \frac{p \ln p}{P-p}(PI-C) + \frac{P \ln P}{P-p}(C-pI) + \eta(C)$$
$$\le \frac{1}{2p}(C-pI)(PI-C).$$

Applications for Trace Inequalities

If $\{e_i\}_{i\in I}$ is an orthonormal basis of H, we say that $A \in \mathcal{B}(H)$ is *trace class* provided

$$\|A\|_1 := \sum_{i\in I} \langle |A| e_i, e_i \rangle < \infty. \tag{80}$$

The definition of $\|A\|_1$ does not depend on the choice of the orthonormal basis $\{e_i\}_{i\in I}$. We denote by $\mathcal{B}_1(H)$ the set of trace class operators in $\mathcal{B}(H)$.

The following properties are also well known:

(i) We have

$$\|A\|_1 = \|A^*\|_1$$

for any $A \in \mathcal{B}_1(H)$;

(ii) $\mathcal{B}_1(H)$ is an *operator ideal* in $\mathcal{B}(H)$, i.e.

$$\mathcal{B}(H)\,\mathcal{B}_1(H)\,\mathcal{B}(H) \subseteq \mathcal{B}_1(H);$$

(iii) $(\mathcal{B}_1(H), \|\cdot\|_1)$ is a *Banach space*.

We define the *trace* of a trace class operator $A \in \mathcal{B}_1(H)$ to be

$$\operatorname{tr}(A) := \sum_{i\in I} \langle Ae_i, e_i\rangle, \tag{81}$$

where $\{e_i\}_{i\in I}$ is an orthonormal basis of H. Note that this coincides with the usual definition of the trace if H is finite-dimensional. We observe that the series (81) converges absolutely and it is independent from the choice of basis.

The following results collect some properties of the trace:

(i) If $A \in \mathcal{B}_1(H)$, then $A^* \in \mathcal{B}_1(H)$ and

$$\operatorname{tr}\left(A^*\right) = \overline{\operatorname{tr}(A)};$$

(ii) If $A \in \mathcal{B}_1(H)$ and $T \in \mathcal{B}(H)$, then $AT,\ TA \in \mathcal{B}_1(H)$ and

$$\operatorname{tr}(AT) = \operatorname{tr}(TA) \text{ and } |\operatorname{tr}(AT)| \le \|A\|_1 \|T\|; \tag{82}$$

(iii) $\operatorname{tr}(\cdot)$ is a bounded linear functional on $\mathcal{B}_1(H)$ with $\|\operatorname{tr}\| = 1$;

(iv) $\mathcal{B}_{fin}(H)$, the space of operators of *finite rank*, is a dense subspace of $\mathcal{B}_1(H)$.

In the recent paper [4] we have showed amongst other that

$$\begin{aligned}(0 \le)\, S(A|B) - \frac{\ln m}{M-m}(MA-B) - \frac{\ln M}{M-m}(B-mA) \\ \le \ln S\left(\frac{M}{m}\right)A,\end{aligned} \tag{83}$$

$$\begin{aligned}(0 \le)\, S(A|B) - \frac{\ln m}{M-m}(MA-B) - \frac{\ln M}{M-m}(B-mA) \\ \le \frac{4}{(M-m)^2}\left(K\left(\frac{M}{m}\right) - 1\right)(B-mA)A^{-1}(MA-B)\end{aligned} \tag{84}$$

and

$$\begin{aligned}&\frac{1}{2M^2}(B-mA)A^{-1}(MA-B) \\ &\le S(A|B) - \frac{\ln m}{M-m}(MA-B) - \frac{\ln M}{M-m}(B-mA)\end{aligned} \tag{85}$$

$$\leq \frac{1}{2m^2}(B-mA)A^{-1}(MA-B)$$

for positive invertible operators A and B that satisfy the condition (13).

Observe that, if $A,\ B \in \mathcal{B}_1(H)$ with $\operatorname{tr}(A) = \operatorname{tr}(B) = 1$ and satisfy (13), then we must assume $m \leq 1 \leq M$ and by trace properties we have

$$\begin{aligned}\operatorname{tr}\left[(B-mA)A^{-1}(MA-B)\right] &= \operatorname{tr}\left[(m+M)B - mMA - BA^{-1}B\right]\\ &= m+M-mM-\operatorname{tr}\left(A^{-1}B^2\right)\\ &= (M-1)(1-m) - \chi^2(B,A),\end{aligned}$$

where $\chi^2(B,A) =: \operatorname{tr}\left(A^{-1}B^2\right) - 1 \geq 0$.

We also have

$$\frac{\ln m}{M-m}(M-1) + \frac{\ln M}{M-m}(1-m) = \ln\left(m^{\frac{M-1}{M-m}} M^{\frac{1-m}{M-m}}\right).$$

We can state the following result:

Proposition 1 (Dragomir, 2015 [5]) *Let $A,\ B \in \mathcal{B}_1(H)$ with $\operatorname{tr}(A) = \operatorname{tr}(B) = 1$ that satisfy (13) for some $m,\ M$ with $0 < m < 1 < M$. Then we have the inequalities*

$$(0 \leq) \operatorname{tr} S(A|B) - \ln\left(m^{\frac{M-1}{M-m}} M^{\frac{1-m}{M-m}}\right) \leq \ln S\left(\frac{M}{m}\right), \tag{86}$$

$$\begin{aligned}&(0 \leq) \operatorname{tr} S(A|B) - \ln\left(m^{\frac{M-1}{M-m}} M^{\frac{1-m}{M-m}}\right) \\ &\leq \frac{4}{(M-m)^2}\left(K\left(\frac{M}{m}\right) - 1\right)\left[(M-1)(1-m) - \chi^2(B,A)\right]\end{aligned} \tag{87}$$

and

$$\begin{aligned}&\frac{1}{2M^2}\left[(M-1)(1-m) - \chi^2(B,A)\right]\\ &\leq \operatorname{tr} S(A|B) - \ln\left(m^{\frac{M-1}{M-m}} M^{\frac{1-m}{M-m}}\right)\\ &\leq \frac{1}{2m^2}\left[(M-1)(1-m) - \chi^2(B,A)\right].\end{aligned} \tag{88}$$

Observe that

$$\frac{m\ln m}{M-m}(M-1) + \frac{M\ln M}{M-m}(1-m) = \ln\left(m^{\frac{m(M-1)}{M-m}} M^{\frac{M(1-m)}{M-m}}\right),$$

then by taking the trace in the inequalities (66) and (75) we can state the following result as well:

Proposition 2 (Dragomir, 2015 [5]) *Let* $A, B \in \mathcal{B}_1(H)$ *with* $\operatorname{tr}(A) = \operatorname{tr}(B) = 1$ *that satisfy (13) for some* m, M *with* $0 < m < 1 < M$*. Then we have the inequalities*

$$\begin{aligned}(0 \leq) \ln\left(m^{\frac{m(M-1)}{M-m}} M^{\frac{M(1-m)}{M-m}}\right) + \operatorname{tr} S(B|A) \\ \leq \frac{\ln M - \ln m}{M - m}\left[(M-1)(1-m) - \chi^2(B,A)\right]\end{aligned} \tag{89}$$

and

$$\begin{aligned}&\frac{1}{2M}\left[(M-1)(1-m) - \chi^2(B,A)\right] \\ &\leq \ln\left(m^{\frac{m(M-1)}{M-m}} M^{\frac{M(1-m)}{M-m}}\right) + \operatorname{tr} S(B|A) \\ &\leq \frac{1}{2m}\left[(M-1)(1-m) - \chi^2(B,A)\right].\end{aligned} \tag{90}$$

References

1. S.S. Dragomir, Bounds for the normalized Jensen functional. Bull. Aust. Math. Soc. **74**(3), 417–478 (2006)
2. S.S. Dragomir, A note on Young's inequality. Rev. R. Acad. Cienc. Exactas Fís. Nat. Ser. A Math. doi:10.1007/s13398-016-0300-8, http://rgmia.org/papers/v18/v18a126.pdf
3. S.S. Dragomir, A note on new refinements and reverses of Young's inequality. Trans. J. Math. Mech. **8**(1), 45–49 (2016) [http://rgmia.org/papers/v18/v18a131.pdf]
4. S.S. Dragomir, Some inequalities for relative operator entropy, Preprint RGMIA Res. Rep. Coll. **18** (2015), Art. 145. [http://rgmia.org/papers/v18/v18a145.pdf]
5. S.S. Dragomir, Further inequalities for relative operator entropy. Preprint RGMIA Res. Rep. Coll. **18** (2015), Art. 160. [http://rgmia.org/papers/v18/v18a160.pdf]
6. J.I. Fujii, E. Kamei, Uhlmann's interpolational method for operator means. Math. Jpn. **34**(4), 541–547 (1989)
7. J.I. Fujii, E. Kamei, Relative operator entropy in noncommutative information theory. Math. Jpn. **34**(3), 341–348 (1989)
8. S. Furuichi, On refined Young inequalities and reverse inequalities. J. Math. Inequal. **5**, 21–31 (2011)
9. S. Furuichi, Refined Young inequalities with Specht's ratio. J. Egypt. Math. Soc. **20**, 46–49 (2012)
10. S. Furuichi, Precise estimates of bounds on relative operator entropies. Math. Inequal. Appl. **18**, 869–877 (2015)
11. S. Furuichi, N. Minculete, Alternative reverse inequalities for Young's inequality. J. Math Inequal. **5**(4), (2011) 595–600
12. T. Furuta, J.M. Hot, J. Pečarić, Y. Seo, *Mond-Pečarić Method in Operator Inequalities. Inequalities for Bounded Selfadjoint Operators on a Hilbert space*. Monographs in Inequalities, vol. 1 (Element, Zagreb, 2005), pp. xiv+262, pp.+loose errata. ISBN: 953-197-572-8

13. I.H. Kim, Operator extension of strong subadditivity of entropy. J. Math. Phys. **53**, 122204 (2012)
14. F. Kittaneh, Y. Manasrah, Improved Young and Heinz inequalities for matrix. J. Math. Anal. Appl. **361**, 262–269 (2010)
15. F. Kittaneh, Y. Manasrah, Reverse Young and Heinz inequalities for matrices. Lin. Multilin. Alg. **59**, 1031–1037 (2011)
16. P. Kluza, M. Niezgoda, Inequalities for relative operator entropies. Electron. J. Lin. Alg. **27**, 851–864 (2014). Art. 1066
17. F. Kubo, T. Ando, Means of positive operators. Math. Ann. **264**, 205–224 (1980)
18. W. Liao, J. Wu, J. Zhao, New versions of reverse Young and Heinz mean inequalities with the Kantorovich constant. Taiwan. J. Math. **19**(2), 467–479 (2015)
19. G.V. Milovanović, M.T. Rassias, *Analytic Number Theory, Approximation Theory, and Special Functions* (Springer, New York, 2014)
20. M.S. Moslehian, F. Mirzapour, A. Morassaei, Operator entropy inequalities. Colloq. Math. **130**, 159–168 (2013)
21. M. Nakamura, H. Umegaki, A note on the entropy for operator algebras. Proc. Jpn. Acad. **37**, 149–154 (1961)
22. I. Nikoufar, On operator inequalities of some relative operator entropies. Adv. Math. **259**, 376–383 (2014)
23. W. Specht, Zer Theorie der elementaren Mittel. Math. Z. **74**, 91–98 (1960)
24. M. Tominaga, Specht's ratio in the Young inequality. Sci. Math. Jpn. **55**, 583–588 (2002)
25. A. Uhlmann, Relative entropy and the Wigner-Yanase-Dyson-Lieb concavity in an interpolation theory. Commun. Math. Phys. **54**(1), 21–32 (1977)
26. G. Zuo, G. Shi, M. Fujii, Refined Young inequality with Kantorovich constant. J. Math. Inequal. **5**, 551–556 (2011)

On the Use of Elliptic Regularity Theory for the Numerical Solution of Variational Problems

Axel Dreves, Joachim Gwinner, and Nina Ovcharova

Abstract In this article we show the crucial role of elliptic regularity theory for the development of efficient numerical methods for the solution of some variational problems. Here we focus on a class of elliptic multiobjective optimal control problems that can be formulated as jointly convex generalized Nash equilibrium problems (GNEPs) and on nonsmooth boundary value problems that stem from contact mechanics leading to elliptic variational inequalities (VIs).

Keywords Complementarity problem • Dual mixed formulation • Elliptic boundary value problem • Jointly convex generalized Nash equilibrium problem • Lagrange multiplier • Multiobjective optimal control • Normalized Nash equilibrium • Obstacle problem • Saddle point formulation • Signorini problem • Smooth domain • Unilateral contact • Variational inequality

AMS 90C29, 90C33, 49J21, 49N60

Introduction

As noted in the survey paper [44], elliptic regularity theory is of essential importance for the derivation of error estimates of the finite element method (FEM) for the numerical solution of nonsmooth boundary value problems formulated as variational inequalities. This is now well documented in the literature starting from the pioneering work of Falk [18]. More recent examples of this research direction are the paper [39] on the *h*-FEM treatment of unilateral crack problems and other nonsmooth constraints and the paper [25] on *hp*-FEM convergence for unilateral contact problems with Tresca friction in plane linear elastostatics.

A. Dreves • J. Gwinner • N. Ovcharova (✉)
Department of Aerospace Engineering, Universität der Bundeswehr München,
München, Germany
e-mail: axel.dreves@unibw.de; joachim.gwinner@unibw.de; nina.ovcharova@unibw.de

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_11

This article is concerned with other applications of elliptic regularity theory. First we consider a class of elliptic multiobjective optimal control problems formulated as jointly convex generalized Nash equilibrium problems. As will be detailed below, a rather straightforward variational formulation of such a problem leads to a generalized Nash equilibrium problem (GNEP), where however each player has to satisfy different constraints that depend on the control of the other players. Thus one obtains more involved quasi-variational inequalities, in contrast to variational inequalities that can be obtained when considering normalized solutions of jointly convex GNEPs as is shown in the recent paper [17], based on the regularity of the underlying elliptic boundary value problem.

Then we turn to Signorini mixed boundary value problems, unilateral frictionless contact problems and other nonsmooth boundary value problems that can be formulated as variational inequalities with a coercive bilinear form. To get rid of relatively complicated constraints as e.g., inequality constraints and to obtain simpler nonnegativity constraints or box constraints one can introduce Lagrange multipliers similar as in constrained optimization in finite dimensions. In addition to simplification for better numerical treatment, there is also an intrinsic interest in Lagrange multipliers as dual variables. Often in applications they have a clear physical meaning and are more of interest than the primal variables; speaking in the language of continuum mechanics, the engineer is often more interested in the stresses and strains than in the displacements. This motivates multifield variational formulations and multiple saddle point problem formulations, see [19, 26, 27]. While for linear elliptic boundary value problems the passage from the primal variational formulation to a dual mixed formulation or a saddle point problem form involving a Lagrange multiplier is a standard procedure and while there are the well-established mixed finite element methods [4, 11] for their numerical treatment, such a procedure for nonsmoothly constrained problems has to overcome several difficulties. First, the standard approach to existence of Lagrange multipliers for inequality constrained optimization in infinite dimensional spaces relies on the Hahn–Banach separation theorem and needs an interior point condition (Slater condition) with respect to the ordering cone in the image space. However, the topological interior of such an ordering cone in standard function (Hilbert or Banach) spaces, as e.g. the interior of the cone L^p_+ of non-negative L^p functions is empty. So one may resort to the nonempty quasirelative interior of L^p_+ and one may impose a Slater-like condition, that is, the existence of a feasible point that lies in the quasirelative interior of L^p_+. However, as a counterexample of Daniele and Giuffrè [13] shows, this condition is not sufficient, and extra more complicated assumptions or related involved conditions that are actually equivalent are needed to ensure the existence of a Lagrange multiplier, see [9, 13–15].

Therefore we proceed in another way and show how by a simple formula one obtains a Lagrange multiplier in the dual of the preimage space thus even reducing the variational inequality to a complementarity problem. By this simple approach, the Lagrange multiplier lives in the dual of the Sobolev space of the variational problem, thus at first, is a general measure which may be singular. Here regularity theory comes into play to conclude that the Lagrange multiplier is indeed an

L^p function. Thus from an inequality constraint, one finally obtains a Lagrange multiplier in the cone L^p_+ of non-negative L^p functions. This approach works also with not necessarily symmetric bilinear forms, when the equivalence to convex quadratic optimization is lost; it even works for nonlinear operators. Moreover, we can combine such dual mixed formulations for variational problems with inequality constraints via non-negative Lagrange multipliers with mixed formulations for variational inequalities of second kind where the Lagrange multiplier is in a simple box set. This applies to unilateral contact problems with Tresca friction.

The outline of this article is as follows. The next section provides a review of elliptic regularity theory dealing with the linear Dirichlet problem, the scalar unilateral boundary value problem (obstacle problem), and frictionless unilateral contact of linear elastostatics. In section "From Elliptic Multiobjective Optimal Control to Jointly Convex Generalized Nash Equilibria" we consider a class of elliptic multiobjective optimal control problems and show following [17] how based on elliptic regularity theory, these problems can be reformulated as so-called jointly convex generalized Nash equilibria. In section "Lagrange Multipliers, Convex Duality Theory, and Mixed Formulations of Nonsmooth Variational Problems" we present a direct approach to mixed formulations of some nonsmooth variational problems and of associated variational inequalities. The article ends with some conclusions and an outlook to some open problems.

A Review of Elliptic Regularity Theory

In this section we review the elliptic regularity theory that is needed for the understanding of the subsequent sections.

Regularity of Linear Scalar Dirichlet Problem

In this subsection we are concerned with the regularity of the solution of the Dirichlet problem with a (scalar) linear second-order elliptic operator L; that is, the operator L is of the form (summation convention employed)

$$Lu = D_i(-a^{ij}(x)\, D_j u) + a(x)u,$$

where the coefficients a^{ij} $(i,j = 1,\dots,d)$ and a are assumed to be bounded, measurable functions on a domain $\Omega \subset \mathbb{R}^d$ and moreover, a is non-negative and there exists a positive number α such that

$$a^{ij}(x)\, \xi_i \xi_j \geq \alpha|\xi|^2, \forall x \in \Omega, \xi \in \mathbb{R}^d.$$

A simple example is $L = -\Delta$, the negative Laplacian on $\mathbb{R}^d$; on the other hand, lower order terms involving $D_i u$ can easily be included in the definition of L. The operator L above gives rise to the bilinear form

$$\mathscr{L}(u,v) = \int_{\Omega} [a^{ij}(x) D_j u D_i v + a(x)uv]\, dx.$$

Let in addition f be (locally) integrable on Ω and φ belong to $H^1(\Omega)$, the Sobolev space of all L^2 functions on Ω with weak L^2 derivatives, see [1]. Then a function $u \in H^1(\Omega)$ is called a weak solution of the Dirichlet problem:

$$Lu = f, \quad u = \varphi \text{ on } \partial\Omega,$$

if $u - \varphi \in H_0^1(\Omega)$ and u satisfies

$$\mathscr{L}(u,v) = \int_{\Omega} fv\, dx,\ \forall v \in C_0^1(\Omega).$$

The following example of a domain with a reentrant corner taken from the book of Braess [10] shows that even for smooth data f, φ we cannot expect the solution to be in $H^2(\Omega)$, not to mention in $C^2(\Omega)$, what is suggested by a classic treatment of partial differential equations.

Example 1 Let

$$\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 \ :\ |x| < 1, x_1 < 0 \text{ or } x_2 > 0\}.$$

Identify $\mathbb{R}^2$ with $\mathbb{C}$. Let $z = x_1 + ix_2 = \rho \exp(i\theta)$ and consider

$$w(z) = z^{2/3}; \quad u(x) = \operatorname{Im} w(z) = \rho^{2/3} \sin\left(\frac{2}{3}\theta\right).$$

So u is harmonic and $u \in H^1(\Omega)$ solves

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega; \\ u(\exp(i\theta)) &= \sin\left(\tfrac{2}{3}\theta\right) && \text{for } 0 \le \theta \le \tfrac{3}{2}\pi, \\ u &= 0 && \text{elsewhere on } \partial\Omega. \end{aligned}$$

Since $w'(z) = \frac{2}{3} z^{-\frac{1}{3}}$, even the first derivatives of u are not bounded for $z \to 0$.

There are two options for a domain to obtain regularity $u \in H^2(\Omega)$: smoothness of the boundary $\partial\Omega$ or convexity of the domain. For the first let us recall from the monograph of Gilbarg and Trudinger [20, Theorem 8.12]

Theorem 1 *Suppose that $\partial\Omega$ is of class C^2. Moreover assume the coefficients a^{ij} are uniformly Lipschitz continuous in Ω and for the data assume $f \in L^2(\Omega)$ and*

$\varphi \in H^2(\Omega)$ such that $u - \varphi \in H^1_0(\Omega)$ with a weak solution u of the above Dirichlet problem. Then also $u \in H^2(\Omega)$.

For such a regularity result and for its direct proof we can also refer to the monograph of Aubin [3, Chap. 7, Sect. 1-8, Theorem 1-1] and to the monograph of Kinderlehrer and Stampacchia [38, Chap. IV, Appendix A].

Regularity results for the Dirichlet problem for elliptic operators, respectively, for the Laplacian on convex domains and on more general so-called semiconvex domains (here a bounded domain is semiconvex, if for any $x \in \partial\Omega$ there exists an open ball $B_x \subset \mathbb{R}^d \setminus \bar{\Omega}$ with $\bar{B}_x \cap \bar{\Omega} = \{x\}$) are established in the work of Kadlec [35] and of Mitrea et al. [42]. Let us also mention the regularity results for solutions of the equations of linear elasticity in convex plane polygonal domains by Bacuta and Bramble [5].

Regularity of the Scalar Unilateral Boundary Value Problem

Let us turn to the regularity of scalar unilateral boundary value problems, in particular of Signorini boundary value problems. We are also concerned with the regularity of domain obstacle problems, since domain obstacle and boundary obstacle (Signorini) problems are related as follows.

Let Γ_D, Γ_S be two disjoint smooth and open subset of $\partial\Omega$ such that $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_S$. Let A be a linear elliptic operator defined by $Au = -D_j(a^{ij} D_i u)$ with coefficients a^{ij} as above giving the bilinear form

$$a(u, v) = \int_\Omega a^{ij}(x)\, D_i u(x)\, D_j v(x)\, dx.$$

Let $\psi \in H^1(\Omega)$ with $\psi \leq 0$ on Γ_S, let $\tilde{\psi}$ be the unique solution of the Dirichlet problem

$$\Delta\tilde{\psi} = f \text{ in } \Omega, \qquad \tilde{\psi} = \psi \text{ on } \partial\Omega$$

and assume that $\tilde{\psi} \in H^2(\Omega)$. Let

$$V = \{v \in H^1(\Omega) :\; v = 0 \text{ on } \Gamma_D\}$$

and define the closed convex subsets of V:

$$\begin{aligned} K &= \{v \in V :\; v \geq \psi \text{ on } \Gamma_S\}, \\ \tilde{K} &= \{v \in V :\; v \geq \tilde{\psi} \text{ in } \Omega, v = \psi \text{ on } \Gamma_S\}. \end{aligned}$$

Then there holds the following

Theorem 2 *If u is the solution of the VI (domain obstacle problem)*

$$u \in \tilde{K}, \quad a(u, v-u) \geq \int_\Omega f(v-u)\, dx \quad \forall v \in \tilde{K},$$

then u resolves the VI (Signorini problem)

$$u \in K, \quad a(u, v-u) \geq \int_\Omega f(v-u)\, dx \quad \forall v \in K.$$

For its proof see the proof of Theorem 9.3 in [38, Chap. IV]. In virtue of Theorem 2 we can conclude from the regularity result [38, Chap. IV, Theorem 2.3] for the domain obstacle problem the following regularity result for the Signorini problem with $A = -\Delta$:

Theorem 3 *Suppose $f \in L^s(\Omega)$ and $\max(-\Delta\tilde{\psi} - f, 0) \in L^s(\Omega)$ for some $s > d$. Then the solution of the above Signorini problem with $A = -\Delta$ lies in $H^{2,s}(\Omega) \cap C^{1,\lambda}(\bar{\Omega}), \lambda = 1 - (d/s)$. Hence, $\Delta u \in L^s(\Omega)$.*

There is a refinement concerning the regularity of the domain obstacle problem at the boundary by Jensen [33]. He has proven the local regularity $W^{2,\infty}$ of the solution at boundary points. However, Kinderlehrer [37] has provided the following example of a scalar Signorini problem with a solution that fails to be in $H^2(\Omega)$. Here even the boundary obstacle is zero, but $\partial\Gamma_D \cap \partial\Gamma_S \neq \emptyset$ with the Dirichlet part Γ_D and the Signorini part Γ_S.

Example 2 Let

$$\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 \mid |x| < 1, x_2 > 0\}$$

with the mutually disjoint, open boundary parts

$$\begin{aligned} \Gamma_S &= \{x = (x_1, 0) \mid -1 < x_1 < 0\}, \\ \Gamma_N &= \{x = (x_1, 0) \mid 0 < x_1 < 1\}, \\ \Gamma_D &= \{x = (x_1, x_2) \mid |x| = 1, x_2 > 0\}. \end{aligned}$$

Let $z = x_1 + ix_2 = \rho \exp(i\theta)$ and consider

$$u(x) = -\operatorname{Re} z^{1/2} = -\rho^{1/2} \cos(\theta/2).$$

So u is harmonic and $u(x) = 0$ for $x_1 < 0$, $x_2 = 0$. By the Cauchy–Riemann differential equations,

$$\frac{\partial}{\partial \nu}\, u(x_1, 0) = -\frac{\partial}{\partial x_2}\, u(x_1, 0) = -\frac{\partial}{\partial x_1}\, \operatorname{Im} z^{1/2} = \begin{cases} 0 & \text{if } x_1 > 0,\, x_2 = 0 \\ \dfrac{1}{2}|x_1|^{-\frac{1}{2}} & \text{if } x_1 < 0,\, x_2 = 0. \end{cases}$$

Hence, $u \in H^1(\Omega)$ satisfies $-\Delta u = 0$ in Ω and the Neumann, respectively, Dirichlet boundary conditions

$$\frac{\partial u}{\partial \nu} = 0 \text{ on } \Gamma_N,\ u = -\cos\frac{\theta}{2} \text{ on } \Gamma_D,$$

and the Signorini boundary conditions

$$u\,\frac{\partial u}{\partial \nu} = 0,\ u \geq 0,\ \frac{\partial u}{\partial \nu} \geq 0 \text{ on } \Gamma_S.$$

Thus u solves the VI

$$u \in K, \qquad \int_\Omega \nabla u \cdot \nabla(v-u)\,dx \geq 0,\ \forall v \in K,$$

where

$$K = \{v \in H^1(\Omega) \mid v \geq 0 \text{ on } \Gamma_S,\ v = -\cos\frac{\theta}{2} \text{ on } \Gamma_D\}.$$

Note that $\dfrac{\partial^2}{\partial x_1^2} = \cos^2\theta\,\dfrac{\partial^2}{\partial \rho^2} + \ldots,\ \dfrac{\partial^2}{\partial x_2^2} = \sin^2\theta\,\dfrac{\partial^2}{\partial \rho^2} + \ldots,$

$$\iint_\Omega |u_{\rho\rho}|^2\,dx = (1/16)\int_0^\pi\int_0^1 \rho^{-3}\,\rho\,\cos^2(\theta/2)d\rho\,d\theta; \int_0^1 \rho^{-2}\,d\rho = \infty$$

so u cannot lie in $H^2(\Omega)$.

Variational Formulation of Frictionless Unilateral Contact Problem of Linear Elastostatics

Before we continue our review of elliptic regularity theory addressing frictionless unilateral contact problems we introduce some notation from continuum mechanics and describe the variational form of unilateral contact problems as variational inequalities (of first kind, following the terminology of [21]).

Let us assume Hooke's law and small deformations of a non-homogeneous, anisotropic body. For notational simplicity we focus on the case of plane elasticity; the three-dimensional case poses no additional difficulty in deriving the variational formulation. So let $\Omega \subset \mathbb{R}^2$ be a bounded plane domain with Lipschitz boundary Γ ($\Gamma \in C^{0,1}$), occupied by an elastic body, and let $\underline{x} = (x_1, x_2)$ be a Cartesian coordinate system. Then $\underline{n} = (n_1, n_2)$, the unit outward normal to Γ, exists almost everywhere and $\underline{n} \in [L^\infty(\Gamma)]^2$, see, e.g., [36, Theorem 5.4].

With the displacement vector $\underline{v} = (v_1, v_2)$ to lie in the Sobolev space $[H^1(\Omega)]^2$ the linearized strains are given by

$$\varepsilon_{ij}(v) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) \quad (i,j = 1,2) \tag{1}$$

and Hooke's law relating strains and stresses reads

$$\tau_{ij} = E_{ijkl}\ \varepsilon_{kl} \quad (i,j = 1,2)\,, \tag{2}$$

where we use the summation convention over a repeated index within the range 1, 2 and where the elasticity coefficients $E_{ijkl} \in L^\infty(\Omega)$ satisfy

$$\begin{gathered} E_{ijkl} = E_{klij} = E_{jikl}\,; \\ \exists c_0 > 0 :\ E_{ijkl}\ \varepsilon_{ij}\ \varepsilon_{kl}\ \geq\ c_0\,\varepsilon_{ij}\ \varepsilon_{ij} \quad \forall\, \varepsilon_{ij} = \varepsilon_{ji}\,. \end{gathered} \tag{3}$$

With the given vector $\underline{F} = (F_1, F_2) \in [L^2(\Omega)]^2$ the stress field has to satisfy the equilibrium equations

$$\frac{\partial \tau_{ij}}{\partial x_j} + F_i = 0 \quad (i = 1,2)\,. \tag{4}$$

The traction vector $\underline{b}$ on the boundary, where

$$b_i = \tau_{ij}\, n_j$$

can be decomposed into the normal component

$$b_n = b_i\, n_i = \tau_{ij}\, n_i\, n_i$$

and the tangential component

$$b_t = b_i\, t_i = \tau_{ij}\, t_i\, n_j\,,$$

where $\underline{t} = (t_1, t_2) = (-n_2, n_1)$ is the unit tangential vector. Likewise the displacement v can be decomposed (see [36, Chap. 5], [16] for the relevant trace theorems):

$$v_n = v_i n_i\,,\ \ v_t = v_i t_i\,.$$

To describe the boundary conditions, let $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_S$, where the open parts Γ_D, Γ_N, and Γ_S are mutually disjoint. Eventually nonzero displacements $\underline{D} \in [H_1(\Gamma_D)]^2$, respectively, tractions $\underline{T} \in [L_2(\Gamma_N)]^2$ are prescribed on Γ_D, resp. Γ_N, i.e.,

$$v_i = D_i \qquad \text{on} \quad \Gamma_D\,, \tag{5}$$

$$b_i = T_i \qquad \text{on} \quad \Gamma_N\,, \tag{6}$$

whereas on Γ_S the frictionless unilateral contact conditions (Signorini's conditions for v_n and b_n)

$$v_n \leq g,\ b_n \leq 0,\ (v_n - g)\, b_n = 0,\ b_t = 0 \tag{7}$$

with a given gap function $g \in L_2(\Gamma_S)$ are imposed. To make the contact problem meaningful we assume $meas(\Gamma_S) > 0$. Here we also require $meas(\Gamma_D) > 0$, hence rigid body motions are excluded and the variational problem becomes coercive.

Now the problem (1), (2), (4)–(7) can be formulated as the following variational inequality (VI): Find $u \in K$ such that

$$\beta(u, v-u) \geq \lambda(v-u) \quad \forall v \in K\,, \tag{8}$$

where we introduce the bilinear form, respectively the linear form

$$\beta(\underline{v}, \underline{w}) = \int_\Omega E_{ijkl}\, \varepsilon_{ij}(\underline{v})\, \varepsilon_{kl}(\underline{w})\, dx\,,$$

$$\lambda(\underline{v}) = \int_\Omega F_i v_i\, dx + \int_{\Gamma_N} T_i v_i\, ds$$

on the function space

$$V = \left\{\underline{v} \in [H^1(\Omega)]^2 \mid \underline{v} = 0 \text{ on } \Gamma_D\right\}$$

and the convex closed subset

$$K = \{\underline{v} \in V \mid v_i = D_i \text{ on } \Gamma_D;\ v_n \leq g \text{ on } \Gamma_S\}\ .$$

One may reduce the inhomogeneous inequality constraint $v_n = v_i n_i \leq g$ to the homogeneous inequality constraint $\tilde{v}_n = \tilde{v}_i n_i \leq 0$, thus simplify to a convex cone constraint by subtraction of some appropriate extension $\underline{g}$ of $g \in L_2(\Gamma_S)$ to $[H^1(\Omega)]^2$. However, this simple reduction for unilateral constraints does not work with more general bilateral constraints of the form $g_a \leq v_n = v_i n_i \leq g_b$, when the extended real-valued boundary obstacles have domains that intersect, i.e., dom $g_a \cap$ dom $g_b \neq \emptyset$, in particular in a three-dimensional situation.

Regularity of Frictionless Unilateral Contact Problem of Linear Elastostatics

In view of his example given above Kinderlehrer [37] could prove by a difference quotient technique that the solution u of the Signorini problem is in H^2 except perhaps near points of $\partial\Gamma_S \cup \partial\Gamma_N \cup \partial\Gamma_D \subset \Omega$, more precisely the following result for the $d-$dimensional mixed Signorini boundary value problem in the case $g = 0$, what is by the remark above, no loss of generality concerning regularity.

Theorem 4 ([37, Theorem 2.2]) *Suppose for the data* $\underline{F} \in [L^2(\Omega)]^d$, $\underline{T} \in [H^1(\Gamma_D)]^d$, $\underline{D} \in [H^2(\Gamma_D)]^d$. *Set* $\Omega_\delta = \{x \in \Omega \mid dist\ (x, \partial\Gamma_S \cup \partial\Gamma_N \cup \partial\Gamma_D) > \delta\}$ *for* $\delta > 0$. *Then for each* $\delta > 0$ *there hold* $u \in (H^2(\Omega_\delta))^d$ *and the Signorini conditions (7) pointwise a.e. on* Γ_S.

Sobolev imbedding of H^2 in spaces of Hölder continuous functions implies the

Corollary 1 *Under the assumptions of the data as in the above theorem, there holds*

$$for\ d = 2,\ u \in [C^{0,\alpha}(\bar{\Omega}_\delta)]^2\, for\ \ some\ 0 < \alpha < 1,$$
$$and\ \ for\ d = 3,\ u \in [C^{0,\frac{1}{2}}(\bar{\Omega}_\delta)]^2.$$

By the theory of pseudodifferential operators Schumann [47] extended the latter result to $[C^{1,\alpha}(\Omega \cup \Gamma)]^2$ regularity of the solution u; the precise value of α is not known. To conclude this section, we refer to the survey [48] of Schumann who gives an excellent overview of the mathematical methods to prove regularity results for variational inequalities and unilateral problems in elasticity.

From Elliptic Multiobjective Optimal Control to Jointly Convex Generalized Nash Equilibria

In this section we consider a class of elliptic multiobjective optimal control problems and show following [17] how based on elliptic regularity theory, these problems can be reformulated as so-called jointly convex generalized Nash equilibria.

The Concept of Jointly Convex Generalized Nash Equilibrium Problems

Let $V_\nu, \nu = 1, \ldots, N$ be real separable Hilbert spaces or more general reflexive, separable Banach spaces endowed with norms $\|\cdot\|_\nu$, and define $V := V_1 \times \ldots \times V_N$. Further, let X be a nonempty, closed, and convex subset of V and assume that the objective functions $\theta_\nu : V_1 \times \ldots \times V_N \to \mathbb{R}$, $\theta_\nu(\cdot, x^{-\nu}) : V_\nu \to \mathbb{R}$ are convex for any fixed $x^{-\nu}$, where we use the notation $x = (x^1, \ldots, x^N) = (x^\nu, x^{-\nu})$ to emphasize the role of the variable x^ν, but this notation does not mean a permutation. In this setting the infinite dimensional jointly convex generalized Nash equilibrium problem (GNEP for short) has the following form

$$\min_{x^\nu} \theta_\nu(x^\nu, x^{-\nu}) \quad \text{subject to (s.t.)} \quad (x^\nu, x^{-\nu}) \in X \tag{9}$$

for all $\nu = 1, \ldots, N$. The reason for calling this problem jointly convex is that the strategies must belong to a common convex set X, instead of each player having his own strategy set $X_\nu(x^{-\nu})$ depending on the rivals' strategy $x^{-\nu}$. We call $\bar{x}$ a *generalized Nash equilibrium*, if $\bar{x} \in X$ satisfies

$$\theta_\nu(\bar{x}^\nu, \bar{x}^{-\nu}) \le \theta_\nu(x^\nu, \bar{x}^{-\nu}), \quad \forall (x^\nu, \bar{x}^{-\nu}) \in X$$

for all $\nu = 1, \ldots, N$. Note that the concept of GNEPs goes back to the 1954 paper [2] of Arrow and Debreu.

Next let us introduce the Nikaido–Isoda function

$$\Psi(x, y) := \sum_{\nu=1}^{N} [\theta_\nu(x^\nu, x^{-\nu}) - \theta_\nu(y^\nu, x^{-\nu})],$$

see the 1955 paper [43] of Nikaido and Isoda, to define normalized solutions of a jointly convex GNEP. The point $\bar{x} \in X$ is called a *normalized Nash equilibrium*, or a *normalized solution of the jointly convex GNEP* if

$$\Psi(\bar{x}, y) \le 0, \quad \forall y \in X.$$

Thus we get a characterization of some solutions, namely the normalized solutions of jointly convex GNEPs via a variational inequality in contrast to more involved quasi-variational inequalities that characterize the solutions of GNEPs in general form, necessarily jointly convex. Therefore, computing normalized solutions of jointly-convex GNEPs is typically much easier than obtaining solutions of GNEPs in general form. Since for a normalized Nash equilibrium $\bar{x}$ we have for all $\nu = 1, \ldots, N$ and all $(y^\nu, \bar{x}^{-\nu}) \in X$

$$\theta_\nu(\bar{x}^\nu, \bar{x}^{-\nu}) - \theta_\nu(y^\nu, \bar{x}^{-\nu}) = \Psi(\bar{x}, (y^\nu, \bar{x}^{-\nu})) \le \sup_{y \in X} \Psi(\bar{x}, y) \le 0,$$

every normalized solution is also a generalized Nash equilibrium, i.e., for all $\nu = 1, \ldots, N$ it holds that

$$\theta_\nu(\bar{x}^\nu, \bar{x}^{-\nu}) \le \theta_\nu(y^\nu, \bar{x}^{-\nu}) \qquad \forall (y^\nu, \bar{x}^{-\nu}) \in X;$$

however, the converse is not true.

Primal Formulation of Elliptic Multiobjective Optimal Control Problems

Let $\Omega \subset \mathbb{R}^d\,(d = 2, 3)$ be a bounded Lipschitz domain and let $V := H_0^1(\Omega)$ denote the Sobolev space of all L^2 functions on Ω with weak L^2 derivatives and zero boundary values. Let $U^\nu := L^2(\Omega)$ be the space for the controls u^ν for all $\nu = 1, \dots, N$. We have the weights $\gamma_\nu > 0, \beta_\nu > 0$, the given data $f, g^\nu \in L^2(\Omega)$, $(\nu = 1, \dots, N)$; $a^\nu, b^\nu \in \mathbb{R}$ with $a^\nu \leq b^\nu$ $(\nu = 1, \dots, N)$, $a^0, b^0 \in H^1(\Omega)$ with $a^0(x) < b^0(x)$ and some continuous, compact, and linear operators $\chi^\nu : V \to L^2(\Omega)$ $(\nu = 1, \dots, N)$. Then we consider the following problem

$$
\begin{aligned}
(I)\ \min_{y,u^\nu} \ & \frac{1}{2}\,\|\chi^\nu y - g^\nu\|^2_{L^2(\Omega)} + \frac{\gamma_\nu}{2}\,\|u^\nu\|^2_{L^2(\Omega)} \\
\text{s.t.}\ \ & Ly = \sum_{\mu=1}^{N} \beta_\mu u^\mu + f, \quad y|_{\partial\Omega} = 0, \\
& a^0(x) \leq y(x) \leq b^0(x), \text{a.e. in } \Omega, \\
& a^\nu \leq u^\nu(x) \leq b^\nu, \qquad \text{a.e. in } \Omega,
\end{aligned}
$$

for all $\nu = 1, \dots, N$. In this problem every player ν minimizes his own cost function through his individual control variable u^ν and the common state variable y. The state is determined by the controls of all players via a partial differential equation (pde) given by a linear elliptic partial differential operator L of second order as introduced in the previous section.

To provide a functional analytic meaning we can write the above pde constraint in variational form as

$$
y \in V : \quad \mathscr{L}(y, w) = \langle \sum_{\mu=1}^{N} \beta_\mu u^\mu + f, w\rangle_{L^2(\Omega)}, \quad \forall w \in V\,.
$$

Using the continuous embedding from $H_0^1(\Omega)$ in $L^2(\Omega)$, the state constraints $a^0 \leq y \leq b^0$ are to be understood in the L^2 sense as the control constraints $a^\nu \leq u^\nu \leq b^\nu$, which imply that the controls are actually L^∞ functions, since $a^\nu, b^\nu \in \mathbb{R}$.

The elliptic multiobjective optimal control problem (I) is, however, not a GNEP, since the state y is a common optimization variable for all players. If we introduce different state variables y^ν for each player $\nu \in \{1, \dots, N\}$, and if we could guarantee that all the states are equal, we get a GNEP. But we do not get a jointly convex GNEP, since the players then have different constraints $Ly^\nu = \sum_{\mu=1}^{N} \beta_\mu u^\mu + f$, depending on the controls of the other players. For the numerical solution of these GNEPs in general form one can use its optimality conditions that are equivalent to quasi-variational inequalities, and are much harder to solve than VIs. Also the number of algorithms for the solution of quasi-variational inequalities is rather limited. Therefore our next aim is to develop a jointly convex reformulation.

Reduced Multicontrol Formulation of the Elliptic Multiobjective Optimal Control Problems

Since by the Lax–Milgram theorem $L : H_0^1(\Omega) \to H^{-1}(\Omega)$ is an isomorphism, we can use the inverse $L^{-1} : H^{-1}(\Omega) \to H_0^1(\Omega)$ to define the multicontrol to state map

$$S(u) := L^{-1}\left(\sum_{\mu=1}^{N} \beta_\mu u^\mu + f\right),$$

and this is a continuous map affine linearly dependent on $(u^1, \ldots, u^N)$. Since $L^2(\Omega)$ is compactly embedded in $H^{-1}(\Omega)$, see [1], this is even a completely continuous map from $[L^2(\Omega)]^N$ to $H_0^1(\Omega)$. Hence, we obtain the equivalent reduced problem

$$\begin{aligned} (II) \min_{u^\nu} \frac{1}{2} \|\chi^\nu S(u^\nu, u^{-\nu}) - g^\nu\|_{L^2(\Omega)}^2 + \frac{\gamma_\nu}{2} \|u^\nu\|_{L^2(\Omega)}^2 & \\ \text{s.t. } a^0(x) \le S(u^\nu, u^{-\nu})(x) \le b^0(x), & \text{ a.e. in } \Omega, \\ a^\nu \le u^\nu(x) \le b^\nu, & \text{ a.e. in } \Omega, \end{aligned}$$

for all $\nu = 1, \ldots, N$, which is a jointly convex GNEP. A similar problem was first considered in [30] as a GNEP and using a penalty approach and a strict uniform feasible response assumption, the existence of a solution was shown. Further, using the Nikaido–Isoda function, this reformulation (II) was used in [31] to show existence of a Nash equilibrium for the equivalent problem (I). Moreover it was shown that one can solve these reformulations (II) (even for parabolic and not only elliptic pdes) via a primal–dual path-following method based on the Nikaido–Isoda function.

Let us stress that (II) is already a jointly convex GNEP. However, every evaluation of $S(u^\nu, u^{-\nu})$ requires the solution of the pde. To avoid this we give a third equivalent formulation to (I).

A Multistate Formulation of the Elliptic Multiobjective Optimal Control Problems

Now we assume that $\partial\Omega$ is of class C^2 or Ω is convex. Then regularity theory for elliptic equations with Dirichlet boundary conditions, as exposed in the previous section, guarantees that the solution $S(u)$ is even in $H^2(\Omega)$. Therefore,

$$w^\nu := L^{-1}(\beta_\nu u^\nu) \in H_0^1(\Omega) \cap H^2(\Omega) =: W$$

for all $\nu = 1, \dots, N$. These w^ν will become our new optimization variables. Indeed we now have

$$y = S(u) = L^{-1}f + \sum_{\mu=1}^{N} w^\mu, \tag{10}$$

and, since $Lw^\nu \in L^2(\Omega)$, the equation

$$u^\nu = -\frac{1}{\beta_\nu} Lw^\nu \tag{11}$$

holds in $L^2(\Omega)$ for all $\nu = 1, \dots, N$. Thus we arrive at the equivalent problem

$$\begin{aligned}
(III)\ \min_{w^\nu \in W} \frac{1}{2} & \left\| \chi^\nu \left(\sum_{\mu=1}^{N} w^\mu \right) + \chi^\nu L^{-1} f - g^\nu \right\|^2_{L^2(\Omega)} + \frac{\gamma_\nu}{2\beta_\nu^2} \left\| Lw^\nu \right\|^2_{L^2(\Omega)} \\
\text{s.t.}\ \ a^0(x) \le & \left(L^{-1}f + \sum_{\mu=1}^{N} w^\mu \right)(x) \le b^0(x), \quad \text{a.e. in } \Omega, \\
a^\nu \beta_\nu \le & \left(Lw^\nu \right)(x) \le b^\nu \beta_\nu, \qquad \text{a.e. in } \Omega,
\end{aligned}$$

for all $\nu = 1, \dots, N$. Now, defining the common feasible set

$$\begin{aligned}
\widetilde{W} := \Bigg\{ (w^1, \dots, w^N) \in W^N \ \Bigg| \ & a^\nu \beta_\nu \le \left(Lw^\nu \right)(x) \le b^\nu \beta_\nu \ (\forall \nu = 1, \dots, N), \\
& a^0(x) \le \left(L^{-1}f + \sum_{\mu=1}^{N} w^\mu \right)(x) \le b^0(x) \text{ a.e. in } \Omega \Bigg\},
\end{aligned}$$

and the cost functions

$$\theta_\nu(w^\nu, w^{-\nu}) := \frac{1}{2} \left\| \chi^\nu \left(\sum_{\mu=1}^{N} w^\mu \right) + \chi^\nu L^{-1} f - g^\nu \right\|^2_{L^2(\Omega)} + \frac{\gamma_\nu}{2\beta_\nu^2} \left\| Lw^\nu \right\|^2_{L^2(\Omega)},$$

our elliptic multiobjective optimal control problem in the novel formulation (*III*) writes as the jointly convex GNEP:

$$\min_{w^\nu} \theta_\nu(w^\nu, w^{-\nu}) \quad \text{s.t.} \quad (w^\nu, w^{-\nu}) \in \widetilde{W}$$

for all $\nu = 1, \dots, N$. Solving this jointly convex GNEP gives us $(w^1, \dots w^N)$ from which we can easily compute the state variable y via (10) and the controls $(u^1, \dots, u^N)$ via (11), thus gaining the complete solution of our original problem (I). It was demonstrated in [17] that one can solve this reformulation (III) using a relaxation method that computes a best-response function and performs a line search exploiting a merit function, again based on the Nikaido–Isoda function.

Lagrange Multipliers, Convex Duality Theory, and Mixed Formulations of Nonsmooth Variational Problems

In this section we provide mixed formulations of some nonsmooth variational problems and of associated variational inequalities. To achieve this goal we pursue a direct relatively simple approach to Lagrange multipliers that, however, heavily hinges on elliptic regularity theory. To put this approach in perspective we first shortly review the standard approach to Lagrange multipliers in convex duality theory that is based on the Hahn–Banach separation theorem.

A Short Review of Convex Infinite Dimensional Duality Theory in Function Spaces

The standard approach to prove existence of Lagrange multipliers for inequality constrained optimization problems in infinite dimensional spaces is based on the Hahn–Banach separation theorem and thus needs interior point conditions, in particular a nonvoid interior of the ordering cone associated with the inequality constraint. In function spaces of continuous functions endowed with the maximum norm with applications, e.g., to Chebychev approximation one can work with the topological interior of the ordering cone, see, e.g., [32]. However, the cone of non-negative L^p functions and hence the ordering cone in the Sobolev spaces—relevant for pde constrained optimization—have empty topological interior. To overcome this difficulty one can resort to the concept of the so-called quasi-relative interior of a convex set introduced by Borwein and Lewis [8]. Therefore next we give the definition of this concept and a short review of corresponding recent results on Lagrangian duality.

Let C be a nonvoid subset of a real normed space X. Let cl C, co C, cone C denote the topological closure, convex hull, conical hull of C, respectively. Then for a given point $x \in C$, the set

$$T_C(x) = \{y \in X : y = \lim_{n\to\infty} t_n(x_n - x),\ t_n > 0,\ x_n \in C\ (\forall n \in \mathbb{N}),\ \lim_{n\to\infty} x_n = x\}$$

is called the *tangent cone (contingent cone)* to C at x. If C is convex, then $T_C(x) =$ cl cone $(C - x)$. With the dual space X^* and the duality form $(.,.)$, the *normal cone* to C at $x \in C$ is defined by

$$N_C(x) = \{x^* \in X^* : (x^*, y - x) \leq 0, \; \forall y \in C\}.$$

Now the *quasi-interior* of a convex subset C of X is the set

$$\text{qi } C = \{x \in C \; : \; \text{cl cone } (C - x) = X\}$$

and there holds the characterization, see [14], for x in the convex set C:

$$x \in \text{qi } C \Leftrightarrow N_C(x) = \{\, 0_{X^*}\}.$$

Due to Borwein and Lewis [8] is the following refinement of the notion of the quasi-interior: The *quasirelative interior* of a convex subset C of X is the set

$$\text{qri } C = \{x \in C \; : \; \text{cl cone } (C - x) \text{ is a linear subspace of } X\}$$

and there holds the characterization, see [14], for x in the convex set C:

$$Tx \in \text{qri } C \Leftrightarrow N_C(x) \text{ is a linear subspace of } X^*.$$

These are useful concepts in L^p function spaces and thus in Sobolev spaces as shown by the following example.

Example 3 Consider the Banach space $X = L^2(T, \mu)$ with $1 \leq p < \infty$ on a measure space (T, μ) and the closed convex cone $C = \{z \in X : z(t) \geq 0 \; \mu\text{–a.e.}\}$. Then the characteristic function of T, $1 = 1_T$ lies in qi C, hence in qri C. Indeed, by Lebesgue's theorem of majorized convergence, any $x \in X$ can be approximated by the sequence $\{x_n\}$ of truncations,

$$x_n(t) = \begin{cases} x(t) \text{ if } x(t) \geq -n \; a.e.; \\ -n \;\; \text{elsewhere}, \end{cases}$$

and clearly $x_n \in n(C - 1)$.

Now let us turn to inequality constrained convex optimization and Lagrangian duality theory. Consider the following primal optimization problem:

$$\text{(P)} \quad \inf_{x \in R} f(x),$$

where

$$R = \{x \in S : g(x) \in -C\},$$

is assumed to be nonempty and S a nonempty subset of X; Y is another normed space partially ordered by a convex cone C; $f : S \to \mathbb{R}$ and $g : S \to Y$ are two maps such that the map $(f, g) : S \to \mathbb{R} \times Y$, defined by $(f, g)(x) = (f(x), g(x)), \forall x \in S$ is convex-like with respect to the cone $\mathbb{R}_+ \times C \subset \mathbb{R} \times Y$, that is the set $(f, g)S + \mathbb{R}_+ \times C$ is convex. Then the Lagrangian is

$$L(x, \ell) = f(x) + (\ell, g(x)), \quad x \in S, \ell \in C^*$$

and the Lagrange dual problem to (P) reads

$$\text{(D)} \quad \sup_{\ell \in C^*} \inf_{x \in S} [f(x) + (\ell, g(x))],$$

where $C^* = \{x^* \in X^* : (x^*, x) \geq 0, \forall x \in C\}$ is the dual cone to C. While for the optimal values of (P) and (D), $\inf \text{(P)} = \inf_x \sup_\ell L(x, \ell) \geq \sup_\ell \inf_x L(x, \ell) = \sup \text{(D)}$ trivially holds, one is interested in the equality of these optimal values and moreover in the existence of a *Lagrange multiplier*, that is, an optimal solution ℓ in (D). This is called *strong duality*.

In the favorable situation when the topological interior of the ordering cone, int C, is not empty, the approach to strong duality in infinite dimensions via the Hahn–Banach separation theorem requires the easily verifiable Slater condition as a constraint qualification (see the important paper of Jeyakumar and Wolkowicz [34]), that is, the existence of a feasible point $\tilde{x} \in R$ such that $g(\tilde{x}) \in -\text{int } C$.

Thus one may be inclined to transfer this approach to the situation when the topological interior of C is empty by replacing "int" by "qri." However, this fails, as the following example due to Daniele and Giuffrè [13] shows.

Example 4 Let $X = S = Y = l^2$, the Hilbert space of all real sequences $x = (x_n)_{n \in \mathbb{N}}$ with $\sum_{n=1}^{\infty} x_n^2 < \infty$ and $C = l_+^2$ the cone of all non-negative sequences in l^2. Define $f : l^2 \to \mathbb{R}$ and $g : l^2 \to l^2$, respectively, by

$$f(x) = \sum_{n=1}^{\infty} \frac{x_n}{n}, \quad (g(x))_n = -\frac{x_n}{2^n}, \quad \forall n \in \mathbb{N}.$$

Then the feasible set $T = \{x \in l^2 \mid -g(x) \in l_+^2\} = l_+^2$. One has $\text{cl}(l_+^2 - l_+^2) = l^2, l_{++}^2 := \text{qri } l_+^2 = \{x \in l^2 \ : \ x_n > 0, \forall n \in \mathbb{N}\} \neq \emptyset$. Take $\tilde{x} \in l_+^2, \tilde{x}_n = \frac{1}{n}$, then $-(g(\tilde{x}))_n = \frac{1}{n2^n}, -g(\tilde{x}) \in l_{++}^2$. Further, $\inf \text{(P)} = 0$ and $x = 0_{l^2}$ is the optimal solution of (P). On the other hand, for $\ell \in l_+^2$ we have

$$\begin{aligned}\inf_{x \in l^2} [f(x) + (\ell, g(x))] &= \inf_{x \in l^2} \left[\sum_{n=1}^{\infty} \frac{x_n}{n} - \sum_{n=1}^{\infty} \ell_n \frac{x_n}{2^n} \right] \\ &= \inf_{x \in l^2} \sum_{n=1}^{\infty} \left[\frac{1}{n} - \frac{\ell_n}{2^n} \right] x_n\end{aligned}$$

$$= \begin{cases} 0 & \text{if } \ell_n = \dfrac{2^n}{n} \ \forall n \in \mathbb{N}, \\ -\infty & \text{otherwise.} \end{cases}$$

However, ℓ with $\ell_n = \frac{2^n}{n}$ does not belong to l^2. Hence, $\sup(\mathrm{D}) = -\infty$ and the optimal values do not coincide.

This example can also be given in a function space using the well-known isometry of l^2 and $L^2(0, 2\pi)$ based on Fourier expansion.

So in addition to a qri Slater-like condition one needs extra conditions to ensure strong duality. To this aim Boţ, Csetnek, and Moldovan [9] introduce the following conic extension of (P) in the image space:

$$\begin{aligned} \mathscr{E}_{\inf(\mathrm{P})} &= \{(\inf(\mathrm{P}) - f(x) - r, -g(x) - y) : x \in S, r \geq 0, y \in C\} \\ &= (\inf(\mathrm{P}), 0_Y) - (f, g)\, S - \mathbb{R}_+ \times C, \end{aligned}$$

where as in classic convex duality theory only $\inf(\mathrm{P}) \in \mathbb{R}$ is required, but not the existence of an optimal solution to (P). Note that by feasibility of (P), $R \neq \emptyset$ implies $\inf(\mathrm{P}) < \infty$ and in the case $\inf(\mathrm{P}) = -\infty$ strong duality trivially holds.

In this way Boţ, Csetnek, and Moldovan [9] could prove the following strong duality result.

Theorem 5 ([9, Theorem 4.1]) *Suppose that $cl(C - C) = Y$ and there exists some $\tilde{x} \in S$ such that $g(\tilde{x}) \in -qri\, C$. If*

$$(0, 0_Y) \notin qri\ co[\mathscr{E}_{\inf(P)} \cup \{(0, 0_Y)\}], \tag{12}$$

then strong duality holds.

A Direct Approach to Lagrange Multipliers and Dual Mixed Formulations of Inequality Constrained Optimization and of VIs of the First Kind

We start with convex quadratic optimization in infinite dimensional spaces. Let V be a real Hilbert space and let Q be another real Hilbert space (for simplicity identified with its dual Q'). Let $A \in \mathscr{L}(V, V')$ with $A = A', A \geq 0$ (i.e., $\langle Av, v\rangle \geq 0, \forall v \in V$). Further, let $B \in \mathscr{L}(V, Q)$ and let $f \in V', g \in Q$ be fixed elements. Moreover let an order $\leq$ defined in Q via a convex closed cone $C \subset Q$ via $q \geq 0$, iff $q \in C$. With these data consider the convex quadratic optimization problem

$$(CQP) \qquad \begin{cases} \text{minimize}\, f(v) = \frac{1}{2}\langle Av, v\rangle - \langle f, v\rangle \\ \text{subject to}\ Bv \leq g\,. \end{cases}$$

This gives rise to the bilinear form $a(u, v) := \langle Au, v\rangle$ and the convex closed set

$$K(g) := \{v \in V \mid Bv \leq g\},$$

which is translated from the cone

$$K_0 := \{v \in V \mid Bv \leq 0\}.$$

As is well known, a solution u of (CQP) is characterized by the following VI of the first kind—following the terminology in [21]:

$$(VI - 1) \qquad u \in K(g),\ a(u, v - u) \geq \langle f, v - u\rangle,\ \forall v \in K(g).$$

Here we present a simple approach—different from the approach reviewed above—to Lagrange multipliers. Assume that there exists a preimage of g under B, $B\tilde{g} = g$. This allows to work with the duality on $V \times V'$, obtain readily the existence of a Lagrange multiplier in the dual cone

$$K_0^+ = \{\kappa \in V' \ :\ \langle \kappa, w\rangle \geq 0,\ \forall w \in K_0\}$$

and arrive at the following characterization.

Proposition 1 *Let $u \in K(g)$. Then u solves the above $(VI - 1)$, iff there exists $\lambda \in K_0^+$ such that $(u, \lambda) \in V \times V'$ solves the mixed system*

$$(MP - 1) \qquad \begin{cases} a(u, v) = \langle \lambda, v\rangle + \langle f, v\rangle \\ \langle \mu - \lambda, u - \tilde{g}\rangle \geq 0\,, \end{cases}$$

for all $v \in V, \mu \in K_0^+$. Further, there holds the complementarity condition

$$\langle \lambda, u - \tilde{g}\rangle = 0.$$

Proof Let $u \in K(g)$ solve the $(VI - 1)$: $a(u, v - u) \geq \langle f, v - u\rangle$, $\forall v \in K(g)$. Define $\lambda \in V'$ by $\lambda(v) = a(u, v) - f(v)$. Then $(MP - 1)_1$ holds. Further, for any $v \in K_0$, $\tilde{v} := v + u$ lies in $K(g)$ and hence,

$$\lambda(v) = a(u, \tilde{v} - u) - f(\tilde{v} - u) \geq 0.$$

Thus $\lambda \in K_0^+$. Since $\tilde{g} \in K(g), u - \tilde{g} \in K_0$,

$$\langle \mu - \lambda, u - \tilde{g}\rangle = \langle \mu, u - \tilde{g}\rangle - [a(u, u - \tilde{g}) - f(u - \tilde{g})] \geq 0$$

for any $\mu \in K_0^+$ and therefore $(MP - 1)$ holds. The complementarity condition follows from $(MP - 1)_2$ by the choices $\mu = 2\lambda$, $\mu = 0$.

Vice versa, let $v \in K(g)$, hence $v - \tilde{g} \in K_0$. This implies by the complementarity condition

$$\langle \lambda, v - u \rangle = \langle \lambda, v - \tilde{g} \rangle - \langle \lambda, u - \tilde{g} \rangle \geq 0.$$

Hence, we arrive at

$$a(u, v - u) = (f + \lambda)(v - u) \geq f(v - u).$$

□

By the proof above it is clear that $(MP - 1)$ is equivalent to the following complementarity problem: Find $(u, \lambda) \in V \times V'$ such that

$$(CP - 1) \qquad \begin{cases} \lambda = Au - f \\ \lambda \in K_0^+, u - \tilde{g} \in K_0 \\ \langle \lambda, u - \tilde{g} \rangle = 0\,. \end{cases}$$

Moreover the proof shows that the characterization above holds also with not necessarily symmetric bilinear forms, when the equivalence to convex quadratic optimization is lost; it even holds for nonlinear operators A mapping a Banach space V to its dual V'.

This approach applies to domain obstacle problems, where the linear map B is the imbedding map [1], say from $H^1(\Omega)$ to $L^2(\Omega)$ for linear scalar elliptic operators L or more generally from $W^{m,p}(\Omega)$ to some $L^q(\Omega)$. It also applies to boundary obstacle problems or unilateral contact problems with the Signorini condition on some boundary part Γ_c in appropriate function spaces, where the linear map B is the trace map γ [16] to the boundary part Γ_c. By this simple approach, the Lagrange multiplier lives in the dual of the Sobolev space of the variational problem, thus at first, is a general measure which may be singular. Here regularity theory—see the review in the second section of this paper—comes into play to conclude that the Lagrange multiplier is indeed a L^p function. Thus from an inequality constraint, one finally obtains a Lagrange multiplier λ in the cone L^p_+ of non-negative L^p functions on the domain Ω. Thus we obtain the recent result [15, Theorem 3.3] of Daniele, Giuffrè, Maugeri, and Raciti. When in the (scalar) mixed Signorini problem with a linear elliptic pde, there exists a multiplier ℓ to the inequality constraint $\gamma v \leq g \Leftrightarrow v|\Gamma_c \leq g$ a.e. that lives in the dual Q' to the image space $Q = L^2(\Gamma_c)$, thus lies in $L^2_+(\Gamma_c)$, then the multipliers ℓ and λ are related by $\lambda = \gamma^* \ell$, where γ^* denotes the adjoint of the trace map $\gamma : H^1(\Omega) \to L^2(\Gamma_c)$.

Indeed, this direct simple approach to Lagrange multipliers and mixed formulations is used in an efficient numerical treatment of domain obstacle problems. Based on such mixed formulations the very effective biorthogonal basis functions with local support, due to Lamichhane and Wohlmuth [41], can be employed for approximation of the Lagrange multipliers in the hp-adaptive FEM for elliptic obstacle problems, see the recent paper [6] of Banz and Schröder.

A Direct Approach to Lagrange Multipliers for VIs of Second Kind

Here we consider nonsmooth optimization problems of the form

$$(NOP) \qquad \min_{v\in V} f(v) = \frac{1}{2}\langle Av, v\rangle - \langle f, v\rangle + \varphi(v),$$

where φ is convex, even positively homogeneous on V, but not differentiable in the classic sense. A prominent example encountered with given friction or Tresca friction in solid mechanics is

$$\varphi_g(v) = \int_{\Gamma_c} g|v|\, ds \quad (g \in L^\infty(\Gamma_c), g > 0)\,.$$

An optimal solution of (NOP) is characterized as a solution of the VI of the second kind:

$$(VI-2) \qquad u \in V, \qquad \langle Au, v-u\rangle + \varphi(v) - \varphi(u) \geq f(v-u),\ \forall v \in V.$$

For the above example of φ_g use

$$\varphi_g(v) = \int\limits_{\Gamma} g\,|v|\, d\Gamma = \sup\left\{ \int\limits_{\Gamma} g\ v\ \mu\ \ d\Gamma \ \middle|\ \mu \in L^2(\Gamma), |\mu| \leq 1 \right\},$$

where sup is attained by $\mu = \text{sign}\, v$, set

$$M := \{\mu \in L^2(\Gamma), |\mu| \leq 1\}$$

and arrive—as it is shown in more general terms in Proposition 2 below—at the mixed problem:

Find $u \in V = H^1(\Omega), \lambda \in M$ such that for all $v \in V, \mu \in M$

$$\begin{cases} \langle Au, v\rangle + \int\limits_{\Gamma} g\ v\ \lambda\ d\Gamma &= \langle f, v\rangle, \\ \int\limits_{\Gamma} g\ u\ (\lambda - \mu)\ d\Gamma &\geq 0. \end{cases}$$

To reveal the duality structure, introduce

$$M(g) = \{\mu \in L^2(\Gamma), |\mu| \leq g \text{ a.e.}\}.$$

Although, with $g \in L^\infty(\Gamma)$, this set is clearly contained in $L^\infty(\Gamma)$, we stick to the easier treatable L^2 duality. Thus $(VI-2)$ is equivalent—as it is shown in more general terms in Proposition 2 below—to the mixed problem:

Find $u \in V = H^1(\Omega), \lambda \in M(g)$ such that for all $v \in V, \mu \in M(g)$

$$\begin{cases} \langle Au, v\rangle_{V^* \times V} + \langle \lambda, v\rangle_{L^2(\Gamma)} = \langle f, v\rangle_{V^* \times V}, \\ \langle u, \lambda - \mu\rangle_{L^2(\Gamma)} \qquad\qquad \geq 0. \end{cases}$$

Indeed, in the more general setting of a reflexive Banach space V , a map $A : V \to V^*$,$f \in V^*$ and a sublinear functional $\varphi : V \to \mathbb{R}$, we have the following result using the convex weakly $*$-compact subdifferential

$$P := \partial\varphi(0) = \{q \in V^*, \langle q, v\rangle \leq \varphi(v), \forall v \in V\}.$$

Proposition 2 *$u \in V$ solves the above $(VI-2)$, iff there exists $p \in P$ such that $(u, p) \in V \times V^*$ solves the mixed system*

$$(MP-2) \qquad \begin{cases} \langle Au, v\rangle + \langle p, v\rangle = \langle f, v\rangle \\ \langle p - q, u\rangle \geq 0\ , \end{cases}$$

for all $v \in V, q \in P$.

Proof Let $u \in V$ solve $(VI-2)$. Then the choice $v = 0$ gives

$$\langle Au, u\rangle + \varphi(u) \leq f(u), \tag{13}$$

whereas the choice $v = tw, w \in V, t > 0, t \to \infty$ gives for all $w \in V$,

$$\langle Au, w\rangle + \varphi(w) \geq f(w), \tag{14}$$

hence, from (13) and (14) we get

$$\langle Au, u\rangle + \varphi(u) = f(u). \tag{15}$$

Note that (15) and (14) imply $(VI-2)$, hence these assertions are equivalent to $(VI-2)$.

Define $p \in V^*$ by $p = f - Au$. Then $(MP-2)_1$ trivially holds. Further, from (14), for any $w \in V$, $\varphi(v) \geq \langle p, w\rangle$, hence p lies in $\partial\varphi(0) = P$. Finally from (15), $\varphi(u) = \langle p, u\rangle$, hence $(MP-2)_2$ follows.

Vice versa, $(MP-2)_2$ implies $\varphi(u) = \langle p, u\rangle$, hence together with $(MP-2)_1$ and the choice $v = u$ gives (15). Since $\varphi(v) \geq \langle p, v\rangle$, from $(MP-2)_1$ we arrive at (14). □

Similarly as discussed in the previous subsection, the regularity of the multiplier p hinges on the regularity of the datum f and in particular on the regularity of the solution u of the $(VI-2)$ via the map A.

To apply the above general result to the friction-type functional φ_g we only have to set $V = H^1(\Omega), \varphi = \varphi_g \circ \gamma$ with the linear continuous trace operator γ that maps $H^1(\Omega)$ onto $H^{\frac{1}{2}}(\Gamma)$ dense in $L^2(\Gamma)$ and use the subdifferential chain rule [46]

$$\partial\varphi = \partial(\varphi_g \circ \gamma) = \gamma^* \partial\varphi_g \gamma.$$

Note that this chain rule holds as an equality, since φ_g is real-valued and so the constraint qualification $0 \in \text{int}\ (\text{range}\ \gamma - \text{dom}\ \varphi)$ is trivially satisfied.

To conclude this subsection let us mention other duality relations and mixed formulations useful in numerical treatment of variational inequalities of the second kind. By (L^1, L^∞) duality and density one obtains

$$\varphi_g(v) = \int_{\Gamma_c} g|v|\, ds = \sup\left\{ \int_{\Gamma_c} g\ v\ \mu\ ds \,\middle|\, \mu \in \mathbf{C}(\Gamma), |\mu| \le 1 \right\}.$$

This is used in convergence proof of Finite Element Methods and Boundary Element Methods, see [24, 25].

Another way to cope with the nondifferentiable functional φ_g is to decompose the modulus function $|\rho| = \rho^+ + \rho^-$ with the positive part $\rho^+ = \max(\rho, 0) \ge 0$ and the negative part $\rho^- = \max(-\rho, 0) \ge 0$. This leads to inequality constrained problems considered in the previous subsection what is not elaborated here further.

A Direct Approach to Lagrange Multipliers for More General VIs

To conclude this section we deal with the more general

$$(VI-3) \qquad u \in K, \qquad \langle A(u), v-u\rangle + \varphi(v) - \varphi(u) \ge f(v-u),\ \forall v \in K,$$

where as above $f \in V^*$, $\varphi : V \to \mathbb{R}$ is sublinear and now $K \subset V$ is a convex closed cone with vertex at zero. A VI of this form occurs in unilateral contact of a linear elastic body with a rigid foundation under the Tresca friction law, if the initial gap between body and foundation is zero, see [25]. The more general setting, in particular for non-zero gap, with K and φ convex would encompass also the VIs of first kind studied before, but needs additional arguments. Therefore we prefer this simpler homogeneous setting to elucidate the direct approach to Lagrange multipliers. In this setting we have the following result in a general locally convex topological vector space V for a not necessarily linear operator $A : V \to V^*$.

Theorem 6 *Let $u \in K$. Then u solves the above $(VI-3)$, iff there exist $p \in P = \partial\varphi(0)$ and $\lambda \in K^-$ such that the complementarity condition $\langle \lambda, u\rangle = 0$ holds and $(u, p, \lambda) \in V \times V^* \times V^*$ solves the mixed system*

$$(MP-3) \qquad \begin{cases} \langle A(u), v\rangle + \langle p+\lambda, v\rangle = \langle f, v\rangle \\ \langle p-q, u\rangle \geq 0\,, \end{cases}$$

for all $v \in V, q \in P$.

Proof Let $u \in K$ solve $(VI-3)$. Then we first proceed as in the proof of Proposition 2. The choice $v = 0$ gives

$$\langle A(u), u\rangle + \varphi(u) \leq f(u), \tag{16}$$

whereas the choice $v = tw, w \in K, t > 0, t \to \infty$ gives for all $w \in K$,

$$\langle A(u), w\rangle + \varphi(w) \geq f(w), \tag{17}$$

hence from (16) and (17) we get

$$\langle A(u), u\rangle + \varphi(u) = f(u). \tag{18}$$

Note that (18) and (17) imply $(VI-3)$, hence these assertions are equivalent to $(VI-3)$.

Define $\ell \in V^*$ by $\ell = f - A(u)$. From (17) we find

$$\varphi(v) \geq \langle \ell, w\rangle\,, \quad \forall w \in K. \tag{19}$$

Now we claim that $\ell \in K^- + P$ and that hence, $(MP-3)_1$ holds. Note that both K^- and P are convex closed sets, moreover P is weakly* compact in V^*. So the claim can be shown by an indirect argument employing the separation theorem. Here we use that

$$\varphi(w) = \max_{q\in P}\langle q, w\rangle$$

and thus (19) means that for any $w \in K$ there exists $q \in P$ such that $\langle q, w\rangle \geq \langle \ell, w\rangle$. Therefore by the extension lemma [23, Theorem 2.2] (which is a refined version of the famous Fan–Glicksberg–Hoffman theorem of alternative and is proved from a fixed point theorem or from the separation theorem) there exists $p \in P$ such that $\langle q, w\rangle \geq \langle \ell, w\rangle$ holds for all $w \in K$. Now define $\lambda = \ell - p$, hence $\lambda \in K^-$ and $\ell = \lambda + p \in K^- + P$ as claimed.

From (18) we obtain

$$\langle p, u\rangle \leq \varphi(u) = \langle \lambda + p, u\rangle,$$

hence, $\langle \lambda, u \rangle \geq 0$. Since $u \in K, \lambda \in K^{-}$, the complementarity condition $\langle \lambda, u \rangle = 0$ follows. This gives with (18) that $\varphi(u) = \langle p, u \rangle$, hence we arrive at $(MP-3)_2$.

Vice versa, $(MP-3)_1$ implies together with the complementarity condition and $(MP-3)_2$

$$\langle f - A(u), u \rangle = \langle \lambda + p, u \rangle = \langle p, u \rangle = \varphi(u) \,,$$

hence (18). In view of $\lambda \in K^{-}$ we conclude from $(MP-3)_1$ that for any $w \in K$,

$$\langle f - A(u), w \rangle = \langle \lambda + p, w \rangle \leq \langle p, w \rangle \leq \varphi(w) \,,$$

hence (17). □

Conclusions and Outlook

We have seen the crucial role of elliptic regularity theory in two instances. First with elliptic multiobjective optimal control formulated as jointly convex GNEP the regularity of the solution of the underlying pde was needed to arrive at a reformulation that was the basis for an efficient numerical solution method. In this approach we had to require that the domain where the elliptic pde lives is convex or sufficiently smooth. On the other hand, real-world domains may have reentrant corners or are only piecewise smooth. This leads to the question how this approach can be refined using the well-known elliptic theory in nonsmooth domains [22], abandoning classic Sobolev spaces, and working instead with weighted Sobolev spaces [40].

Then we have presented a direct approach to Lagrange multipliers in inequality constrained and related nonsmooth boundary value problems which gives an immediate link between the regularity of the Lagrange multiplier and the regularity of the solution of the problem. As already the one-dimensional obstacle problem demonstrates, there is a threshold of smoothness, however, that in general cannot be overstepped even if the data are arbitrarily smooth. The regularity theory for frictionless unilateral contact reported from the work [37] has shown the influence of the switching points, where the boundary conditions change, on the smoothness of the solution. So one may be interested in a more detailed analysis in weighted Sobolev spaces [40] that takes the switching points in account.

Finally let us point out that we have here considered frictionless monotone unilateral contact problems. Nonmonotone contact problems can be put in primal form as hemivariational inequalities (HVIs). While the theory of HVIs is well developed, the numerical solution of these problems is in its infancy; here, we can refer to [7, 12, 28, 29, 45, 49] (ordered according to publication date). So one may ask for mixed formulations with appropriate Lagrange multipliers that would allow the development of mixed finite element procedures for the efficient solution of these nonconvex variational problems.

References

1. R.A. Adams, J.J.F. Fournier, *Sobolev Spaces* (Elsevier, Amsterdam, 2003)
2. K.J. Arrow, G. Debreu, Existence of an equilibrium for a competitive economy. Econometrica **22**, 265–290 (1954)
3. J.-P. Aubin, *Approximation of Elliptic Boundary-Value Problems*. Pure and Applied Mathematics, vol. XXVI (Wiley-Interscience, New York, 1972)
4. I. Babuška, G.N. Gatica, On the mixed finite element method with Lagrange multipliers. Numer. Methods Partial Differ. Equ. **19**(2), 192–210 (2003)
5. C. Bacuta, J.H. Bramble, Regularity estimates for solutions of the equations of linear elasticity in convex plane polygonal domains. Z. Angew. Math. Phys. **54**(5), 874–878 (2003)
6. L. Banz, A. Schröder, Biorthogonal basis functions in hp-adaptive FEM for elliptic obstacle problems. Comput. Math. Appl. **70**(8), 1721–1742 (2015)
7. M. Barboteu, K. Bartosz, P. Kalita, An analytical and numerical approach to a bilateral contact problem with nonmonotone friction. Int. J. Appl. Math. Comput. Sci. **23**(2), 263–276 (2013)
8. J.M. Borwein, A.S. Lewis, Partially finite convex programming, Part I: quasi relative interiors and duality theory. Math. Programm. **57**(1–3), 15–48 (1992)
9. R.I. Boţ, E.R. Csetnek, A. Moldovan, Revisiting some duality theorems via the quasirelative interior in convex optimization. J. Optim. Theory Appl. **139**(1), 67–84 (2008)
10. D. Braess, *Finite Elements* (Cambridge University Press, Cambridge, 2007)
11. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer, New York, 1991)
12. J. Czepiel, P. Kalita, Numerical solution of a variational-hemivariational inequality modelling simplified adhesion of an elastic body. IMA J. Numer. Anal. **35**(1), 372–393 (2015)
13. P. Daniele, S. Giuffrè, General infinite dimensional duality and applications to evolutionary network equilibrium problems. Optim. Lett. **1**(3), 227–243 (2007)
14. P. Daniele, S. Giuffrè, G. Idone, A. Maugeri, Infinite dimensional duality and applications. Math. Ann. **339**(1), 221–239 (2007)
15. P. Daniele, S. Giuffrè, A. Maugeri, F. Raciti, Duality theory and applications to unilateral problems. J. Optim. Theory Appl. **162**(3), 718–734 (2014)
16. Z. Ding, A proof of the trace theorem of Sobolev spaces on Lipschitz domains. Proc. Am. Math. Soc. **124**(2), 591–600 (1996)
17. A. Dreves, J. Gwinner, Jointly convex generalized Nash equilibria and elliptic multiobjective optimal control. J. Optim. Theory Appl. **168**(3), 1065–1086 (2016)
18. R.S. Falk, Error estimates for the approximation of a class of variational inequalities. Math. Comput. **28**, 963–971 (1974)
19. G.N. Gatica, N. Heuer, S. Meddahi, On the numerical analysis of nonlinear twofold saddle point problems. IMA J. Numer. Anal. **23**(2), 301–330 (2003)
20. D. Gilbarg, N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics (Springer, Berlin, 2001)
21. R. Glowinski, *Numerical Methods for Nonlinear Variational Problems* (Springer, Berlin, 2008)
22. P. Grisvard, *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics, vol. 69 (SIAM, Philadelphia, PA, 2011)
23. J. Gwinner, An extension lemma and homogeneous programming. J. Optim. Theory Appl. **47**(3), 321–336 (1985)
24. J. Gwinner, On the p-version approximation in the boundary element method for a variational inequality of the second kind modelling unilateral contact and given friction. Appl. Numer. Math. **59**(11), 2774–2784 (2009)
25. J. Gwinner, *hp*-FEM convergence for unilateral contact problems with Tresca friction in plane linear elastostatics. J. Comput. Appl. Math. **254**, 175–184 (2013)
26. J. Gwinner, Three-field modelling of nonlinear nonsmooth boundary value problems and stability of differential mixed variational inequalities. Abstr. Appl. Anal. (2013). doi:http://dx.doi.org/10.1155/2013/108043. ID 108043

27. J. Gwinner, Multi-field modeling of nonsmooth problems of continuum mechanics,differential mixed variational inequalities and their stability, in *Applied Mathematics in Tunisia*. Springer Proceedings of Mathematical Statistics, vol. 131 (Springer, Cham, 2015), pp. 119–139
28. J. Gwinner, N. Ovcharova, From solvability and approximation of variational inequalities to solution of nondifferentiable optimization problems in contact mechanics. Optimization **64**(8), 1683–1702 (2015)
29. M. Hintermüller, V.A. Kovtunenko, K. Kunisch, Obstacle problems with cohesion: a hemivariational inequality approach and its efficient numerical solution. SIAM J. Optim. **21**(2), 491–516 (2011)
30. M. Hintermüller, T. Surowiec, A PDE-constrained generalized Nash equilibrium problem with pointwise control and state constraints. Pac. J. Optim. **9**(2), 251–273 (2013)
31. M. Hintermüller, T. Surowiec, A. Kämmler, Generalized Nash equilibrium problems in banach spaces: theory, Nikaido–Isoda-based path-following methods, and applications. SIAM J. Optim. **25**(3), 1826–1856 (2015)
32. J. Jahn, *Introduction to the Theory of Nonlinear Optimization* (Springer, Berlin, 1996)
33. R. Jensen, Boundary regularity for variational inequalities. Indiana Univ. Math. J. **29**(4), 495–504 (1980)
34. V. Jeyakumar, H. Wolkowicz, Generalizations of Slater's constraint qualification for infinite convex programs. Math. Programm. Ser. B **57**(1), 85–101 (1992)
35. J. Kadlec, The regularity of the solution of the Poisson problem in a domain whose boundary is similar to that of a convex domain. Czechoslovak Math. J. **14**(89), 386–393 (1964)
36. N. Kikuchi, J.T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods* (SIAM, Philadelphia, PA, 1988)
37. D. Kinderlehrer, Remarks about Signorini's problem in linear elasticity. Ann. Scuola Norm. Sup. Pisa Cl. Sci. **8**(4), 605–645 (1981)
38. D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications* (SIAM, Philadelphia, PA, 2000)
39. D. Knees, A. Schröder, Global spatial regularity for elasticity models with cracks, contact and other nonsmooth constraints. Math. Methods Appl. Sci. **35**(15), 1859–1884 (2012)
40. A. Kufner, A.-M. Sändig, *Some Applications of Weighted Sobolev Spaces* (Teubner, Leipzig, 1987)
41. B.P. Lamichhane, B.I. Wohlmuth, Biorthogonal bases with local support and approximation properties. Math. Comput. **76**(257), 233–249 (2007)
42. D. Mitrea, M. Mitrea, L. Yan, Boundary value problems for the Laplacian in convex and semiconvex domains. J. Funct. Anal. **258**(8), 2507–2585 (2010)
43. H. Nikaidô, K. Isoda, Note on non-cooperative convex games. Pac. J. Math. **5**, 807–815 (1955)
44. M.A. Noor, K.I. Noor, T.M. Rassias, Some aspects of variational inequalities. J. Comput. Appl. Math. **47**(3), 285–312 (1993)
45. N. Ovcharova, J. Gwinner, A study of regularization techniques of nondifferentiable optimization in view of application to hemivariational inequalities. J. Optim. Theory Appl. **162**(3), 754–778 (2014)
46. R.T. Rockafellar, *Conjugate Duality and Optimization* (SIAM, Philadelphia, PA, 1974)
47. R. Schumann, Regularity for Signorini's problem in linear elasticity. Manuscripta Math. **63**(3), 255–291 (1989)
48. R. Schumann, Regularity for variational inequalities—a survey of results, in *From Convexity to Nonconvexity*. Nonconvex Optimization and its Applications, vol. 55 (Kluwer Academic Publishers, Dordrecht, 2001), pp. 269–282
49. G.E. Stavroulakis, E.S. Mistakidis, Numerical treatment of hemivariational inequalities in mechanics: two methods based on the solution of convex subproblems. Comput. Mech. **16**(6), 406–416 (1995)

Strong and Weak Convexity of Closed Sets in a Hilbert Space

Vladimir V. Goncharov and Grigorii E. Ivanov

Abstract We give a brief survey of the geometrical and topological properties of two classes of closed sets in a Hilbert space, which strengthen and weaken the convexity concept, respectively. We prove equivalence of various characterizations of these sets, which are partially new while partially known in the literature but accompanied with different proofs. Along with the uniform notions dating back to Efimov, Stechkin, Vial, Clarke, Stern, Wolenski, and others we pay attention to some local and pointwise constructions, which can be interpreted through positive and negative scalar curvatures. In the final part of the paper we give several applications to geometry of Hilbert spaces, to set-valued analysis, and to time optimal control problem.

Keywords Strong convexity • Weak convexity • Proximal smoothness • Prox-regularity • Proximal normal cone • Variational inequality • Minkowski operations • Continuous selections • Hausdorff continuity

Mathematcal Subject Classification (2010): 46C05, 49J52, 49J53, 52A01

V.V. Goncharov (✉)
CIMA, Universidade de Évora, Rua Romão Ramalho 59, 7000-671 Évora, Portugal

Institute of Systems Dynamics and Control Theory of Siberian Branch of RAS,
Lermontov str. 134, 664033 Irkutsk, Russia
e-mail: goncha@uevora.pt

G.E. Ivanov
Moscow Institute of Physics and Technology, Institutski str. 9, Dolgoprudny,
141700 Moscow Region, Russia
e-mail: g.e.ivanov@mail.ru

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_12

Introduction

Convexity is one of the fundamental concepts, being used in formulation of the basic mathematical principles as well as in numerous applications in optimization, theory of games, economics, engineering, design, and in other fields of science and technology. In particular, convexity stands as the main hypothesis in various mathematical statements, numerical algorithms, etc. However, in the second half of the twentieth century many attempts to weaken this hypothesis were made in order to extend well-known results. For instance, the concept of *paraconvexity* (see section "Weakly Convex Sets" for the exact definition) was introduced in 1958 by Michael [48] as an adequate substitution of the traditional convexity in continuous selections theorems.

On the other hand, in the same year very fruitful generalization of convexity was given by Federer (see [31]) who defined and studied properties of so named sets with positive reach in finite dimensions. The idea of such extension is based on the fact that a convex closed set $A \subset \mathbb{R}^n$ is completely characterized by the *Chebyshev property*: each point $x \in \mathbb{R}^n$ admits a unique metric projection onto A. So, Federer introduced the characteristics reach (A, a) of a closed set $A \subset \mathbb{R}^n$ as the maximal radius of a ball centered at $a \in A$ such that every point from this ball admits a unique metric projection onto A, and considered the class of sets A with reach $(A, a) > 0$, $a \in A$. In other terms this property can be expressed through a weakened variational inequality (which is equivalent to the "*quasimonotonicity*" of the normal cone) or through regularity of the distance function near the set A. For the first time the various features of the closed sets with positive reach were collected in [31, Theorem 4.8]. Let us observe that the paraconvexity by Michael and the generalized convexity considered by Federer are completely different even in $\mathbb{R}^2$. Indeed, the sets having form of the letters **V**, **X**, **Y**, or **Z** are paraconvex but have no positive reach due to angles, while the (very smooth) letter **U** is of positive reach but not paraconvex (see [31, Examples 1.1 and 1.2]). Notice that the *positive reach property* is very similar to (but does not coincide with) the so-called *exterior sphere condition* known from Differential Geometry and having a lot of applications in various fields of Partial Differential Equations, Optimal Control, etc. Roughly speaking, it means that a set can be continuously rolled outside by some ball with (fixed or variable) positive radius. Apparently, for the first time the sets (in particular, convex surfaces in $\mathbb{R}^n$) with such property were considered a little bit earlier by Reshetnyak (see [59]).

Some decades later, in the series of works by De Giorgi et al. (see, e.g., [27, 28]) a generalization of convex functions is appeared as a suitable tool to study evolution equations. So, a new class of so-called φ-convex functions was introduced. The name is due to the fact that for a φ-convex function $f(\cdot)$ the slope of its proximal gradients is controlled by a continuous real nonnegative function φ (this slope is equal to 0 whenever the function $f(\cdot)$ is convex, and the proximal subdifferential is reduced to subdifferential in the sense of Convex Analysis). Afterwards, the closed sets whose indicator functions are φ-convex, naturally called *φ-convex* (or *p-convex*) *sets*, were considered in Hilbert setting for various goals, e.g., for studying geodesics (see [16, 17]). It turned out that the notions of φ-convex sets and sets with positive reach according to Federer essentially coincide even in infinite dimensional Hilbert spaces. Namely, the equivalence of three main properties (variational inequality with a continuous function $\varphi(\cdot)$; existence, uniqueness, and continuity of the metric projection in a neighborhood of the set; differentiability of the distance function near the set) was established for the first time by Clarke et al. (see [18]) in a particular case when $\varphi \equiv$ const (equivalently, when an open neighborhood around the set admits the form of uniform tube). General case instead was treated subsequently in the works [19, 53] and others. In particular, proving that the proximal smoothness of the distance function implies φ-convexity of the set, Colombo used in [19] a nice argument involving solutions of a differential inclusion with maximally monotone operator that extends to infinite dimensions the original idea by Federer (he used Peano's Theorem for ordinary differential equations). Furthermore, it was proved that φ-convexity is equivalent to very clear geometric property: given $\bar{x} \in A$ for any $x, y \in A$ close to $\bar{x}$ a convex combination $\lambda x + (1 - \lambda) y$, $\lambda \in]0, 1[$, must be distant from the set A not more than of the order $O\left(\|x - y\|^2\right)$. Notice that the necessity of this condition was proved much earlier in [16]. Let us pay a special attention to the fact that in infinite dimensions for φ-convexity it is not enough to require the (local) existence and uniqueness of the metric projection, but this projection should be continuous. The question whether (local or global) Chebyshev property implies continuity of the projection due to authors' knowledge remains open up to now.

Summarizing everything said above we have a class of closed sets larger than the family of convex ones, which can be named by the different ways (emphasizing one of their properties): φ-convex (or p-convex), proximally convex, proximally smooth, prox-regular sets, $O(2)$-convex sets (the last term is due to Shapiro [60]), and so on. We refer the reader also to the nice survey [21] on theory and applications of prox-regular sets.

Independently, also in 1950s (like Michael and Federer) the soviet mathematicians Efimov and Stechkin had introduced in [30] another geometric concept of generalized convexity even in Banach setting. Their definition refers to the representation of a convex closed set as intersection of a family of (closed) semispaces. But in the place of semispaces they considered the complements of (open) balls of the fixed radius. Two decades later Vial defined one more geometric concept of weak convexity (see [63]) employing a generalization of the (straight line) segment joining two points. In fact, given $x, y \in A$, and a real number $R > 0$ the intersection of all closed balls of the radius R (if any), which contain both x and y, is called *R-strong* (or *R-spherical*) *segment* joining these points. In what follows we denote it by $D_R(x, y)$. So, a set A is said to be *weakly convex* (*by Vial*) if there exists $R > 0$ such that given $x, y \in A$ with $0 < \|x - y\| < 2R$ the intersection $D_R(x, y) \cap A$ contains a point different from both x and y.

Observe that the Vial's definition of weak convexity allows to formulate its strong counterpart. Namely, a set $A \subset H$ (H is a normed space, which will be assumed Hilbert everywhere in our paper if nothing said in contrary) is called *strongly convex* (*by Vial*) if there exists $R > 0$ such that given $x, y \in A$ with $\|x - y\| \leq 2R$ one has $D_R(x, y) \subset A$. In what follows we will see that this definition of strong convexity is equivalent to a lot of other geometric as well as analytic properties. For instance, one of them written in terms of variational inequality (see Theorem 2.1 (*g*) below) was introduced by Pliś as early as the 1970s (see [52]). Later on various applications of this property were given (see, e.g., [22]). Among equivalent characterizations of strongly convex sets let us emphasize the following: a set $A \subset H$ is strongly convex iff it can be represented as intersection of a family of closed balls of given radius $R > 0$. Such equivalence recalls the duality between two approaches in definition of convex sets: on one hand, a set A is convex whenever for each $x, y \in A$ it contains the entire segment

$$[x, y] := \{\lambda x + (1 - \lambda) y : 0 \leq \lambda \leq 1\}$$

(the direct approach) while, on the other hand, A is convex iff it can be represented as the intersection of a family of semispaces (the dual approach). Notice that a semispace can be seen as a ball of the radius $R = +\infty$. Furthermore, the geometric definition by Efimov and Stechkin [30] (see above) can be considered as a version of the dual approach to the weak convexity introduced by Vial in [63], although there is no total equivalence between them (the weak convexity by Vial implies the generalized convexity by Efimov and Stechkin at least in Hilbert spaces while vice versa is not true).

Let us mention that already at the beginning of 1960s Danzer et al. introduced in [24] an abstract convexity by means of intersection of a certain family of "elementary" sets (from some class $\mathcal{M}$), or, equivalently, through separation of points from a set by elements of the class $\mathcal{M}$. However, the cases when $\mathcal{M}$ is the family of (closed) semispaces (usual convexity) or the class of closed balls of fixed radius (strong convexity) seem to be more productive. We do not touch also the variety of other abstract as well as axiomatic concepts of convexity, which can be introduced in sets endowed with different structures. For details we refer to the book [61] and to the ample bibliography therein.

Returning to our matter, let us notice that the theories of *φ-convex* (more traditionally *proximally smooth* or *prox-regular*) and *weakly convex* (*by Vial)* sets have been developed independently for a long time. Besides the works mentioned above let us emphasize the contribution by Polovinkin, Balashov, and by the authors (see [3–12, 19, 20, 33, 34, 37–44, 54–56], etc.). A lot of applications of these classes of sets (and of strongly convex sets as well) was found in various fields of Nonlinear and Multivalued Analysis, Differential Equations, Numerical Calculus, Optimization, Theory of Games, etc. However, it was proved that the weak convexity by Vial is nothing else than the uniform prox-regularity, i.e., the smoothness of the distance function in an uniform tube of a radius $R > 0$ around the set (in other terms, φ-convexity with the function $\varphi(x) \equiv \text{const} = \frac{1}{2R}$).

Let us note that the Vial's definitions allow to emphasize better geometrical and topological properties of closed sets. In particular, the symmetry between weak and strong convexity has very nice and clear theoretical consequences as well as fruitful applications. However, this approach is essentially uniform and does not permit to study the local properties of weakly convex closed sets. To this objective the prox-regularity approach, enlarging a little bit of the weak convexity notion, seems to be more appropriate. Moreover, following the same logic (based on variational inequalities and proximal normals) one can define the strong convexity concepts (roughly speaking, localize the Vial's strong convexity). The recently introduced metric curvatures for convex closed (solid) sets in a Hilbert space (see [33]) partially fulfill this task.

In what follows we will consider mainly subsets of a Hilbert space, though recently the works on weak convexity (prox-regularity) appeared in Banach setting as well (see, e.g., [13, 14, 43]).

The paper is organized as follows. The next section "Strongly Convex Sets" is devoted to the main notions and results concerning strongly convex sets, while in section "Weakly Convex Sets" the authors give a symmetric sketch of theory of weakly convex sets. Furthermore, each of these sections is divided into two parts: in the first one the uniform (strong or weak) convexity is studied, equivalence of various characterizations is proved and the related questions are discussed; in the second part instead some local (and pointwise) constructions are introduced, and the connections with uniform concepts are established. Finally, in the last section "Balance between Weak and Strong Convexity" some applications of strong and weak convexity to the geometry of Hilbert spaces, to multivalued mappings and their continuous selections, and to a minimum time control problem are given. Notice that in all the applications two kinds of sets are involved, and the weak convexity of one set is somehow balanced by the strong convexity of another.

Strongly Convex Sets

Let H be a real Hilbert space with the inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$. Given $R>0$ and $c\in H$ we denote by $B_R(c)$ the closed ball of the radius R centered at the point c. By $\operatorname{int} A$, $\overline{A}$, and ∂A we mean as usual the interior, the closure, and the boundary of $A\subset H$. For a closed set $A\subset H$ let us give the list of main notations being used in the paper.

- $\operatorname{diam} A := \sup\limits_{x,y\in A}\|x-y\|$ is the *diameter* of A.
- $\sigma(p,A) := \sup_{a\in A}\langle p,a\rangle$, $p\in H$, is the *support function* of A.
- $N(a,A) := \{p\in H : \langle p,a\rangle = \sigma(p,A)\}$ is the *normal cone* to A at a point $a\in A$ (whenever the set A is convex).
- $d_A(x) := \inf\limits_{a\in A}\|x-a\|$ is the *distance* from a point $x\in H$ to A.
- $U_A(r) := \{x\in H : d_A(x) < r\}$ is the (open) *neighborhood* of radius $r>0$ around A.
- $P_A(x) := \{a\in A : \|x-a\| = d_A(x)\}$ is the *metric projection* of a point $x\in H$ onto A.

Besides the distance and the metric projection onto a closed set we will consider also the so-called *antidistance*

$$f_A(x) := \sup_{a \in A} \|x - a\|$$

from a point $x \in H$ to $A \subset H$, *antineighborhood* of radius $r > 0$

$$aU_A(r) := \{a \in H : f_A(a) > r\}$$

and *antiprojection* (the set of farthest points)

$$aP_A(x) := \{a \in A : \|x - a\| = f_A(x)\},$$

$x \in H$.

In what follows we use the *Minkowski operations* (*sum* and *difference*) between sets $A, B \subset H$ defined as

$$A + B := \{a + b : \ a \in A, b \in B\}, \ A \overset{*}{-} B := \{x \in H : x + B \subset A\} \tag{1}$$

and the *Hausdorff* (or *Hausdorff-Pompeiu*) *distance*

$$h(A, B) := \max\left\{\sup_{a \in A} d_B(a), \ \sup_{b \in B} d_A(b)\right\}. \tag{2}$$

Notice that obviously $U_A(r) = A + \operatorname{int} B_r(x)$, while $aU_A(r) = (H \setminus B_r(x)) + A$ (the latter equality was proved in [6]).

Let us associate to a closed convex set $A \subset H$ the *modulus of convexity* δ_A : $[0, \operatorname{diam} A) \to [0, +\infty)$ due to Polyak (see [57]):

$$\delta_A(\varepsilon) := \sup\left\{\delta > 0 : B_\delta\left(\frac{a_1 + a_2}{2}\right) \subset A \ \forall a_1, a_2 \in A, \|a_1 - a_2\| = \varepsilon\right\}. \tag{3}$$

It characterizes degree of the *uniform strict convexity* (*rotundity*) of the set A and equals zero whenever the boundary of A has at least one affine piece. In particular, for $A = B_R(c)$, $R > 0$, $c \in H$, by the simple geometric reasoning we immediately obtain

$$\delta_A(\varepsilon) = R - \sqrt{R^2 - \frac{\varepsilon^2}{4}}, \ 0 \le \varepsilon \le 2R.$$

In the second part of this section we consider another modulus of convexity (rotundity), which characterizes the strict convexity of a set locally while now let us introduce the basic uniform concepts and prove the equivalent assertions regarded to the strongly convex sets.

Uniform Strong Convexity

Definition 2.1 A subset $A \subset H$ is said to be **strongly convex** if there exist $R > 0$ (called **radius of strong convexity**) and a set $C \subset H$ such that

$$A = \bigcap_{c \in C} B_R(c).$$

In particular, the R-**strongly convex segment** with endpoints $x, y \in H$, $\|x - y\| \leq 2R$, is defined as the intersection of all balls $B_R(c)$, $c \in H$, containing x and y, and is denoted by $D_R(x, y)$.

Remark 2.1 If subset $A \subset H$ is strongly convex, then it is closed convex and bounded.

Theorem 2.1 (Equivalent Characterizations of a Strongly Convex Set) *Given a nonempty closed convex bounded set $A \subset H$ and a number $R > 0$ the following assertions are equivalent:*

(a) *A is strongly convex with radius R;*
(b) $\operatorname{diam} A \leq 2R$ *and* $D_R(a_1, a_2) \subset A$ *for all* $a_1, a_2 \subset A$*;*
(c) *for each two-dimensional affine subspace $L \subset H$ the set $A \cap L$ is either empty or strongly convex in the space L with radius R;*
(d) $\operatorname{diam} A \leq 2R$ *and* $\omega \subset A$ *for each circumference arc ω of radius R with length $\leq \pi R$ and with the endpoints belonging to A;*
(e) *(support principle) for all $a \in \partial A$ and $p \in N(a, A)$, $\|p\| = 1$, the inclusion*

$$A \subset B_R(a - Rp)$$

holds;
(f) *there exists a closed convex set $A_1 \subset H$ with $A + A_1 = B_R(0)$;*
(g) *given $a_1 \in \partial A$ and $p_1 \in N(a_1, A)$, $\|p_1\| = 1$, the inequality*

$$\langle p_1, a_1 - a_2 \rangle \geq \frac{1}{2R} \|a_1 - a_2\|^2$$

holds for all $a_2 \in A$;
(h) *for all $a_i \in \partial A$, $p_i \in N(a_i, A)$, $\|p_i\| = 1$, $i = 1, 2$, one has*

$$\langle p_1 - p_2, a_1 - a_2 \rangle \geq \frac{1}{R} \|a_1 - a_2\|^2;$$

(i) *for all $a_i \in \partial A$, $p_i \in N(a_i, A)$, $\|p_i\| = 1$, $i = 1, 2$, one has*

$$\|a_1 - a_2\| \leq R \|p_1 - p_2\|$$

or, in other terms, the support function $\sigma(\cdot, A)$ is Fréchet differentiable on the unit sphere and its gradient $\nabla\sigma(\cdot, A)$ satisfies the Lipschitz condition

$$\|\nabla\sigma(p_1, A) - \nabla\sigma(p_2, A)\| \le R\,\|p_1 - p_2\| \quad \forall p_1, p_2 \in \partial B_1(0)\,;$$

(j) *for all $a_i \in \partial A$, $p_i \in N(a_i, A)$, $\|p_i\| = 1$, $i = 1, 2$, one has*

$$\langle p_1 - p_2, a_1 - a_2\rangle \le R\,\|p_1 - p_2\|^2\,;$$

(k) *for any $a \in \partial A$ there exists $\varepsilon \in (0, R)$ such that the set $A \cap B_\varepsilon(a)$ is strongly convex with the radius R;*

(l) *(local support principle) for any $a \in \partial A$ there exist $\varepsilon \in (0, R)$ and $p \in H$, $\|p\| = 1$, such that $A \cap B_\varepsilon(a) \subset B_R(a - Rp)$;*

(m) $\operatorname{diam} A \le 2R$ *and*

$$\delta_A(\varepsilon) \ge R - \sqrt{R^2 - \frac{\varepsilon^2}{4}}, \;\; 0 < \varepsilon < \operatorname{diam} A;$$

(n) *the inequality*

$$\liminf_{\varepsilon \to 0+} \frac{\delta_A(\varepsilon)}{\varepsilon^2} \ge \frac{1}{8R}$$

holds;

(o) *for each $R_1 > R$ the antidistance function $f_A(\cdot)$ is Fréchet differentiable on the antineighborhood $aU_A(R_1)$;*

(p) *for each $R_1 > R$ the antiprojection operator $x \mapsto aP_A(x)$ is single-valued and continuous on the antineighborhood $aU_A(R_1)$;*

(q) *for each $R_1 > R$*

$$\|b_1 - b_2\| \le \frac{R}{R_1 - R}\,\|x_1 - x_2\| \quad \forall b_i \in aP_A(x_i)\,, \qquad \forall x_i \in aU_A(R_1)\,,\; i = 1, 2; \tag{4}$$

(r) *there exists $R_1 > R$ such that (4) holds;*

(s) *for all $x_i \in H$, $a_i \in P_A(x_i)$, $i = 1, 2$, one has*

$$\|a_1 - a_2\| \le \frac{R}{\sqrt{(R + d_A(x_1))\,(R + d_A(x_2))}} \times \sqrt{\|x_1 - x_2\|^2 - (d_A(x_1) - d_A(x_2))^2}.$$

Proof We prove the equivalence by the following scheme:

$$\begin{tikzcd}
(q) \arrow[dr, leftrightarrow] & (r) \arrow[d, leftrightarrow] & (s) & (k) \arrow[d, leftrightarrow] & (l) & \\
(o) \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & (a) \arrow[rr] \arrow[ur, leftrightarrow] \arrow[dr] & & (b) \arrow[r, leftrightarrow] \arrow[ur, leftrightarrow] \arrow[d] \arrow[dr, leftrightarrow] & (m) & \\
(p) & & (c) \arrow[r] & (d) \arrow[d] & (n) & \\
& (f) \arrow[uu] & & (e) \arrow[ll] \arrow[r] & (g) \arrow[r] & (h) \arrow[d] \\
& & & (j) \arrow[u] & & (i) \arrow[ll]
\end{tikzcd}$$

$(a) \Rightarrow (b)$ follows directly from the definitions of strongly convex set and strongly convex segment.

$(b) \Rightarrow (d)$ Let ω be an arc of a circumference of radius R with the length $\leq \pi R$ and the endpoints $a_1, a_2 \in A$. Then by the definition of strongly convex segment and by the statement (b) we get $\omega \subset D_R(a_1, a_2) \subset A$;

$(a) \Rightarrow (c)$ Let $A = \bigcap_{x \in C} B_R(x)$. Then $A \cap L = \bigcap_{x \in C}(B_R(x) \cap L)$. If the set $B_R(x) \cap L$ is nonempty, then it is either a point, or a disc with radius $\leq R$. So, $B_R(x) \cap L$ can be represented as an intersection of discs (balls in L) with radius R.

$(c) \Rightarrow (d)$ follows from the planarity of an arc.

$(d) \Rightarrow (e)$ (See Fig. 1). Fix any $a \in \partial A$ and $p \in N(a, A)$, $\|p\| = 1$. Suppose that there exists a point $a_2 \in A \setminus B_R(a - Rp)$. Let L be the affine hull of the vectors a_2, a, and $a - Rp$. Take a point $q \in L - a$ such that $\langle p, q \rangle = 0$ and $\langle q, a_2 - a \rangle \geq 0$. By the assertion (d) we, in particular, have $\|a_2 - a\| \leq 2R$. Furthermore, $a_2 \neq a$ since $a_2 \notin B_R(a - Rp)$. Let a_1 be a point of the circumference $(\partial B_R(a - Rp)) \cap L$

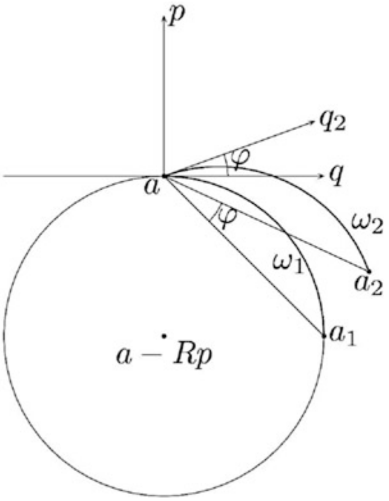


Fig. 1 Proof of the implication (d) ⇒ (e) in Theorem 2.1

with $\|a-a_1\| = \|a-a_2\|$ and $\langle q, a_1 - a\rangle \geq 0$. Denote by φ the angle between the vectors $a_2 - a$ and $a_1 - a$. Then $\varphi \in \left(0, \frac{\pi}{2}\right]$ due to the fact that $\langle p, a_1 - a\rangle \leq 0$, $\langle p, a_2 - a\rangle \leq 0$ (remind that $a_1 \in \partial B_R(a - Rp)$ and $p \in N(a, A)$, $a_2 \in A$) and that $\langle q, a_1 - a\rangle \geq 0, \quad \langle q, a_2 - a\rangle \geq 0$ by our construction. Observe that $\varphi \neq 0$ because $a_2 \neq a_1$. Let us denote by ω_1 the shortest arc of the circumference $(\partial B_R(a-Rp)) \cap L$ with endpoints a and a_1 and consider the rotation of the plane $L-a$ around the point a on the angle φ in the direction from q to p. This rotation maps the point a_1 to a_2 and the arc ω_1 to some arc connecting a and a_2 (let ω_2). Then the image q_2 of q is tangent to ω_2 at the point a similarly as q is tangent to ω_1. Moreover, q_2 is directed from a towards a_2. By the assertion (d) we get $\omega_2 \subset A$, and it follows from $p \in N(a, A)$ that $\langle p, x - a\rangle \leq 0$ for all $x \in \omega_2$. Consequently, $\langle p, q_2\rangle \leq 0$ since the vector q_2 (being tangent to ω_2) is limit of secants passing through points satisfying the same inequality. On the other hand, q is perpendicular to p while q_2 is obtained by rotation from q to p on the angle $\varphi \in \left(0, \frac{\pi}{2}\right]$. So, the angle between q_2 and p is acute, which is contradiction.

$(e) \Rightarrow (f)$ Set $A_1 := B_R(0) \overset{*}{-} A$. By the definition of the Minkowski difference $A + A_1 \subset B_R(0)$. Suppose that the inverse inclusion is false, i.e., there exists a point $x \in B_R(0)$ such that $x \notin A_1 + A$. Since the set $A_1 + A$ is convex and closed (by the weak compactness) it follows by Hahn–Banach separation theorem that there exists a vector p with $\|p\| = 1$ such that $\sigma(p, A_1) + \sigma(p, A) < \langle p, x\rangle \leq R$. Due to the weak compactness of A we can associate to p a point $a \in A$ with $\sigma\,(p, A) = \langle p, a\rangle$. By the statement (e) we have $A \subset B_R(a - Rp)$. Hence, by the definition of the Minkowski difference, $Rp - a \in B_R(0) \overset{*}{-} A = A_1$. Consequently, $\langle p, Rp - a\rangle \leq \sigma(p, A_1)$. So, $R \leq \sigma(p, A_1) + \langle p, a\rangle = \sigma(p, A_1) + \sigma(p, A)$. This contradicts the inequality $\sigma(p, A_1) + \sigma(p, A) < R$ above.

$(f) \Rightarrow (a)$ Let A_1 be closed convex set from the assertion (f). Then by applying Hahn–Banach separation theorem and the definition of the Minkowski operations we have $A + A_1 \overset{*}{-} A_1 = A$, i.e., $A = B_R(0) \overset{*}{-} A_1 = \bigcap_{c \in -A_1} B_R(c)$.

$(e) \Rightarrow (g)$ Fix arbitrary $a_1 \in \partial A$, $p_1 \in N(a_1, A)$, $\|p_1\| = 1$, and $a_2 \in A$. By the statement (e) we have $a_2 \in B_R(a_1 - Rp_1)$, i.e.,

$$\|a_1 - Rp_1 - a_2\|^2 = \|a_1 - a_2\|^2 - 2R\,\langle p_1, a_1 - a_2\rangle + R^2 \leq R^2,$$

which yields (g).

$(g) \Rightarrow (h)$ Let us fix $a_i \in \partial A$, $p_i \in N(a_i, A)$, $\|p_i\| = 1$, $i = 1, 2$. From (g) we have

$$2R\langle p_1, a_1 - a_2\rangle \geq \|a_1 - a_2\|^2, \qquad 2R\langle p_2, a_2 - a_1\rangle \geq \|a_1 - a_2\|^2.$$

Adding the above inequalities, we obtain the desired result.

The implications $(h) \Rightarrow (i)$ and $(i) \Rightarrow (j)$ follow successively from the Cauchy–Schwarz inequality.

$(j) \Rightarrow (e)$ Fix arbitrary $a \in A$ and $p \in N(a, A)$, $\|p\| = 1$, and put $c = a - Rp$. We should prove that $A \subset B_R(c)$. Let us denote $R_0 := \sup_{a \in A} \|a - c\|$. Then the desired inclusion holds true iff $R_0 \le R$. Suppose that $R_0 > R$ and choose a sequence $\{x_k\} \subset A$ such that $R < \|x_k - c\| \to R_0$ as $k \to \infty$. For all $k \in \mathbb{N}$ let us define

$$p_k = \frac{x_k - c}{\|x_k - c\|},$$

and $a_k \in A$ be such that $\sigma(p_k, A) = \langle p_k, a_k \rangle$. Since $p_k \in N(a_k, A)$ and $\|p_k\| = 1$, from the statement (*j*) it follows that

$$\langle p_k - p, a_k - a \rangle \le R\|p_k - p\|^2 \qquad \forall k \in \mathbb{N}. \tag{5}$$

On the other hand, by the choice of a_k we have $\langle p_k, x_k \rangle \le \langle p_k, a_k \rangle$. Consequently, passing to the limits,

$$\liminf_{k\to\infty} \langle p_k, a_k - c \rangle \ge \liminf_{k\to\infty} \langle p_k, x_k - c \rangle = \liminf_{k\to\infty} \|x_k - c\| = R_0$$

and

$$\limsup_{k\to\infty} \|a_k - c - R_0 p_k\|^2 \le \limsup_{k\to\infty} \left(2R_0^2 - 2R_0\langle p_k, a_k - c\rangle\right) \le 0.$$

It means that

$$\varepsilon_k := \|a_k - c - R_0 p_k\| \to 0, \quad k \to \infty. \tag{6}$$

Since

$$\begin{aligned}
\langle p_k - p, a_k - a \rangle &\ge \langle p_k - p, c + R_0 p_k - a \rangle - \|p_k - p\| \cdot \|a_k - c - R_0 p_k\| \\
&\ge \langle p_k - p, R_0 p_k - Rp \rangle - 2\varepsilon_k \\
&= (R_0 + R)(1 - \langle p, p_k \rangle) - 2\varepsilon_k \\
&= \frac{R_0 + R}{2}\|p_k - p\|^2 - 2\varepsilon_k,
\end{aligned}$$

by (5) we obtain

$$\frac{R_0 + R}{2}\|p_k - p\|^2 - 2\varepsilon_k \le R\|p_k - p\|^2.$$

Thus, $(R_0 - R)\|p_k - p\|^2 \le 4\varepsilon_k \to 0$ as $k \to \infty$ (see (6)), and we deduce that $p_k \to p$ as $k \to \infty$ (recall that $R_0 > R$). Then

$$\langle p, a - x_k \rangle = \langle p, Rp + c - x_k \rangle = R - \|x_k - c\|\langle p, p_k \rangle \to R - R_0 < 0,$$

contradicting the fact that $p \in N(a, A)$ and $x_k \in A$.

For the further equivalences $(b) \Leftrightarrow (k)$, $(b) \Leftrightarrow (l)$, $(b) \Leftrightarrow (m)$, $(b) \Leftrightarrow (n)$, $(a) \Leftrightarrow (o)$, $(o) \Leftrightarrow (p)$, $(a) \Leftrightarrow (q)$, $(a) \Leftrightarrow (r)$, and $(a) \Leftrightarrow (s)$ we refer, respectively, to [64, Theorem 1.1], [64, Theorem 1.2], [12, Theorem 2.1], [12, Theorem 2.1] (see also [64, Proposition 5.2]), to [40, Theorem 1], [40, Remark 5], [7, Theorems 2, 5], [4, Theorem 2.2], and to [5, Theorem 2.2]. ■

Remark 2.2 Observe that we need the boundedness assumption in Theorem 2.1 because the item (j) doesn't imply that the set A is bounded [e.g., each affine subspace of H satisfies (j)] while the boundedness immediately follows from each of the variational inequalities (g)–(i). On the other hand, from Theorem 2.1 (m) one immediately deduces that each strongly convex set $A \subset H$ has nonempty interior unless it is not singleton.

Remark 2.3 The estimations in the statements (g)–(j), (m), (n), (q), (s) are exact. Namely, they become equalities whenever $A \subset H$ is a ball of the radius R.

Remark 2.4 If a set $A \subset H$ is strongly convex with the radius $R > 0$, then it is strongly convex with any radius $R_1 > R$ as well. Vice versa, if $A \subset H$ is strongly convex with each radius $R_1 > R$ ($R > 0$ is fixed), then A is strongly convex with the radius R. It follows from the statement (e) of Theorem 2.1.

Remark 2.4 implies that for any strongly convex $A \subset H$ there exists the *minimal (or sharp) radius of the strong convexity.*

Theorem 2.2 *Given a closed convex bounded set $A \subset H$ and a number $R > 0$ the following assertions are equivalent:*

(a) *R is the minimal radius of strong convexity of A;*
(b)

$$\lim_{\varepsilon \to +0} \frac{\delta_A(\varepsilon)}{\varepsilon^2} = \frac{1}{8R};$$

(c) *there exists $R_1 > R$ such that $\frac{R}{R_1 - R}$ is the minimal Lipschitz constant for the antiprojection operator $aP_A(\cdot)$ on the antineighborhood $aU_A(R_1)$;*
(d) *for each $R_1 > 0$ the metric projection operator $P_A(\cdot)$ is Lipschitz continuous with the minimal constant $\frac{R}{R+R_1}$ on the set $H \setminus U_A(R_1)$.*

Proof The equivalences $(a) \Leftrightarrow (b)$, $(a) \Leftrightarrow (c)$, and $(a) \Leftrightarrow (d)$ were proved, respectively, in [12, Theorem 2.1], [4, Corollary 2.3], and in [5, Theorem 2.1 and Corollary 2.1]. ■

Theorem 2.3 (Hölderian Dependence of the Metric Projection and the Antiprojection Upon the Set) *Let $A_1 \subset H$ be a strongly convex set with the radius $R > 0$; $A_2 \subset H$ be an arbitrary (may be nonconvex) set; $h := h(A_1, A_2)$ be the Hausdorff-Pompeiu distance between these sets, and $x \in H$ be an arbitrary point. Then* *(a) for the metric projections* $a_i \in P_{A_i}(x)$ *(*$i = 1, 2$*) one has*

$$\|a_1 - a_2\| \leq \sqrt{\frac{4Rhd}{R+d} + h^2}$$

where $d := d_{A_1}(x)$*;*

(b) if $R_1 := f_{A_1}(x) > R$*, then the antiprojections* $b_i \in aP_{A_i}(x)$ *(*$i = 1, 2$*) satisfy the inequality*

$$\|b_1 - b_2\| \leq \sqrt{\frac{4RR_1h}{R_1 - R} + h^2}.$$

Proof (a) If $d = 0$, then the desired inequality is obvious. Assume that $d > 0$. Since $p := \frac{x-a_1}{d} \in N(a_1, A_1)$ and $\|p\| = 1$, by the assertion (e) of Theorem 2.1 we have $A_1 \subset B_R(a_1 - Rp)$. Consequently, $a_2 \in A_2 \subset B_{R+h}(a_1 - Rp)$, i.e., $\|a_1 - Rp - a_2\| \leq R + h$. Squaring this inequality and reducing similar terms give

$$\|a_1 - a_2\|^2 - 2R\langle p, a_1 - a_2\rangle \leq 2Rh + h^2. \tag{7}$$

On the other hand, writing the inequality $d_{A_2}(x) \leq d_{A_1}(x) + h = d + h$ as $\|x - a_1 + a_1 - a_2\| = \|x - a_2\| \leq d + h$, by the same reasoning as above, we obtain

$$2\langle x - a_1, a_1 - a_2\rangle + \|a_1 - a_2\|^2 \leq 2dh + h^2,$$

or, after multiplying by $\frac{R}{d}$,

$$2R\langle p, a_1 - a_2\rangle + \frac{R}{d}\|a_1 - a_2\|^2 \leq 2Rh + \frac{R}{d}h^2.$$

Summing the latter inequality and (7) we arrive at

$$\left(1 + \frac{R}{d}\right)\|a_1 - a_2\| \leq 4Rh + \left(1 + \frac{R}{d}\right)h^2$$

that completes proving of the assertion (a). Proof of the statement (b) is analogous. ■

Remark 2.5 The estimations of Theorem 2.3 are exact in the sense that there exist sets and points satisfying the assumptions of the theorem such that the inequalities become equalities.

Local Strong Convexity

For the sake of simplicity of the further constructions let us associate to a convex closed bounded set $A \subset H$ the *Minkowski functional* (or *gauge function*)

$$\rho_A(x) := \inf\left\{\lambda > 0 : \lambda^{-1}x \in A\right\}, \quad x \in H, \tag{8}$$

which is somehow dual to the support function. Namely, $\rho_A(x) = \sigma\left(x, A^0\right)$, $x \in H$, where

$$A^0 := \{p \in H : \langle p, x\rangle \leq 1 \ \ \forall x \in A\}$$

is the *polar set* for A. Observe also that $\rho_A(x) < +\infty \ \forall x \in H$ iff $0 \in \operatorname{int} A$ (or, equivalently, A^0 is bounded). In what follows to avoid confusion we will assume that $0 \in \operatorname{int} A$ although the results concerning localization of the strong convexity remain true without this assumption as well (it is enough to require only that $\operatorname{int} A \neq \emptyset$ and make some translation).

There are various ways to measure the strict convexity of a set and, respectively, various "moduli of strict convexity." One of them is given by the formula (3) and essentially uniform. "Localizing" this modulus we can consider the deviation of the (sublinear) gauge function from a linear one near a fixed point (see, for instance, [25]), namely

$$\begin{aligned}\hat{\delta}_A(\varepsilon, a) &:= \inf\{\rho_A(a) + \rho_A(b) - \rho_A(a+b) : b \in \partial A, \|a-b\| \geq \varepsilon\}\\ &= 2\inf\left\{1 - \rho_A\left(\frac{a+b}{2}\right) : b \in \partial A, \|a-b\| \geq \varepsilon\right\}, \quad a \in \partial A.\end{aligned}$$

Notice that this modulus is strongly related with the modulus (3). In particular, if $B_r(0) \subset A \subset B_R(0)$, then

$$\frac{2}{R}\delta_A(\varepsilon) \leq \inf_{a \in \partial A} \hat{\delta}_A(\varepsilon, a) \leq \frac{2}{r}\delta_A(\varepsilon), \qquad \varepsilon > 0.$$

For our objectives instead the following *modulus of rotundity* is more convenient:

$$\mathfrak{C}_A(\varepsilon, a, p) := \inf\{\langle p, a-b\rangle : b \in A, \|a-b\| = \varepsilon\}, \varepsilon > 0. \tag{9}$$

Here $a \in \partial A$, and $p \in N(a, A)$ is normalized by such a way that $p \in \partial A^0$. This normalization for p (in the place of the condition $\|p\| = 1$) is chosen to emphasize the duality between the geometric properties of the sets A and A^0. Observe that the condition $p \in N(a, A)$ is equivalent to $\langle p, a\rangle = 1$ provided that $a \in \partial A$, $p \in \partial A^0$. Furthermore, it is convenient to consider the so-called *duality mapping* $\mathfrak{J}_A : \partial A^0 \to \partial A$,

$$\mathfrak{J}_A(p) := \{a \in \partial A : \langle p, a\rangle = 1\}, \quad p \in \partial A^0,$$

which is "autoreverse" in the sense that $\mathfrak{J}_A^{-1}(a) = \mathfrak{J}_{A^0}(a)$ for all $a \in \partial A$. For instance, if $B = B_R(0)$, then given $a \in \partial B$ the unique element $p \in \partial B^0$ with $a \in \mathfrak{J}_B(p)$ is $\frac{a}{R^2}$ and

$$\mathfrak{C}_B(\varepsilon, a, p) = \frac{\varepsilon^2}{2R^2},$$

which is strictly larger than $\hat{\delta}_B(\varepsilon, a) = \delta_B(\varepsilon)$, $\varepsilon > 0$.

By using the modulus (9) we can define the “first order strengthening” of the usual convexity.

Definition 2.2 Given $a \in \partial A$ and $p \in \mathfrak{J}_{A^0}(a)$ we say that the set A is (**locally**) **strictly convex** at the point a with respect to (w.r.t.) the direction p if $\mathfrak{C}_B(\varepsilon, a, p) > 0$ for all $\varepsilon > 0$.

Taking into account that the local strict convexity of A at $a \in \partial A$ w.r.t. p implies that a is an *exposed point* of A, and the vector p *exposes* a (i.e., the hyperplane $\{x \in H : \langle p, x\rangle = \sigma(p, A) = 1\}$ touches A at the point a only), we have $\mathfrak{J}_A(p) = \{a\}$. So, one can speak just about the strict convexity w.r.t. the direction p (do not referring to the unique $a \in \mathfrak{J}_A(p)$). Moreover, the strict convexity as defined here is equivalent to the fact that a is a *strongly exposed point* of A w.r.t. the direction p, i.e., both a is exposed one and each sequence $\{x_n\} \subset A$ with $\langle p, x_n\rangle \to \langle p, a\rangle$, $n \to \infty$, converges to a.

Theorem 2.4 (Equivalent Characterizations of a Locally Strictly Convex Set) *Let $A \subset H$ be a closed convex bounded set containing the origin in the interior. Then the following assertions are equivalent:*

(a) A is strictly convex at $a \in \partial A$ w.r.t. a direction $p \in \partial A^0$;
(b) a is a strongly exposed point of A w.r.t. p;
(c) the duality mapping $\mathfrak{J}_A : \partial A^0 \to \partial A$ is Hausdorff continuous at p with $\mathfrak{J}_A(p) = \{a\}$, $a \in \partial A$;
(d) the support function $\sigma(\cdot, A)$ is Fréchet differentiable at p, and $\nabla\sigma(p, A) = a$ (compare with the second part of the assertion (i) of Theorem 2.1).

Proof $(a) \Leftrightarrow (b)$ follows immediately from the definitions.

For proof of the equivalences $(b) \Leftrightarrow (d)$ and $(d) \Leftrightarrow (c)$ we refer, respectively, to [51, Proposition 5.11] and to [2, Corollary 2, p. 460]. ■

The “second order” strict convexity is given by the following definition:

Definition 2.3 Fix $p \in \partial A^0$ and let $a \in \partial A$ be an unique element of $\mathfrak{J}_A(p)$. The set A is said to be **strongly convex** (or **rotund**) with respect to p (at the point a) if

$$\varkappa_A(a, p) := \frac{1}{\|p\|} \liminf_{\substack{(\varepsilon, x, v) \to (0+, a, p) \\ x \in \mathfrak{J}_A(v),\ v \in \partial A^0}} \frac{\mathfrak{C}_A(\varepsilon, x, v)}{\varepsilon^2} > 0. \tag{10}$$

It follows directly from the definitions that the strong convexity at the point a w.r.t. p implies the strict convexity in the sense of Definition 2.2. Due to Theorem 2.4 (c) in (10) as well as in similar formulas below it is enough to require only the convergence of normals $v \to p$ while the convergence of points $x \to a$ holds automatically. The positive quantity $\varkappa_A(a, p)$ (which can be equal to $+\infty$) characterizes degree of the rotundity of the set A close to the point a in the direction p. The following geometric representation of $\varkappa_A(a, p)$ implies, in particular, that it is invariant w.r.t. translations of A (in other words, it does not depend on location of the origin inside A). So, we call $\varkappa_A(a,p)$ the (scalar) metric *curvature* of the set A w.r.t. $p \in \partial A^0$ (at the point $a \in \partial A$).

Theorem 2.5 *Let A be a convex closed bounded set with $0 \in \operatorname{int} A$, which is, moreover, strongly convex w.r.t. $p \in \partial A^0$, and $a \in \partial A$ be a unique element of $\mathfrak{J}_A(p)$. Then*

$$\mathfrak{R}_A(a,p) := \frac{1}{2\|p\|\,\varkappa_A(a,p)}$$
$$= \limsup_{\substack{(\varepsilon,x,v)\to(0+,a,p)\\ x\in\mathfrak{J}_A(v),\ \ v\in\partial A^0}} \inf\left\{r>0 : A\cap B_\varepsilon(x)\subset B_{r\|v\|}(x-rv)\right\}. \tag{11}$$

Proof Let us denote by R the right-hand side of the equality (11), assume that $R<+\infty$ and prove that $\mathfrak{R}_A(a,p)\le R$. Given any $\rho>R$, by the definition of R, for each $\varepsilon>0$ small enough and for each dual pair (x,v) close to (a,p) we have the inclusion

$$A\cap B_\varepsilon(x)\subset B_{\rho\|v\|}(x-\rho v).$$

Hence, in particular, $\|y-x+\rho v\|^2\le\rho^2\|v\|^2$ for all $y\in A$ with $\|x-y\|=\varepsilon$. After the simple transformations we arrive at

$$\langle v, x-y\rangle \ge \frac{\varepsilon^2}{2\rho}. \tag{12}$$

Passing then to infimum in y, we obtain (see (9))

$$\frac{1}{2\rho}\le\frac{\mathfrak{C}_A(\varepsilon,x,v)}{\varepsilon^2}.$$

The first part of theorem will be proved if we pass in the latter inequality to lower limit as $(\varepsilon,x,v)\to(0+,a,p)$ and then to limit as $\rho\to R+$.

In order to show the opposite inequality let us assume that $R>0$ and take an arbitrary $\rho\in(0,R)$. By definition of the upper limit there exist $\varepsilon>0$ small enough and a dual pair (x,v) enough close to (a,p) such that $r>\rho$ whenever $A\cap B_\varepsilon(x)\subset B_{r\|v\|}(x-rv)$. In particular, the latter inclusion fails for $r=\rho$. Therefore, for some $y\in A$, $0<\|x-y\|\le\varepsilon$, we have $\|y-x+\rho v\|^2>\rho^2\|v\|^2$, or, in other form,

$$\langle v, x-y\rangle < \frac{\|x-y\|^2}{2\rho}.$$

Setting $r = r(\varepsilon) := \|x - y\| > 0$ we deduce from (9) that

$$\frac{\mathfrak{C}_A(r, x, v)}{r^2} < \frac{1}{2\rho}$$

and, passing to lower limit as $\varepsilon \to 0+$, $(x, v) \to (a, p)$ and to limit as $\rho \to R-$ we arrive at the required inequality. ■

Due to Theorem 2.5 the number $\mathfrak{R}_A(a, p) \geq 0$ can be naturally named the *curvature radius* of A at $a \in \partial A$ in the direction p.

Besides the above constructions, which characterize just the local structure of the set close to a fixed boundary point (w.r.t. some direction), for $p \in \partial A^0$, $a \in \mathfrak{J}_A(p)$ one can consider the so-called *scaled curvature*

$$\tilde{\varkappa}_A(a, p) := \frac{1}{\|p\|} \liminf_{\substack{(x,v)\to(a,p)\\ x\in\mathfrak{J}_A(v),\ v\in\partial A^0}} \inf_{\varepsilon>0} \frac{\mathfrak{C}_A(\varepsilon, x, v)}{\varepsilon^2} \tag{13}$$

and, respectively, *scaled curvature radius*

$$\tilde{\mathfrak{R}}_A(a, p) := \frac{1}{2\|p\|\,\tilde{\varkappa}_A(a, p)} =$$

$$\limsup_{\substack{(x,v)\to(a,p)\\ x\in\mathfrak{J}_A(v),\ v\in\partial A^0}} \inf\left\{r > 0 : A \subset B_{r\|v\|}(x - rv)\right\}. \tag{14}$$

They are also invariant w.r.t. translations of the set A, but unlike the local constructions (10) and (11) depend also on the size of A. In particular, the curvature $\tilde{\varkappa}_A(a, p)$ cannot be too large, namely $\tilde{\mathfrak{R}}_A(a, p)$ is not smaller than the *Chebyshev radius* of the set A.

Notice that in above constructions we involve the limit passage in (x, v) in order to guarantee the lower semicontinuity of curvatures (respectively, upper semicontinuity of curvature radii), which is important for applications.

The connection of the curvatures with the definitions of section "*Uniform Strong Convexity*" is given by the following statement.

Theorem 2.6 *Let $A \subset H$ be a convex closed bounded set with $0 \in \operatorname{int} A$ and $R > 0$. Then the following assertions are equivalent:*

(a) the set A is strongly convex with the radius R;
(b) $\tilde{\varkappa}_A(a, p) \geq \frac{1}{2R}$ for each $a \in \partial A$ and $p \in \mathfrak{J}_{A^0}(a)$;
(c) $\varkappa_A(a, p) \geq \frac{1}{2R}$ for each $a \in \partial A$ and $p \in \mathfrak{J}_{A^0}(a)$.

Proof The proving is essentially based on the local support principle (see Theorem 2.1 (l)) and on the formulas (11) and (14).

The inequalities in (b) and (c) can be rewritten as $\|p\|\,\tilde{\mathfrak{R}}_A(a,p) \le R$ and $\|p\|\,\mathfrak{R}_A(a,p) \le R$, respectively. So, the implication (a) $\Rightarrow$ (b) follows from both (14) and Theorem 2.1 (e). By comparing formulas (11) and (14), we see that $\mathfrak{R}_A(a,p) \le \tilde{\mathfrak{R}}_A(a,p)$, and the implication ($b$) $\Rightarrow$ (c) follows.

In order to prove (c) $\Rightarrow$ (a) let us fix any $R_1 > R$ and $a \in \partial A$. Take an arbitrary $q \in N(a,A)$, $\|q\| = 1$, and set $p := \frac{q}{\rho_{A^0}(q)} \in \partial A^0$. We have $a \in \mathfrak{J}_A(p)$, and by virtue of the assertion (c) the inequality $\frac{R_1}{\|p\|} > \mathfrak{R}_A(a,p)$ holds. By using (11), we get

$$\frac{R_1}{\|p\|} > \limsup_{\varepsilon \to 0+} \inf\left\{r > 0 : A \cap B_\varepsilon(a) \subset B_{r\|p\|}(a - rp)\right\}.$$

Consequently, there exists $\varepsilon > 0$ such that $A \cap B_\varepsilon(a) \subset B_{R_1}\left(a - R_1\frac{p}{\|p\|}\right) = B_{R_1}(a - R_1 q)$. It means that the assertion (l) of Theorem 2.1 holds and hence the set A is strongly convex with radius R_1. Since $R_1 > R$ is arbitrary, we conclude (see Remark 2.4) that A is strongly convex with the radius R, and theorem is proved. ■

Up to changing the notations Theorems 2.4–2.6 hold also true if the condition $0 \in \operatorname{int} A$ is substituted by $\operatorname{int} A \neq \emptyset$. The latter hypothesis instead is very natural (see Remark 2.2).

Concluding this section we give a duality result between the (local) strong convexity and "strong" smoothness. To this end let us define two more moduli of rotundity, which are equivalent to (9) but have no a clear geometrical interpretation as (11). Namely, given convex closed bounded $A \subset H$, $0 \in \operatorname{int} A$, and a dual pair $(a,p) \in \partial A \times \partial A^0$, $p \in \mathfrak{J}_{A^0}(a)$, we set

$$\mathfrak{C}_A^-(\varepsilon,a,p) := \inf\{\langle p, a-b\rangle : b \in A, \rho_A(a-b) = \varepsilon\}$$

and

$$\mathfrak{C}_A^+(\varepsilon,a,p) := \inf\{\langle p, a-b\rangle : b \in A, \rho_A(b-a) = \varepsilon\}.$$

Denoting by $\|A\|$ and $\left\|A^0\right\|$ the radii of the spheres circumscribed around A and inscribed into A, respectively, we always have

$$\mathfrak{C}_A^{\pm}\left(\frac{\varepsilon}{\|A\|}, a, p\right) \le \mathfrak{C}_A(\varepsilon,a,p) \le \mathfrak{C}_A^{\pm}\left(\varepsilon\left\|A^0\right\|, a, p\right), \varepsilon > 0.$$

So, the set A is strongly convex w.r.t. p at $a \in \partial A$ iff

$$\varkappa_A^{\pm}(a,p) := \frac{1}{\|p\|} \liminf_{\substack{(\varepsilon,x,v)\to(0+,a,p) \\ x\in\mathfrak{J}_A(v),\ v\in\partial A^0}} \frac{\mathfrak{C}_A^{\pm}(\varepsilon,x,v)}{\varepsilon^2} > 0.$$

Observe that $\varkappa_A^{\pm}(a,p)$ are not "true" curvatures of A, because they depend on the structure of the set A "in large" (not only close to a). In particular, the invariantness w.r.t. translations fails, and $\varkappa_A^{\pm}(a,p)$ depend on location of the origin inside A.

Example 2.1 Fix $a \in H$ with $\|a\| < 1$ and consider the unit ball $B = B_1(a)$, containing the origin as an interior point. Let x and y be the points of intersection of the sphere ∂B with the line $\{\lambda a : \lambda \in \mathbb{R}\}$ such that the vector $\overrightarrow{Ox}$ has the same direction as a, while $\overrightarrow{Oy}$ has the opposite direction. If p and q are the (unique) normals to B at the points x and y, respectively, normalized so that $p, q \in \partial B^0$, then we have

$$\varkappa_A^{\pm}(x,p) = \frac{1-\|a\|}{2(1+\|a\|)} \quad \text{and} \quad \varkappa_A^{\pm}(y,q) = \frac{1+\|a\|}{2(1-\|a\|)},$$

while the "true" curvature $\varkappa_A$ in both points (as well as in all $x \in \partial B$) is equal to $1/2$ (the curvature radius $= 1$).

To treat the differentiability properties of the boundary ∂A one uses the so-called *modulus of smoothness*

$$\mathfrak{S}_A(t,a,p) := \sup\{\rho_A(a+tz) - \rho_A(a) - t\langle p,z\rangle : z \in A\}, t \in \mathbb{R}, \tag{15}$$

where as usual (a,p) is a dual pair, $a \in \partial A$, $p \in \mathfrak{J}_{A^0}(a)$. A convex closed bounded set $A \subset H$ with $0 \in \operatorname{int} A$ is said to be *smooth* at $a \in \partial A$ if there exists unique vector $p \in \mathfrak{J}_{A^0}(a)$ such that

$$\lim_{t \to 0} \frac{\mathfrak{S}_A(t,a,p)}{t} = 0.$$

Recalling Theorem 2.4 ((*a*) ⇔ (*d*)) we see that $A \subset H$ is strictly convex w.r.t. $p \in \partial A^0$ (at the unique $a \in \mathfrak{J}_A(p)$) iff the polar set A^0 is smooth at p (w.r.t. unique vector $a \in \partial A$, which is nothing else than the *Fréchet gradient* $\nabla\sigma(p,A)$).

The refinement of the smoothness property above is obtained through the duality formula between the rotundity and smoothness moduli. Such kind a formula was obtained first by Lindenstrauss in 1963 (see [46, Theorem 1]) to lighten the duality between the norms in conjugate Banach spaces. We give a nonsymmetric version of this formula.

Theorem 2.7 (Lindenstrauss Formula) *Let $A \subset H$ be a convex closed bounded set with $0 \in \operatorname{int} A$, and $(a,p) \in \partial A \times \partial A^0$ be such that $\langle p,a\rangle = 1$. Then for each $t > 0$*

$$\mathfrak{S}_{A^0}(\pm t,p,a) = \sup\left\{\varepsilon t - \mathfrak{C}_A^{\pm}(\varepsilon,a,p) : \varepsilon > 0\right\}. \tag{16}$$

Proof Let us prove only one of the equalities (16), namely for $\mathfrak{C}_A^{+}(\varepsilon,a,p)$, since proving of the second is similar.

Fix $t > 0$. Then by the definition of the modulus of smoothness (see (15)) and taking into account that $\rho_{A^0}(\cdot) = \sigma(\cdot, A)$ we can associate to each $\eta > 0$ some points $x \in A$ and $v \in A^0$ with

$$\begin{aligned}
\mathfrak{S}_{A^0}(t,p,a) &\le \langle p + tv, x\rangle - \langle p, a\rangle - t\langle v, a\rangle + \eta \\
&\le \langle p, x - a\rangle + t\rho_A(x-a) + \eta \\
&\le \sup_{y\in A}\{t\rho_A(y-a) - \langle p, a-y\rangle\} + \eta \\
&\le \sup_{\varepsilon>0}\ \sup_{y\in A,\ \rho_A(y-a)=\varepsilon}\{\varepsilon t - \langle p, a-y\rangle\} + \eta \\
&\le \sup_{\varepsilon>0}\left\{\varepsilon t - \mathfrak{C}_A^{+}(\varepsilon, a, p)\right\} + \eta,
\end{aligned}$$

and the inequality "$\le$" in (16) follows. Reversing the reasoning one easily shows the opposite inequality as well. ■

Thus the modulus of smoothness (15) is nothing else than the *Legendre-Fenchel transform* of the function defined as

$$\begin{cases} \mathfrak{C}_A^{-}(-\varepsilon, a, p) & \text{if } \varepsilon < 0, \\ 0 & \text{if } \varepsilon = 0, \\ \mathfrak{C}_A^{+}(\varepsilon, a, p) & \text{if } \varepsilon > 0. \end{cases}$$

For the proofs of the following two statements we refer to [33, Propositions 4.3 and 4.4].

Theorem 2.8 *Let a set $A \subset H$ and a dual pair (a,p) be such as in Theorem 2.7. Then*

$$\limsup_{\substack{(t,x,v)\to(0\pm,a,p)\\ x\in\mathfrak{J}_A(v),\ \ v\in\partial A^0}} \frac{\mathfrak{S}_{A^0}(t,v,x)}{t^2} = \frac{1}{4\|p\|\varkappa_A^{\pm}(a,p)}. \tag{17}$$

So, if A is strongly convex w.r.t. $p \in \partial A^0$ (at unique point $a \in \partial A$), then the upper limit (17) is finite, and we say that the polar set A^0 is *strongly smooth* at p.

Theorem 2.9 *In addition to the hypotheses of Theorem 2.8 assume that the boundary of A^0 is of class $\mathcal{C}^2$ near the point $p \in \partial A^0$, and $a = \nabla\sigma(p, A)$. Then the set A is strongly convex w.r.t. p (at the point a), $\varkappa_A^{+}(a,p) = \varkappa_A^{-}(a,p)$, and*

$$\sup_{v\in A^0}\left\langle \nabla^2\sigma(p,A)\,v, v\right\rangle = \frac{1}{2\|p\|\varkappa_A^{\pm}(a,p)}.$$

Weakly Convex Sets

For a nonempty closed (not necessarily convex) set $A \subset H$ in addition to notions of section "Strongly Convex Sets" in what follows we'll use also

- the *proximal normal cone* to A at $a \in A$,

$$\begin{aligned} N^P(a, A) &:= \{p \in H : \text{there exists } \sigma > 0 \\ &\quad \text{such that } \langle p, x - a\rangle \leq \sigma \|x - a\|^2 \quad \text{for all } x \in A\big\} \\ &= \{p \in H : \text{there exists } t > 0 \text{ with } a \in P_A(a + tp)\}; \end{aligned}$$

- *modulus of nonconvexity* defined in [11] as

$$\begin{aligned} \gamma_A(\varepsilon) &:= \inf \left\{ \gamma \geq 0 : B_\gamma \left(\frac{a_1 + a_2}{2} \right) \cap A \neq \emptyset \right. \\ &\quad \text{for all } a_1, a_2 \in A \text{ with } \|a_1 - a_2\| \leq \varepsilon \} \,. \end{aligned} \tag{18}$$

By *curve* in H we mean any continuous mapping $\Gamma : [0, 1] \to H$, saying that the curve Γ *connects* points $\Gamma(0)$ and $\Gamma(1)$. The *image* of Γ is the set $\Gamma([0, 1]) = \{\Gamma(t) : t \in [0, 1]\}$. We say also that the curve Γ *lies in* a set $A \subset H$ and write $\Gamma \subset A$ if $\Gamma([0, 1]) \subset A$. The *length* of Γ is defined as

$$|\Gamma| = \sup \sum_{i=1}^{I} \|\Gamma(t_i) - \Gamma(t_{i-1})\|,$$

where the supremum is taken over all partitions $0 = t_0 < t_1 < \ldots < t_I = 1$ of the segment $[0, 1]$. Given a subset $A \subset H$ and points $x, y \in A$, a curve $\Gamma : [0, 1] \to A$, $\Gamma(0) = x$, $\Gamma(1) = y$, is called the *shortest curve in A connecting x and y* if the length of Γ is minimal among all curves with the same properties.

Let us recall two well-known notions, which will be used in sequel.

A set $A \subset H$ is said to be *α-paraconvex*, $\alpha \in [0, 1]$, if for each $r > 0$ and $x \in H$ with $d_A(x) \underline{< r}$ the convex hull of $A \cap \operatorname{int} B_r(x)$ is contained in the closed neighborhood $\overline{U_A(\alpha r)}$. So, each convex set is 0-paraconvex while the class of 1 -paraconvex sets is too large covering all the sets in a Hilbert space (see [45]). In terminology of Michael (see [48]) a set A is said to be *paraconvex* if it is α-paraconvex with some $0 \leq \alpha < 1$.

A set $A \subset H$ is called *contractible* if there exist a point $x_0 \in A$ and a continuous function $\theta : A \times [0, 1] \to A$ such that $\theta(x, 0) = x$ and $\theta(x, 1) = x_0$ for all $x \in A$.

Now we pass to the class of closed sets with the properties somehow symmetric to those studied in section "Strongly Convex Sets."

Uniform Weak Convexity

Definition 3.1 A subset $A \subset H$ is said to be **weakly convex** if there exists $R > 0$ (called **radius of weak convexity**) such that for all $x, y \in A$ with $0 < \|x - y\| < 2R$ the strongly convex segment $D_R(x, y)$ contains a point $a \in A$ different from both x and y.

Notice that a subset $A \subset H$ is convex iff it is weakly convex with every radius $R > 0$.

Our objective now is to establish some properties equivalent to the weak convexity similarly as it was done for the strong convexity (see Theorem 2.1). Although the equivalence of majority of the statements below is well known (see, e.g., [18, Theorems 4.1, 4.8]), the proofs placed here are new and based on density of the set of points admitting unique projection onto a given closed set. For the first time the latter result was obtained by Stechkin in [62] in uniformly convex Banach spaces (see also [29]).

Theorem 3.1 (Equivalent Characterizations of a Weakly Convex Set) *Given a nonempty closed set $A \subset H$ and a number $R > 0$ the following assertions are equivalent:*

(a) A is weakly convex with the radius R;

(b) for all $a \in \partial A$ and $p \in N^P(a, A)$, $\|p\| = 1$, one has

$$A \cap \operatorname{int} B_R(a + Rp) = \emptyset; \tag{19}$$

(c) given $a_1 \in \partial A$ and $p_1 \in N^P(a_1, A)$, $\|p_1\| = 1$, the inequality

$$\langle p_1, a_1 - a_2 \rangle \geq -\frac{1}{2R}\|a_1 - a_2\|^2 \tag{20}$$

holds for all $a_2 \in A$;

(d) for all $a_i \in \partial A$, $p_i \in N^P(a_i, A)$, $\|p_i\| \leq 1$, $i = 1, 2$, the inequality

$$\langle p_1 - p_2, a_1 - a_2 \rangle \geq -\frac{1}{R}\|a_1 - a_2\|^2$$

holds;

(e) if $x \in U_A(R)$ and a sequence $\{a_k\} \subset A$ satisfies the minimization condition $\|x - a_k\| \to d_A(x)$, then $\{a_k\}$ converges;

(f) the metric projection operator $x \mapsto P_A(x)$ is single-valued and continuous on the neighborhood $U_A(R)$;

(g) the distance function $d_A(\cdot)$ is continuously differentiable on $U_A(R) \setminus A$;

(h) the distance function $d_A(\cdot)$ is Fréchet differentiable on $U_A(R) \setminus A$;

(i) for all $x_i \in U_A(R)$ and $a_i \in P_A(x_i)$, $i = 1, 2$, one has

$$\|a_1 - a_2\| \leq \frac{R}{R - \max\{d_A(x_1), d_A(x_2)\}}\|x_1 - x_2\|; \tag{21}$$

(j) $\gamma_A(\varepsilon) \le R - \sqrt{R^2 - \frac{\varepsilon^2}{4}}$ *for all* $\varepsilon \in (0, R)$*;*

(k) *for any* $x, y \in A$ *with* $0 < \|x - y\| < 2R$ *there exists a unique shortest curve* Γ *in* A*, connecting points* x *and* y*, moreover* $\Gamma \subset D_R(x, y)$*;*

(l) *for any* $x, y \in A$ *with* $0 < \|x - y\| < 2R$ *there exists a curve* Γ *lying in* A*, connecting points* x *and* y *and such that*

$$|\Gamma| \le 2R \arcsin\left(\frac{\|x - y\|}{2R}\right).$$

Proof We prove the equivalence by the following scheme:

$$\begin{tikzcd}
(j) \arrow[r, leftrightarrow] & (a) \arrow[r, leftrightarrow] & (b) \arrow[r] \arrow[dl] & (e) \arrow[d] & (h) \arrow[l] \\
(k) \arrow[ur, leftrightarrow] \arrow[d, leftrightarrow] & (c) \arrow[r] & (d) \arrow[u] \arrow[d] & (f) \arrow[ul] \arrow[r] & (g) \arrow[u] \\
(l) & & (i) \arrow[ur] & &
\end{tikzcd}$$

$(b) \Rightarrow (c)$ Fix arbitrary $a_1 \in \partial A$, $p_1 \in N^P(a_1, A)$, $\|p_1\| = 1$, and $a_2 \in A$. By the statement (b) we have $a_2 \notin \operatorname{int} B_R(a_1 + Rp_1)$, i.e.,

$$R^2 \le \|a_1 + Rp_1 - a_2\|^2 = R^2 + 2R\langle p_1, a_1 - a_2\rangle + \|a_1 - a_2\|^2,$$

which yields the inequality (20).

$(c) \Rightarrow (d)$ Let us fix $a_i \in \partial A$, $p_i \in N^P(a_i, A)$, $\|p_i\| \le 1$, $i = 1, 2$. From (c) we have

$$2R\langle p_1, a_1 - a_2\rangle + \|a_1 - a_2\|^2 \ge 0, \qquad 2R\langle p_2, a_2 - a_1\rangle + \|a_2 - a_1\|^2 \ge 0.$$

Adding the above inequalities, we obtain the desired result.

$(d) \Rightarrow (b)$ Let $a \in \partial A$ and $p \in N^P(a, A)$, $\|p\| = 1$. By the definition of the proximal normal cone the number

$$\tau := \sup\{t \in (0, R] : a \in P_A(a + tp)\} \tag{22}$$

is positive. We have

$$A \cap \operatorname{int} B_\tau(a + \tau p) = \emptyset. \tag{23}$$

Let us prove that $\tau = R$, in which case the equality (19) holds. Indeed, suppose that $\tau < R$ and choose $r \in (\tau, \min\{2\tau, R\})$. We denote $x_0 := a + rp$ and $d_0 := d_A(x_0)$. Since $\|a + \tau p - x_0\| = r - \tau < \tau$, it follows that $x_0 \in \operatorname{int} B_\tau(a + \tau p)$, and by (23) we get $x_0 \notin A$. Consequently, $d_0 > 0$. The inequality $r > \tau$ together with (22) imply that $a \notin P_A(a + rp)$, i.e., $d_0 < r$. By Stechkin's theorem [62] there exist sequences $\{x_k\}$ and $\{a_k\}$ with $x_k \to x_0$ and $P_A(x_k) = \{a_k\}$ for all $k \in \mathbb{N}$. Bearing in mind that $d_A(x_k) \to d_A(x_0) = d_0 \in (0, r)$, we assume without loss of generality that

$d_A(x_k) \in (0, r)$ for all $k \in \mathbb{N}$. Denote $d_k := d_A(x_k)$, $p_k := \frac{x_k - a_k}{d_k}$, and $b_k := a_k - a$, $k \in \mathbb{N}$. Since $p \in N^P(a, A)$ and $p_k \in N^P(a_k, A)$, it follows from the assertion (d) that $R\langle p_k - p, a_k - a\rangle + \|a_k - a\|^2 \geq 0$. Multiplying by d_k after simple transformations we arrive at

$$R\langle x_k - x_0 - b_k + (r - d_k)p, b_k\rangle + d_k\|b_k\|^2 \geq 0 \quad \forall k \in \mathbb{N}. \tag{24}$$

On the other hand, it follows from (23) that $\|a + \tau p - a_k\| \geq \tau$ because $a_k \in A$, i.e., $\langle p, b_k\rangle \leq \frac{\|b_k\|^2}{2\tau} \leq \frac{\|b_k\|^2}{r}$. Combining this with (24), we obtain

$$R\langle x_k - x_0, b_k\rangle \geq \frac{(R - r)d_k}{r}\|b_k\|^2 \quad \forall k \in \mathbb{N}. \tag{25}$$

Taking into account that $\|b_k\| = \|a_k - a\| \geq \|x_0 - a\| - \|x_k - x_0\| - \|a_k - x_k\| = r - d_k - \|x_k - x_0\|$ and passing to the limit in (25), we arrive at $0 \geq \frac{(R-r)d_0}{r}(r - d_0) > 0$, which is a contradiction.

$(b) \Rightarrow (e)$ Let $x \in U_A(R)$ and a sequence $\{a_k\} \subset A$ satisfy the minimization condition $\|x - a_k\| \to d_A(x)$. We assume that $x \notin A$, since otherwise it is nothing to prove. By Stechkin's theorem [62] there exist sequences $\{x_k\}$ and $\{b_k\}$ with $x_k \to x$ and $P_A(x_k) = \{b_k\}$ for all $k \in \mathbb{N}$. Let us denote $d := d_A(x)$, $d_k := d_A(x_k)$. Since $d_k \to d > 0$, we assume without loss of generality that $d_k > 0$ for all $k \in \mathbb{N}$. Denoting $p_k := \frac{x_k - b_k}{d_k}$, by (b) we get $A \cap \operatorname{int} B_R(b_k + Rp_k) = \emptyset$. Hence $\|b_k + Rp_k - a_j\| \geq R$ or, in other terms, $\|x_k - a_j + (R - d_k)p_k\| \geq R$ for all $j, k \in \mathbb{N}$. Let us denote $q_j := \frac{x - a_j}{\|x - a_j\|}$. Bearing in mind that $x_k \to x$, $d_k \to d$ and $\|x - a_k\| \to d$, we get

$$\liminf_{k,j \to \infty} \|dq_j + (R - d)p_k\| \geq R. \tag{26}$$

Taking into account that $\|q_j\| = \|p_k\| = 1$, we have

$$\|dq_j + (R - d)p_k\|^2 = d^2 + (R - d)^2 + 2d(R - d)\langle p_k, q_j\rangle = R^2 - d(R - d)\|p_k - q_j\|^2.$$

Thus, it follows from (26) that $\limsup_{k,j\to\infty} \|p_k - q_j\| \leq 0$ because $d < R$ (recall that $x \in U_A(R)$). Consequently, $\lim_{k,j\to\infty} \|q_k - q_j\| = 0$. So, $\{q_k\}$ is a Cauchy sequence and therefore converges. Hence $\{a_k\}$ converges as well.

$(e) \Rightarrow (f)$ Fix any $x_0 \in U_A(R)$. Choose a sequence $\{a_k\} \subset A$ with $\|x_0 - a_k\| \to d_A(x_0)$. By the assertion (e) it converges to some $a_0 \in H$. Then clearly $a_0 \in P_A(x_0)$. If $a \in A$ is another projection of x onto A, then again by the assertion (e) we deduce that $a = a_0$. So, $P_A(x_0) = \{a_0\}$, i.e., the metric projection operator is single-valued on $U_A(R)$. Suppose that $x_k \to x_0 \in U_A(R)$, $P_A(x_k) = \{a_k\}$ for all $k \in \mathbb{N}$. We have $\|x_k - a_k\| = d_A(x_k) \to d_A(x_0)$ and hence $\|x_0 - a_k\| \to d_A(x_0)$. According to (e) we obtain $a_k \to a_0$, i.e., the metric projection operator is continuous on $U_A(R)$.

$(f) \Rightarrow (b)$ Let us prove first an auxiliary statement. Namely, suppose that $x \in U_A(R) \setminus A$, $a \in P_A(x)$ and $v \in H$ is such that $\langle x - a, v\rangle > 0$. Then

$$d_A(x + tv) > d_A(x) \tag{27}$$

for all sufficiently small $t > 0$. Indeed, the assertion (f) implies that for $t > 0$ small enough there exists $a_t \in A$, $a_t \to a$ as $t \to 0+$, such that $P_A(x + tv) = \{a_t\}$. Consequently,

$$d_A(x + tv) - d_A(x) \geq \|x + tv - a_t\| - \|x - a_t\| = t\left\langle \frac{x - a_t}{\|x - a_t\|}, v\right\rangle + o(t)$$

$$= t\left\langle \frac{x - a}{\|x - a\|}, v\right\rangle + o(t), \quad t \to 0+.$$

Taking into account that $\langle x - a, v\rangle > 0$, we have the inequality (27) whenever $t > 0$ is sufficiently small.

To prove (b) let us fix $a \in \partial A$ and $p \in N^P(a, A)$, $\|p\| = 1$. By the definition of the proximal normal cone

$$\tau := \sup\{t > 0 : a \in P_A(a + tp)\} > 0. \tag{28}$$

If $\tau \geq R$, then (19) obviously holds. Suppose that $\tau < R$ and set $y_0 := a + \tau p \notin A$. Passing to limit in the equality $d_A(a + tp) = t$ as $t \to \tau-$, we arrive at $d_A(a + \tau p) = \tau$, i.e., $a \in P_A(y_0)$. Then, due to (27) it follows that $d_A(y_0 + t^*p) > d_A(y_0) = \tau$ for all sufficiently small $t^* > 0$. One can choose $t^* > 0$ so small that $t^* < \tau$ and

$$\tau < d_A(y_0 + t^*p) < R. \tag{29}$$

Denote $r := d_A(y_0 + t^*p) \in (\tau, R)$ and consider the nonempty closed set

$$C := \{x \in H : d_A(x) \geq r\}.$$

Let us fix now $y \in U_A(r) \setminus A$, $c \in P_C(y)$ with $\|c - y\| < r$ and $a \in P_A(c)$. We claim that

$$y \in [a, c]. \tag{30}$$

Indeed, denote $q_1 = \frac{c-a}{\|c-a\|}$, $q_2 = \frac{c-y}{\|c-y\|}$ and assume that $q_1 \neq q_2$ (otherwise (30), clearly, holds). Notice that $0 < \|c - y\| < r = \|c - a\|$. Let us set $v := q_1 - q_2$. Since $\langle c - a, v\rangle = r\langle q_1, q_1 - q_2\rangle = r(1 - \langle q_1, q_2\rangle) > 0$, it follows from (27) that $d_A(c + tv) > d_A(c) = r$ and, therefore, $c + tv \in C$ for sufficiently small $t > 0$. Taking into account that $c \in P_C(y)$, we arrive at $\|c + tv - y\| \geq \|c - y\|$ for $t > 0$ small enough. Squaring this inequality and passing to limit as $t \to 0+$ we have $\langle c - y, v\rangle \geq 0$, i.e., $0 \leq \langle q_2, q_1 - q_2\rangle = \langle q_2, q_1\rangle - 1 < 0$. The contradiction completes the proof of (30).

By Stechkin's theorem [62] there exist sequences $\{y_k\}$ and $\{c_k\}$ with $y_k \to y_0$ and $P_C(y_k) = \{c_k\}$ for all $k \in \mathbb{N}$. Since $\|c_k - y_k\| = d_C(y_k) \to d_C(y_0) \leq \|y_0 + t^*p - y_0\| = t^* < \tau < r$ (see (29) and the definition of r above), we suppose without loss of generality that $\|c_k - y_k\| < r$ for all $k \in \mathbb{N}$. Due to the assertion (f) we have

$P_A(c_k) = \{a_k\}$ for some $a_k \in A$, $k \in \mathbb{N}$. According to (30) $y_k \in [a_k, c_k]$ and hence $P_A(y_k) = \{a_k\}$, $k \in \mathbb{N}$. Using once more the assertion (*f*) and bearing in mind that $P_A(y_0) = \{a\}$, we have $a_k \to a$, $k \to \infty$. Since $y_k \in [a_k, c_k]$, $\|c_k - a_k\| = r$, $y_k \to y_0$ and $\|y_k - a_k\| \to \|y_0 - a\| = d_A(y_0) = \tau$, it follows that $c_k \to c_0 := a + \frac{r}{\tau}(y_0 - a)$, and due to (*f*) we arrive at $P_A(c_0) = \{a\}$. Consequently, $\sup\{t > 0 : a \in P_A(a + tp)\} \geq r > \tau$, which contradicts (28).

(*f*) $\Rightarrow$ (*g*) According to the assertion (*f*) there exists a continuous function $a : U_A(R) \to A$ such that $P_A(x) = \{a(x)\}$ for all $x \in U_A(R)$. Let us show that $p(x) = \frac{x - a(x)}{\|x - a(x)\|}$ is the Fréchet derivative of the distance function $d_A(\cdot)$ at each $x \in U_A(R) \setminus A$. To this end fix arbitrary $x_0 \in U_A(R) \setminus A$. By definition of the distance, for all $x \in U_A(R)$ the inequalities

$$\begin{aligned} \|x - a\,(x)\| - \|x_0 - a\,(x)\| &\leq d_A(x) - d_A(x_0) \\ &\leq \|x - a\,(x_0)\| - \|x_0 - a\,(x_0)\| \end{aligned} \tag{31}$$

take place. On the other hand, due to the uniform differentiability of the norm in a Hilbert space out of zero, for any $\delta > 0$ we have

$$\lim_{x \to x_0} \sup_{a \in H,\ \|x_0 - a\| \geq \delta} \frac{\left| \|x - a\| - \|x_0 - a\| - \left\langle \frac{x_0 - a}{\|x_0 - a\|}, x - x_0 \right\rangle \right|}{\|x - x_0\|} = 0. \tag{32}$$

Combining (31) and (32), by continuity of the function $a(\cdot)$, we deduce that

$$\lim_{x \to x_0} \frac{d_A(x) - d_A(x_0) - \langle p(x_0), x - x_0 \rangle}{\|x - x_0\|} = 0.$$

So, $p(x_0)$ is the Fréchet derivative of the distance function at x_0. Since $p(\cdot)$ is continuous on $U_A(R) \setminus A$, the function $d_A\,(\cdot)$ is continuously differentiable on this set.

(*g*) $\Rightarrow$ (*h*) is immediate.

(*h*) $\Rightarrow$ (*e*) Let $p \in H$ be the Fréchet derivative of the distance function $d_A(\cdot)$ at $x_0 \in U_A(R) \setminus A$. Denote $d_0 := d_A(x_0)$. Let a sequence $\{a_k\} \subset A$ satisfy the minimization condition $\lim_{k \to \infty} \|x_0 - a_k\| = d_0$. For each $k \in \mathbb{N}$ we set $t_k := \max\left\{\sqrt{\|x_0 - a_k\| - d_0}, \frac{1}{k}\right\}$. Since $t_k \to 0$, there is no loss of generality to assume that $t_k < 1$. The definition of the Fréchet derivative implies that

$$\lim_{k \to \infty} \frac{d_A(x_0 + t_k(a_k - x_0)) - d_0 - t_k\langle p, a_k - x_0 \rangle}{t_k} = 0.$$

On the other hand, recalling that $a_k \in A$ we have

$$d_A(x_0 + t_k(a_k - x_0)) \leq \|x_0 + t_k(a_k - x_0) - a_k\| \leq (1 - t_k)(d_0 + t_k^2).$$

Hence, comparing the latter relations, we write

$$\liminf_{k\to\infty} \langle p, x_0 - a_k\rangle \geq \lim_{k\to\infty} \frac{d_0 - (1-t_k)(d_0 + t_k^2)}{t_k} = d_0.$$

Taking into account that $\lim_{k\to\infty} \|x_0 - a_k\| = d_0$, we arrive at

$$\liminf_{k\to\infty} \left\langle p, \frac{x_0 - a_k}{\|x_0 - a_k\|} \right\rangle \geq 1.$$

Consequently,

$$\limsup_{k\to\infty} \left\| p - \frac{x_0 - a_k}{\|x_0 - a_k\|} \right\|^2 = 2 - 2\liminf_{k\to\infty} \left\langle p, \frac{x_0 - a_k}{\|x_0 - a_k\|} \right\rangle \leq 0.$$

So, $\lim_{k\to\infty} \frac{x_0 - a_k}{\|x_0 - a_k\|} = p$, and $\{a_k\}$ converges to $x_0 - d_0 p$. Thus (e) is proved.

$(d) \Rightarrow (i)$ Let $x_i \in U_A(R)$ and $a_i \in P_A(x_i)$, $i = 1, 2$. Denote $r := \max\{d_A(x_1), d_A(x_2)\}$, $p_i := \frac{x_i - a_i}{r}$, $i = 1, 2$. Then $p_i \in N^P(a_i, A)$ and $\|p_i\| \leq 1$. The assertion (d) implies that $R\langle p_1 - p_2, a_1 - a_2\rangle + \|a_1 - a_2\|^2 \geq 0$. Consequently,

$$(R - r)\|a_1 - a_2\|^2 \leq R\langle x_1 - x_2, a_1 - a_2\rangle \leq R\|x_1 - x_2\|\,\|a_1 - a_2\|,$$

which yields (i).

$(i) \Rightarrow (f)$ Fix $x \in U_A(R)$. By Stechkin's theorem [62] there exist sequences $\{x_k\}$ and $\{a_k\}$ with $x_k \to x$ and $P_A(x_k) = \{a_k\}$ for all $k \in \mathbb{N}$. The assertion (i) implies that $\{a_k\}$ is a Cauchy sequence and, therefore, converges to some $a \in A$. Passing to the limit in $d_A(x_k) = \|x_k - a_k\|$, we arrive at $d_A(x) = \|x - a\|$, i.e., $a \in P_A(x)$. Using (i) once more, we deduce that $P_A(x)$ is a singleton depending continuously upon $x \in U_A(R)$.

For the further equivalences $(a) \Leftrightarrow (b)$, $(a) \Leftrightarrow (j)$, and $(a) \Leftrightarrow (k) \Leftrightarrow (l)$ we refer the reader, respectively, to [37, Theorem 6.1], [39, Lemma 1.5.3], and to [39, Theorem 1.14.2]. ■

Similarly to the strongly convex sets (see Remark 2.3) the estimations in the statements (c), (d), (i), (j), (l) are exact. Namely, they become equalities whenever A is the complement of an (open) ball of the radius R.

Remark 3.1 For a closed weakly convex set $A \subset H$ and any point $a \in \partial A$ the proximal normal con $N^P(a, A)$ coincides, in fact, with other normal cones such as limiting (Mordukhovich), Fréchet, Dini, Clarke ones (see, e.g., [53, Corollary 2.2] and [15, Sect. 6]). Consequently, in such case, $N^P(a, A)$ is closed and convex.

Remark 3.2 If for a closed subset $A \subset H$ and a number $R > 0$ there exist $r \in (0, R)$ such that inequality (21) holds for every $x_i \in U_A(r)$ and $a_i \in P_A(x_i)$, $i = 1, 2$, then A is weakly convex with radius R (see [3, Theorem 2.1]).

There is also a symmetric version of Remark 2.4.

Remark 3.3 If a set $A \subset H$ is weakly convex with the radius $R > 0$, then it is weakly convex with any radius $R_1 < R$ as well. Vice versa, if $A \subset H$ is closed and weakly convex with each radius $R_1 < R$ ($R > 0$ is fixed), then A is weakly convex with the radius R. It follows from the statement (b) of Theorem 3.1.

So, for any closed weakly convex set $A \subset H$ there exists the *maximal (or sharp) radius of the weak convexity*, which possesses some equivalent properties.

Theorem 3.2 *Let $A \subset H$ be closed and connected with* $\operatorname{diam} A < 2R$*. Then the following assertions are equivalent:*

(a) R is the maximal radius of weak convexity of A;
(b)

$$\lim_{\varepsilon \to 0+} \frac{\gamma_A(\varepsilon)}{\varepsilon^2} = \frac{1}{8R};$$

(c) for each $R_1 \in (0, R)$ the metric projection operator $P_A(\cdot)$ is Lipschitz continuous with the minimal constant $\frac{R}{R-R_1}$ on the neighborhood $U_A(R_1)$.

By the similar reasoning as in proof of Theorem 2.3, using equivalence $(a) \Leftrightarrow (b)$ of Theorem 3.1, we obtain the following result, analogous to [8, Theorem 2].

Theorem 3.3 (Hölderian Dependence of the Metric Projection Upon the Set) *Let $A_1 \subset H$ be a weakly convex set with radius $R > 0$; $A_2 \subset H$ be an arbitrary nonempty set; $h := h(A_1, A_2)$ be the Hausdorff-Pompeiu distance between these sets and $x \in U_{A_1}(R)$, $d := d_{A_1}(x)$. Then for the metric projections $a_i \in P_{A_i}(x)$ $(i = 1, 2)$ one has*

$$\|a_1 - a_2\| \le \sqrt{\frac{4Rhd}{R-d} + h^2}.$$

Let us give two more properties of (uniformly) weakly convex sets.

Theorem 3.4 ([39, Theorem 1.15.2]) *Let $A \subset H$ be closed and weakly convex with radius $R > 0$. If, moreover,* $\operatorname{diam} A < 2R$*, then A is contractible.*

Theorem 3.5 ([39, Theorem 3.3.2]) *Let $A \subset H$ be closed and weakly convex with radius $R > 0$. Assume that $A \subset B_r(c)$ with some $c \in H$ and $r \in (0, R)$. Then A is α-paraconvex with $\alpha = \frac{r}{R}$.*

The latter property shows that under some supplementary hypothesis a weekly convex set may be paraconvex (compare with [31, Example 1.2]).

There is also a connection with the smoothness of the boundary of a closed set (see [37, Theorem 7.2] and [38, Theorem 3]).

Definition 3.2 We say that a closed (nonconvex, in general) set $A \subset H$ is **smooth with constant** $L > 0$ if $A = \overline{\operatorname{int} A}$ and

$$\|p_1 - p_2\| \le L\|a_1 - a_2\| \qquad \forall a_1, a_2 \in \partial A,\ p_i \in N^P(a_i, A),\ \|p_i\| = 1,\ i = 1, 2.$$

So, the smoothness means uniqueness of the (unit) proximal normal and its lipschitzianity (with a constant L). For convex bounded sets with nonempty interior the latter property specifies the smoothness introduced in section "*Local Strong Convexity*."

Theorem 3.6 *Let a proper subset $A \subset H$ be such that $A = \overline{\operatorname{int} A}$, $R > 0$. Then the following assertions are equivalent:*

(a) A is smooth with constant $1/R$;
(b) both sets A and $H \setminus \operatorname{int} A$ are weakly convex with radius R.

Local Weak Convexity

Let us localize first the notion of the (uniform) weak convexity introduced above.

Definition 3.3 A set $A \subset H$ is said to be **locally weakly convex** with radius $R > 0$ if for any point $a \in \partial A$ there exists $\delta > 0$ such that $A \cap B_\delta(a)$ is weakly convex with radius R.

Due to the following statement any weakly convex set $A \subset H$ is locally weakly convex with the same radius.

Lemma 3.7 *Let $A \subset H$ be weakly convex with radius $R > 0$ and $C \subset H$ be strongly convex with the same radius. Then the intersection $A \cap C$ is weakly convex with the radius R.*

Proof Fix any $x, y \in A \cap C$ with $0 < \|x - y\| < 2R$. According to Definition 3.1 one can find a point $a \in A \cap D_R(x, y)$ different from both x and y. Using the assertion (b) of Theorem 2.1, we get $D_R(x, y) \subset C$. So, $a \in A \cap C \cap D_R(x, y)$, and the set $A \cap C$ is weakly convex with the radius R by definition. ■

In some cases the reverse implication also holds true.

Theorem 3.8 ([39, Theorem 1.16.1]) *Let $A \subset H$ be closed, connected, and locally weakly convex with radius $R > 0$. If, moreover,* $\operatorname{diam} A < 2R$, *then A is weakly convex with the radius R.*

Observe that the assumption $\operatorname{diam} A < 2R$ is essential in Theorem 3.8. For instance, an arc $A = \{(\cos\varphi, \sin\varphi) : \varphi \in [0, 2\pi - \delta]\}$, $\delta \in (0, \pi)$ is locally weakly convex with the radius $R = 1$ being not weakly convex with this radius.

Let us give one more result on local weak convexity.

Theorem 3.9 ([39, Theorem 1.16.2]) *Let $A \subset H$ be compact and locally weakly convex with radius $R > 0$. Then A is weakly convex (with a possibly smaller radius).*

The same pointwise constructions as in section "*Local Strong Convexity*" can be introduced for closed not necessarily convex sets taking just in mind that in such a case there is no duality between boundary points and the (proximal) normals.

In other words, in the definition of the rotundity modulus $\mathfrak{C}_A(\varepsilon, a, p)$ [see (9)] a vector p should be normalized by another way, say $\|p\| = 1$. So, using the same notions of the *modulus of rotundity* and of the scalar *curvature* as in section "*Local Strong Convexity*" (see (9) and (10)) and allowing $\varkappa_A(a, p)$ to have negative values, we arrive at the following definition (we believe that the assumption of boundedness of A and of nonemptiness of its interior are not relevant in this context).

Definition 3.4 Let $A \subset H$ be a nonempty closed set, $a \in \partial A$ and $p \in N^P(a, A)$, $\|p\| = 1$. We say that A is **weakly convex** at the point a w.r.t. the vector p if $\varkappa_A(a, p) > -\infty$.

Let us give also a geometric interpretation of the (negative) curvature, which is similar to Theorem 2.5.

Theorem 3.10 *Under the assumptions above let us suppose that* $\varkappa_A(a, p) < 0$. *Then*

$$\mathfrak{R}_A^-(a, p) := -\frac{1}{2\varkappa_A(a, p)} =$$

$$\liminf_{\substack{(\varepsilon, x, v) \to (0+, a, p) \\ x \in \partial A \\ v \in N^P(x, A) \cap \partial B_1(0)}} \sup\left\{r > 0 : A \cap B_\varepsilon(x) \cap \operatorname{int} B_r(x + rv) = \emptyset\right\}. \tag{33}$$

The proof is similar to that of Theorem 2.5, and we omit it.

Notice that the scaled constructions $\tilde{\varkappa}_A(a, p)$ (see (13)) and

$$\tilde{\mathfrak{R}}_A^-(a, p) := -\frac{1}{2\tilde{\varkappa}_A(a, p)} \tag{34}$$

have a sense also in negative case. Modifying a little bit the proof we arrive at the formula

$$\tilde{\mathfrak{R}}_A^-(a, p) =$$

$$\liminf_{\substack{(x, v) \to (a, p) \\ x \in \partial A,\ v \in N^P(x, A) \cap \partial B_1(0)}} \sup\left\{r > 0 : A \cap \operatorname{int} B_r(x + rv) = \emptyset\right\}. \tag{35}$$

Remark 3.4 It is easy to see the difference between two radii $\mathfrak{R}_A^-(a, p)$ and $\tilde{\mathfrak{R}}_A^-(a, p)$. Let, for instance,

$$A = \left\{(x, y) \in \mathbb{R}^2 : \left(\frac{x}{\alpha}\right)^2 + \left(\frac{y}{\beta}\right)^2 \geq 1\right\}$$

where $\alpha > \beta$, and $a = (0, \beta)$. Then, obviously, there is a unique normal direction to A at the point a (namely, $p = (0, -1)$), and for each $0 < r < \beta$ there exists

$\varepsilon > 0$ such that for every $x \in B_\varepsilon(a) \cap \partial A$ and $v \in N^P(x,A)$, $\|v\| = 1$, one has $A \cap \operatorname{int} B_r(x + rv) = \emptyset$, whereas this is not true for $r > \beta$. Consequently, $\tilde{\mathfrak{R}}^-_A(a,p) = \beta$ (hence $\tilde{\varkappa}_A(a,p) = -\frac{1}{2\beta}$). On the other hand, the real (negative) curvature of A at the point a is larger. In fact, simple calculations give $\varkappa_A(a,p) = -\frac{\beta}{2\alpha^2}$.

Notice, however, that $\varkappa_A(a,p) > -\infty$ iff $\tilde{\varkappa}_A(a,p) > -\infty$ provided $\varkappa_A(a,p) < 0$, so the weak convexity at the point a can be defined by means of $\tilde{\varkappa}_A(a,p)$ or by means of the radius (35) as well. To see this it is enough to show that $\mathfrak{R}^-_A(a,p) > 0$ implies $\tilde{\mathfrak{R}}^-_A(a,p) > 0$. Indeed, assuming the contrary we can choose $\varepsilon > 0$ and $\delta > 0$ such that for each $x \in B_\varepsilon(a)$ and $v \in B_\varepsilon(p) \cap N^P(x,A)$, $\|v\| = 1$,

$$A \cap B_\varepsilon(x) \cap \operatorname{int} B_\delta(x + \delta v) = \emptyset, \tag{36}$$

and, on the other hand, there exist sequences $\{x_n\} \subset \partial A$, $\{v_n\} \subset \partial B_1(0)$, $\{y_n\} \subset A$ with $v_n \in N^P(x_n,A)$, $x_n \to a$, $v_n \to p$, $y_n \in B_{1/n}\left(x_n + \frac{1}{n}v_n\right)$. Thus, $\|x_n - y_n\| \le \|x_n + \frac{1}{n}v_n - y_n\| + \frac{1}{n} \le \frac{2}{n} \to 0$ as $n \to \infty$. So, for sufficiently large n one has $y_n \in A \cap B_\varepsilon(x_n) \cap \operatorname{int} B_\delta(x_n + \delta v_n)$, $x_n \in B_\varepsilon(a)$, $v_n \in B_\varepsilon(p)$, which contradicts the equality (36).

Let us express the property of the (uniform) weak convexity in the sense of Definition 3.1 in terms of the curvatures (compare with Theorem 2.6).

Theorem 3.11 *Let $A \subset H$ be a closed (not necessarily convex, nor bounded) set and $R > 0$. Consider the following assertions:*

(a) the set A is weakly convex with the radius R;
(b) $\tilde{\varkappa}_A(a,p) \ge -\frac{1}{2R}$ for each $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$;
(c) $\varkappa_A(a,p) \ge -\frac{1}{2R}$ for each $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$.

Then $(a) \Leftrightarrow (b) \Rightarrow (c)$.

Proof $(a) \Rightarrow (b)$ Assuming that there exist $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$, with $\tilde{\varkappa}_A(a,p) < -\frac{1}{2R}$ (equivalently, $\tilde{\mathfrak{R}}^-_A(a,p) < R$) due to (35) we find a pair (x,v), $x \in \partial A$, $v \in N^P(x,A)$, $\|v\| = 1$, close to (a,p) such that the relation $A \cap \operatorname{int} B_r(x + rv) = \emptyset$ implies $r < R$. Thus $A \cap \operatorname{int} B_R(x + Rv) \ne \emptyset$ contradicting the weak convexity of A with the radius R (see Theorem 3.1 (b)).

$(b) \Rightarrow (c)$ is obvious.

$(b) \Rightarrow (a)$ Fix $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$, and consider two cases.

Case 1. $\tilde{\varkappa}_A(a,p) \ge 0$. From (13) and (9) we easily obtain that $\langle p, a - y\rangle \ge 0$ $\forall y \in A$. In other words, the set A is contained in the closed semispace P with the normal vector p whereas each ball $\operatorname{int} B_r(a + rp)$, $r > 0$, is contained in the open semispace $H \setminus P$. In particular, $A \cap \operatorname{int} B_{R_1}(a + R_1 p) = \emptyset$ for each $R_1 < R$.

Case 2. $-\frac{1}{2R} \le \tilde{\varkappa}_A(a,p) < 0$. According to (34) this is equivalent to the inequality $\tilde{\mathfrak{R}}^-_A(a,p) \ge R$. Let us fix arbitrary $R_1 < R$. Then by (35) we easily deduce that

$$\sup\{r > 0 : A \cap \operatorname{int} B_r(a + rp) = \emptyset\} > R_1.$$

Hence, there exists $r > R_1$ with $A \cap \operatorname{int} B_r(a + rp) = \emptyset$, and, consequently, $A \cap \operatorname{int} B_{R_1}(a + R_1 p) = \emptyset$.

Joining together the cases 1 and 2 we conclude (see Theorem 3.1 (b)) that the set A is weakly convex with the radius R_1. Finally, by using Remark 3.3 we arrive at the assertion (a). ■

Observe that the reverse implication $(b) \Leftarrow (c)$ is not true in general. This can be illustrated by the following simple example. Given $R > 0$ let us define the closed (connected) set A as the complement of the union of two (open) circles $\operatorname{int} B_{2R}(-R,0)$ and $\operatorname{int} B_{2R}(R,0)$. Then A is weakly convex with the *maximal radius* $\sqrt{3}R$ (it is realized at the points $x = \left(0, \pm\sqrt{3}R\right)$ w.r.t. the normal vectors $v = (0, \mp 1)$) while $\mathfrak{R}_A^-(a,p) \geq 2R$ (see (33)) for all $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$.

Combining the definitions of the scaled curvature (13) and of the modulus of rotundity (9) we write

$$-\tilde{\varkappa}_A(a,p) = \limsup_{\substack{(x,v)\to(a,p) \\ x\in\partial A,\, v\in N^P(a,A)\cap\partial B_1(0)}} \sup_{y\in A} \frac{\langle v, y-x\rangle}{\|y-x\|^2}. \tag{37}$$

Let now a subset $A \subset H$ be nonempty closed and weakly convex at each $a \in \partial A$ w.r.t. each unit normal vector p (see Definition 3.4). Then, bearing in mind that $(a,p) \mapsto -\tilde{\varkappa}_A(a,p)$ is upper semicontinuous real function (see (37)), by Michael's selections theorem [47] there exists a continuous nonnegative real function ψ_A such that $-\tilde{\varkappa}_A(a,p) \leq \psi_A(a,p)$ for all $a \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$. Thus, the inequality

$$\langle v, b-a\rangle \leq \psi_A(a,p)\,\|b-a\|^2$$

holds for all $a, b \in \partial A$ and $p \in N^P(a,A)$, $\|p\| = 1$. This property is equivalent, in fact, to the pointwise weak convexity.

In other words, A is weakly convex at each $a \in \partial A$ w.r.t. each normal $p \in N^P(a,A)$, $\|p\| = 1$, iff there exists a ball centered on the semiline $a+\lambda p$, $\lambda > 0$, with a radius continuously depending on a and p, which touches A at the point a only. This is nothing else than the *exterior sphere condition* (see section "Introduction").

If, in particular, $\psi_A(a,p)$ is majorized by a continuous function $\varphi(\cdot)$ depending only on $a \in \partial A$, then we arrive at the *φ-convexity*, or, in other terminology (not uniform) *prox-regularity*, *proximal smoothness*, etc., which was studied by many authors (see section "Introduction"). In particular, a series of equivalencies of Theorem 3.1 in such nonuniform case was proved in [53, Theorem 4.1] and in [19, Theorem 6.3]. The lack of uniformity here is due to the fact that well-posedness of the metric projection as well as (continuous) differentiability of the distance function take place not in some tube $U_A(R)$ around the set A but in an open neighborhood $U \supset A$ of arbitrary shape.

Balance between Weak and Strong Convexity

In this section we present some results, in which both weakly and strongly convex sets are involved. In such cases nonconvexity of a weakly convex set is compensated with the strong convexity of another one. The balance between radii of the weak and strong convexity here is relevant.

Separation of Sets with a Sphere and the Related Questions

We say that sets $A, C \subset H$ *may be separated with a sphere of radius* $\rho > 0$ if there exists a point $x \in H$ such that $A \cap \operatorname{int} B_\rho(x) = \emptyset$ while $C \subset B_\rho(x)$.

Given $A, C \subset H$ let us consider the so-called *nearest points problem*: to find points $a \in A$ and $c \in C$ with the minimal distance $\|a - c\|$. This problem is said to be *well posed* if any sequences $\{a_k\} \subset A$ and $\{c_k\} \subset C$ with $\|a_k - c_k\| \to \inf_{a \in A,\, c \in C} \|a - c\|$ converge. Note that if the nearest points problem is well posed and the sets A and C are closed, then the minimum $\min_{a \in A,\, c \in C} \|a - c\|$ is attained in a unique pair (a_0, c_0) with $a_0 \in A$ and $c_0 \in C$.

Theorem 4.1 *Given nonempty closed sets $A, C \subset H$ let us assume that C is strongly convex with a radius $r > 0$, and $A \subset H$ is weakly convex with a radius $R > r$. If, moreover,* $\operatorname{int} C \neq \emptyset$ *(equivalently, C is not a singleton, see Remark 2.2), then*

(a) the Minkowski sum $A + C$ is closed and weakly convex with the radius $R - r$;
(b) the sets A and C may be separated with a sphere of any radius $\rho \in [r, R]$ whenever $A \cap \operatorname{int} C = \emptyset$;
(c) the nearest points problem for sets A and C is well posed whenever $\inf_{a \in A,\, c \in C} \|a - c\| < R - r$.

The assertions (a) and (b) of Theorem 4.1 are proved in [39, Theorems 1.12.3, 1.12.4, 1.18.2] while (c) is obtained in [44, Theorem 4.2].

Let us give also a commutative equality for the Minkowski operations that can be treated as a kind of minimax property implying some results for saddle points and differential games (see [38, Theorems 5, 6]).

Theorem 4.2 ([38, Theorem 4]) *Let a proper set $A \subset H$ with $A = \overline{\operatorname{int} A}$ be weakly convex with a radius $R_1 > 0$ and such that $H \setminus \operatorname{int} A$ is weakly convex with a radius $R_2 > 0$. If $C_i \subset H$, $i = 1, 2$, are strongly convex sets with respective radii $r_i \in (0, R_i)$, then*

$$A + C_1 \overset{*}{-} C_2 = A \overset{*}{-} C_2 + C_1.$$

Continuity and Selections of the Intersection Mapping

Given a metric space (T, ρ_T) and a Banach space E let us consider two multivalued mappings $A : T \rightrightarrows E$ and $C : T \rightrightarrows E$ continuous w.r.t. the Hausdorff metric (2) and such that $A(t) \cap C(t) \neq \emptyset$ for all $t \in T$. The question is:

$$\begin{aligned}&\textit{under which hypotheses}\\ &\textit{the intersection mapping}\\ &t \mapsto F(t) := A(t) \cap C(t) \qquad (38)\\ &\textit{will keep the continuity}?\end{aligned}$$

There are simple examples showing that the multifunction $F(\cdot)$ may fail to be continuous even if both mappings A and C admit compact convex values in $E = \mathbb{R}^2$ and one of them is constant.

On the other hand, the problems concerned with continuity of the intersection mapping (as well as with existence of its continuous selection, i.e., of a continuous function $f : T \to E$ such that $f(t) \in F(t)$, $t \in T$) arise in various fields such as theory of dynamic systems including phase constraints and relaxation, differential inclusions, set-valued analysis, and optimal control theory. These and akin questions for mappings with convex values were investigated by Moreau [49], De Blasi and Pianigiani [26], Penot [50], and others. In particular, it was shown that in order to have the continuity of $F(\cdot)$ (in fact, the lower semicontinuity because the upper semicontinuity holds gratis) one can assume, in addition, that $F(t)$ has nonempty interior for all $t \in T$. Assuming that one of the mappings (say $C(\cdot)$) admits *uniformly convex* values (see [57]) it is possible to avoid the latter hypothesis, which seems to be too hard. This was done by Balashov and Repovš in [10] where they proved also existence of an uniformly continuous selection of $F(\cdot)$ (under the uniform continuity hypotheses for both $A(\cdot)$ and $C(\cdot)$). The same authors in [11] reduced the convexity hypothesis for values of $A(\cdot)$ to some conditions in terms of the modulus of nonconvexity. Furthermore, it turns out that the affirmative answer to the question (38) can be given whenever the sets $A(t)$, $t \in T$, are weakly convex, and $C(t)$, $t \in T$, are strongly convex (or vice versa). The first version of this result in a Hilbert space was proved in [39, Theorem 3.2.4] while recently it was generalized for Banach setting (see [41, Theorem 4.3]). Let us give the exact formulation of the Hilbert version. Here as usual (T, ρ_T) is an arbitrary metric space and H is a Hilbert one.

Theorem 4.3 *Suppose that multifunctions $A : T \rightrightarrows H$ and $C : T \rightrightarrows H$ with closed values are both continuous (uniformly continuous) w.r.t. the Hausdorff distance and such that for each $t \in T$ the set $C(t)$ is strongly convex with a radius $r > 0$ while $A(t)$ is weakly convex with a radius $R > r$. If, moreover, $F(t) := A(t) \cap C(t) \neq \emptyset$, $t \in T$, then the mapping $F : T \rightrightarrows H$ is continuous (uniformly continuous) and admits a continuous (respectively, uniformly continuous) selection $f(t) \in F(t)$, $t \in T$.*

An interesting application of the above result is the so-called *splitting problem* introduced by Repovš and Semenov in [58]. Given two nonempty sets $A, C \subset H$ let us consider their Minkowski sum $X = A + C$. The question is: *under which hypotheses do there necessarily exist continuous functions* $a : X \to A$ *and* $c : X \to C$ *such that* $a(x) + c(x) = x$ *for all* $x \in X$*?* As was shown in [9] the answer to this question is, in general, negative even for convex closed sets. However, in [11, Example 4.1] the authors proposed some hypotheses on the sets A and C (one of them may be just weakly convex) guaranteeing (uniformly) continuous splitting. These hypotheses involve rather complicate relation between the moduli $\delta_A(\varepsilon)$ and $\gamma_C(\varepsilon)$ (see (3) and (18), respectively). Nevertheless, a simplified version of their result can be immediately deduced from Theorem 4.3.

Theorem 4.4 (On the Splitting Problem) *Let* $C \subset H$ *be strongly convex with a radius* $r > 0$ *and* $A \subset H$ *be closed and weakly convex with a radius* $R > r$*. Then for* $X := A + B$ *there exist uniformly continuous functions* $a : X \to A$ *and* $c : X \to C$ *such that* $a(x) + c(x) = x$ *for all* $x \in X$*.*

Proof It suffices to apply Theorem 4.3 to the multifunctions $A(x) := A$ and $C(x) := x - C$. ■

Notice that apparently first continuous selections result for continuous multifunctions with weakly convex values appeared in [20, Theorem 3.1].

Recall finally that some investigations in past of continuous selections of mappings with values in spaces of integrable functions adjoin to our subject. For instance, in [35, 36] the authors proved existence of a continuous selection of the intersection of (a finite number of) multifunctions when one of them admits *decomposable values* (see [32]) while others are closed tubes w.r.t. some suitable seminorms around lower semicontinuous decomposable mappings. A generalization for a larger class of multifunctions (defined, moreover, on a paracompact space) was obtained later in [1].

Minimum Time Problem with a Constant Dynamics

In conclusion let us give one more application to well-posedness and regularity in a time-minimum control problem governed by some constant convex dynamics in a Hilbert space H. Namely, given a nonempty closed (not necessarily convex) set $A \subset H$ and a convex closed bounded set $F \subset H$ with $0 \in \operatorname{int} F$ we are interested in minimization of time $\tau > 0$ necessary to reach A from a point $a \in H$ by trajectories of the differential inclusion

$$\dot{x}(t) \in F.$$

Since due to the assumption $0 \in \operatorname{int} F$ the set A can be achieved in some finite time $\tau > 0$, denoting by $\mathfrak{T}_A^F(a)$ the infimum of such instants, we see that $\mathfrak{T}_A^F(a) < +\infty$

for all $a \in H$ while $\mathfrak{T}_A^F(a) = 0$ iff $a \in A$. This value function (or *time-minimum function*) can be represented as

$$\begin{aligned}\mathfrak{T}_A^F(a) &= \inf\{t > 0 : (a + tF) \cap A \neq \emptyset\} \\ &= \inf_{x \in A} \rho_F(x - a),\end{aligned}$$

where $\rho_F(\cdot)$ is the *Minkowski functional* associated with the set F (see (8)). Consider also the set of (boundary) points of A, which are attainable for the minimal time, i.e.,

$$P_A^F(a) := \left\{x \in A : \mathfrak{T}_A^F(a) = \rho_F(x - a)\right\}.$$

In a particular case $F = B_1(0)$ the function $\mathfrak{T}_A^F(\cdot)$ and the set $P_A^F(a)$ are, clearly, usual distance $d_A(\cdot)$ and the metric projection $P_A(a)$, respectively. We want to find conditions guaranteeing that the set $P_A^F(a)$ is a continuous singleton for all a from a neighborhood of A, and the function $\mathfrak{T}_A^F(\cdot)$ is sufficiently regular close to A (compare with Theorem 3.1 (f)–(h)). The first (uniform) result in this direction can be formulated in terms of strong and weak convexity. For details we refer to the proof of Theorem 6.1 (case 2) in [33, pp. 19–21] and to [34, Theorem 3.3].

Theorem 4.5 *Under all the standing assumptions above let us suppose the set A to be weakly convex with a radius $r > 0$ and the gauge F to be strongly convex with a radius $R > 0$. Then $P_A^F(a)$ is a singleton for all a belonging to the generalized tube*

$$U_A^F\left(\frac{r}{R}\right) := \left\{a \in H : \mathfrak{T}_A^F(a) < \frac{r}{R}\right\}. \tag{39}$$

Moreover, the mapping $a \mapsto P_A^F(a)$ is Hölder continuous on (39) with the exponent $1/2$.

Recently [42] the result of Theorem 4.5 was extended for Banach setting.

The Fréchet differentiability of the value function strongly depends on the fact that the “time-minimum projection” $P_A^F(a)$ is nonempty and single-valued. This question was studied in Banach setting, e.g., in [23]. In terms of a balance between strong and weak convexity we arrive at the following statement, which supplements Theorem 4.5.

Theorem 4.6 *In addition to the assumptions of Theorem 4.5, let us suppose that either the gradient $\nabla\rho_F(\cdot)$ is Hölder continuous on the boundary of F with an exponent $0 < \alpha \leq 1$, or the (unit) normal vector to the set A at boundary points is Hölder continuous with an exponent $0 < \alpha \leq 1$ (compare with Definition 3.2 where $\alpha = 1$). Then the value function $\mathfrak{T}_A^F(\cdot)$ is Fréchet differentiable on $U_A^F\left(\frac{r}{R}\right)$ and its derivative is Hölder continuous with the exponent $\frac{\alpha}{2-\alpha}$.*

This is a particular case of [34, Theorem 5.7], where the assumptions on weak and strong convexity are local as well as the regularity hypotheses on A and F can be mixed (switching from one boundary point of A to another).

Remark 4.1 Assuming in the framework of Theorem 4.5 the Hölder regularity of either F or A like in Theorem 4.6 with an exponent $0 < \alpha < 1$ increases the Hölder continuity of the mapping $a \mapsto P_A^F(a)$ up to the exponent $\frac{1}{2-\alpha}$ (see [34, Theorem 3.8]).

By using local and pointwise constructions of sections "*Local Strong Convexity*" and "*Local Weak Convexity*" one can obtain global versions of the above results. For instance, let us formulate the following well-posedness statement in finite dimensional setting.

Theorem 4.7 *Assume that a nonempty closed set $A \subset \mathbb{R}^n$ is weakly convex at each point $x \in \partial A$ w.r.t. each $p \in N^P(A,x)$, $\|p\| = 1$, and a closed bounded set F, $0 \in \operatorname{int} F$, is strongly convex at each $v \in \partial F$ w.r.t. each $q \in \mathfrak{J}_{F^0}(v)$. Then the mapping $a \mapsto P_A^F(a)$ is single-valued and continuous on an open neighborhood $\mathfrak{A}(A)$ of A given by*

$$\mathfrak{A}(A) := \left\{ a \in H : \liminf \left[\tilde{\varkappa}_F(v,q) + \mathfrak{T}_A^F(a)\, \tilde{\varkappa}_A\left(x, \frac{p}{\|p\|}\right)\right] > 0 \right\},$$

where the lower limit is taken in all the variables x, p, v, and q such that $x \in \partial A$, $p \in N^P(A,x)$, $\rho_F(x-a) \to \mathfrak{T}_A^F(a)$, $-p, q \in \partial F^0$, $p+q \to 0$, and $v \in \mathfrak{J}_F(q)$.

For details we refer to Theorem 6.2 in [33], to its proof and to remarks after that.

Acknowledgements Vladimir V. Goncharov financially supported by National Funds of Portugal through FCT—Fundação para a Ciência e a Tecnologia in the framework of the Project "UID/Mat/04674/2013 (CIMA)." Grigorii E. Ivanov supported by the Russian Foundation for basic research, project 16-01-00259-a.

References

1. S.M. Ageev, D. Repovš, On selection theorems with decomposable values. Topol. Meth. Nonlinear Anal. **15**, 385–399 (2000)
2. E. Asplund, R.T. Rockafellar, Gradients of convex functions. Trans. Am. Math. Soc. **139**, 443–467 (1969)
3. M.V. Balashov, Proximal smoothness of a set with the Lipschitz metric projection. J. Math. Anal. Appl. **406**, 360–363 (2013)
4. M.V. Balashov, Antidistance and antiprojection in the Hilbert space. J. Convex Anal. **22**, 521–536 (2015)
5. M.V. Balashov, M.O. Golubev, About the Lipschitz property of the metric projection in the Hilbert space. J. Math. Anal. Appl. **394**, 545–551 (2012)
6. M.V. Balashov, M.O. Golubev, Weak concavity of the antidistance function. J. Convex Anal. **21**, 951–964 (2014)
7. M.V. Balashov, G.E. Ivanov, On farthest points of sets. Math. Notes **80**, 163–170 (2006)
8. M.V. Balashov, G.E. Ivanov, Properties of the metric projection on weakly vial-convex sets and parametrization of set-valued mappings with weakly convex images. Math. Notes **80**, 461–467 (2006)

9. M.V. Balashov, D. Repovš, On the splitting problem for selections. J. Math. Anal. Appl. **355**, 277–287 (2009)
10. M.V. Balashov, D. Repovš, Uniform convexity and the splitting problem for selections. J. Math. Anal. Appl. **360**, 307–316 (2009)
11. M.V. Balashov, D. Repovš, Weakly convex sets and modulus of nonconvexity. J. Math. Anal. Appl. **371**, 113–127 (2010)
12. M.V. Balashov, D. Repovš, Uniformly convex subsets of the Hilbert space with modulus of convexity of the second order. J. Math. Anal. Appl. **377**, 754–761 (2011)
13. F. Bernard, L. Thibault, N. Zlateva, Characterization of proximal regular sets in super reflexive Banach spaces. J. Convex Anal. **13**, 525–559 (2006)
14. F. Bernard, L. Thibault, N. Zlateva, Prox-regular sets and epigraphs in uniformly convex Banach spaces: various regularity and other properties. Trans. Am. Math. Soc. **363**, 2211–2247 (2011)
15. M. Bounkhel, L. Thibault, On various notions of regularity of sets in nonsmooth analysis. Nonlinear Anal. **48**, 223–246 (2002)
16. A. Canino, On p-convex sets and geodesics. J. Differ. Equ. **75**, 118–157 (1988)
17. A. Canino, Local properties of geodesics on p-convex sets. Ann. Mat. Pura Appl. **159**, 17–44 (1991)
18. F.H. Clarke, R.J. Stern, P.R. Wolenski, Proximal smoothness and the lower-$\mathcal{C}^2$ property. J. Convex Anal. **2**, 117–144 (1995)
19. G. Colombo, V.V. Goncharov, Variational inequalities and regularity properties of closed sets in Hilbert spaces. J. Convex Anal. **8**, 197–221 (2001)
20. G. Colombo, V.V. Goncharov, Continuous selections via geodesics. Topol. Meth. Nonlinear Anal. **18**, 171–182 (2001)
21. G. Colombo, L. Thibault, Prox-regular sets and applications, in *Handbook of Nonconvex Analysis*, ed. by D.Y. Gao, D. Motreanu (International Press, Somerville, MA, 2010)
22. G. Colombo, P. Wolenski, Variational analysis for a class of minimal time functions in a Hilbert space. J. Convex Anal. **11**, 335–361 (2004)
23. G. Colombo, V.V. Goncharov, B.S. Mordukhovich, Well-posedness of minimal time problems with constant dynamics in Banach spaces. Set-Valued Var. Anal. **18**, 349–372 (2010)
24. L. Danzer, B. Grünbaum, V. Klee, Helly's theorem and its relatives, in *Convexity*, ed. by V. Klee. Proceedings of Symposia in Pure Mathematics, vol. 7 (American Mathematical Society, Providence, RI, 1963), pp. 101–180
25. F.S. De Blasi, J. Myjak, On a generalized best approximation problem. J. Approx. Theory **94**, 54–72 (1998)
26. F.S. De Blasi, G. Pianigiani, Remarks on Hausdorff continuous multifunction and selections. Comment. Math. Univ. Carol. **24**, 553–561 (1983)
27. E. De Giorgi, M. Degiovanni, A. Marino, M. Tosques, Evolution equations for a class of nonlinear operators. Atti Acad. Naz. Lincei Red. Cl. Sci. Fiz. Mat. Natur. **75**, 1–8 (1983)
28. M. Degiovanni, A. Marino, M. Tosques, Evolution equations with lack of convexity. Nonlinear Anal. **9**, 1401–1443 (1985)
29. M. Edelstein, On nearest points of sets in uniformly convex Banach spaces. J. Lond. Math. Soc. **43**, 375–377 (1968)
30. N.V. Efimov, S.B. Stechkin, Support properties of sets in Banach spaces and Chebyshev sets. Doklady Acad. Sci. USSR **127**, 254–257 (1959)
31. H. Federer, Curvature measures. Trans. Am. Math. Soc. **93**, 418–491 (1959)
32. A. Fryszkowski, Continuous selections for a class of non-convex multivalued maps. Stud. Math. **76**, 163–174 (1983)
33. V.V. Goncharov, F.F. Pereira, Neighbourhood retractions of nonconvex sets in a Hilbert space via sublinear functionals. J. Convex Anal. **18**, 1–36 (2011)
34. V.V. Goncharov, F.F. Pereira, Geometric conditions for regularity in a time-minimum problem with constant dynamics. J. Convex Anal. **19**, 631–669 (2012)
35. V.V. Goncharov, A.A. Tolstonogov, Joint continuous selections of multivalued mappings with nonconvex values, and their applications. Math. USSR Sb. **73**, 319–339 (1992)

36. V.V. Goncharov, A.A. Tolstonogov, Continuous selections of the family of nonconvex-valued mappings with a noncompact domain. Sib. Math. J. **35**, 479–494 (1994)
37. G.E. Ivanov, Weak convexity in the senses of Vial and Efimov-Stechkin. Izv. RAN. Ser. Mat. **69**, 35–60 (2005)
38. G.E. Ivanov, Weakly convex sets and their properties. Mat. Zametki **79**, 60–86 (2006)
39. G.E. Ivanov, *Weakly Convex Sets and Functions: Theory and Applications* (Fizmatlit, Moscow, 2006) (in Russian)
40. G.E. Ivanov, Farthest points and strong convexity of sets. Math. Notes **87**, 355–366 (2010)
41. G.E. Ivanov, Continuity and selections of the intersection operator applied to nonconvex sets. J. Convex Anal. **22**, 939–962 (2015)
42. G.E. Ivanov, Sharp estimates for the moduli of continuity of metric projections onto weakly convex sets. Izvestiya Math. **79**, 668–697 (2015)
43. G.E. Ivanov, Weak convexity of sets and functions in a Banach space. J. Convex Anal. **22**, 365–398 (2015)
44. G.E. Ivanov, M.S. Lopushanski, Well-posedness of approximation and optimization problems for weakly convex sets and functions. J. Math. Sci. **209**, 66–87 (2015)
45. V. Klee, Circumspheres and inner products. Math. Scand. **8**, 363–370 (1960)
46. J. Lindenstrauss, On the modulus on smoothness and divergent series in Banach spaces. Mich. Math. J. **10**, 241–252 (1963)
47. E. Michael, Continuous selections. I. Ann. Math. **63**, 361–382 (1956)
48. E. Michael, Paraconvex sets. Math. Scand. **7**, 372–376 (1959)
49. J.J. Moreau, Intersection of moving convex sets in a normed space. Math. Scand. **36**, 159–173 (1975)
50. J.-P. Penot, Preservation of persistence and stability under intersections and operations. I. Persistence. J. Optim. Theory Appl. **79**, 525–550 (1993)
51. R.R. Phelps, Convex functions, monotone operators and differentiability. Lecture Notes in Mathematics, vol. 1364 (Springer, Berlin, 1989)
52. A. Pliś, Uniqueness of optimal trajectories for non-linear control systems. Ann. Polon. Math. **29**, 397–401 (1975)
53. R.A. Poliquin, R.T. Rockafellar, L. Thibault, Local differentiability of distance functions. Trans. Am. Math. Soc. **353**, 5231–5249 (2000)
54. E.S. Polovinkin, On strongly convex sets. Phys. J. **2**, 43–59 (1996)
55. E.S. Polovinkin, Strongly convex analysis. Matem. Sb. **187**, 103–130 (1996)
56. E.S. Polovinkin, M.V. Balashov, *Elements of Convex and Strongly Convex Analysis* (Fizmatlit, Moscow, 2004) (in Russian)
57. B.T. Polyak, Existence theorems and convergence of minimizing sequences in extremum problems with restrictions. Sov. Math. **7**, 72–75 (1966)
58. D. Repovš, P.V. Semenov, Sections of convex bodies and splitting problem for selections. J. Math. Anal. Appl. **334**, 646–655 (2007)
59. Yu.G. Reshetnyak, On a generalization of convex surfaces. Matem. sbornik **40**(82), 381–398 (1956)
60. A. Shapiro, Existence and differentiability of metric projections in Hilbert spaces. SIAM J. Optim. **4**, 130–141 (1994)
61. I. Singer, *Abstract Convex Analysis*. Wiley Interscience Publication (Canadian Mathematical Society, New York, Toronto, 1997)
62. S.B. Stechkin, Approximate properties of sets in linear normed spaces. Rev. Math. Pures Appl. **8**, 5–18 (1963)
63. J.-P. Vial, Strong and weak convexity of sets and functions. Math. Ops. Res. **8**, 231–259 (1983)
64. A. Weber, G. Reissig, Local characterization of strongly convex sets. J. Math. Anal. Appl. **400**, 743–750 (2013)

Non-equilibrium Solutions of Dynamic Networks: A Hybrid System Approach

Scott Greenhalgh and Monica-Gabriela Cojocaru

Abstract Many dynamic networks can be analyzed through the framework of equilibrium problems. While traditionally, the study of equilibrium problems is solely concerned with obtaining or approximating equilibrium solutions, the study of equilibrium problems not in equilibrium provides valuable information into dynamic network behavior. One approach to study such non-equilibrium solutions stems from a connection between equilibrium problems and a class of parametrized projected differential equations. However, there is a drawback of this approach: the requirement of observing distributions of demands and costs. To address this problem we develop a hybrid system framework to model non-equilibrium solutions of dynamic networks, which only requires point observations. We demonstrate stability properties of the hybrid system framework and illustrate the novelty of our approach with a dynamic traffic network example.

Keywords Dynamic networks • Hybrid systems • Variational inequalities • Equilibrium problems

Introduction

Equilibrium problems are vastly applicable to many networks and their formulation has become fairly complex. To date, most studies of networks formulated as equilibrium problems are concerned with equilibrium solutions, where equilibrium is defined depending on the context of the problem (Wardrop [12–14], Nash/Cournot [10, 11, 16], market [3, 20], and physical/mathematical equilibrium [18, 19]). However, the study of equilibrium problems not in equilibrium provides information absent from the analysis of equilibrium solutions. For instance, in a

S. Greenhalgh (✉)
Queen's University, Kingston, ON, Canada
e-mail: scottyg3@gmail.com

M.-G. Cojocaru
Univerity of Guelph, Guelph, ON, Canada
e-mail: mcojocar@uoguelph.ca

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_13

classic dynamic traffic network equilibrium problem, non-equilibrium solutions describe the adjustment of flows on a network in response to disturbances (like lane closures, accidents, or road construction) in link costs or demand.

One approach to study non-equilibrium solutions is through a class of parametrized projected differential equations, called double layered dynamics (DLD), [5, 10, 11] which tracks the adjustment of demands over time. To do this, DLD requires observing distributions of network information at all times, which in reality may be difficult or impossible to obtain. To overcome this obstacle, we develop a hybrid system framework to model non-equilibrium solutions. Through the hybrid system framework we extend the association between dynamic networks and projected differential equations, through their common connection with variational inequalities (VI). The result is a hybrid system version of non-equilibrium solutions of dynamic networks, which advantageously require only an observed point of network information.

Preliminaries

To begin, we present the foundations for modeling non-equilibrium solutions. We define the frameworks of equilibrium problems for static networks, as described by VI and projected differential equations (PrDE). Next, we provide analogous definitions for the frameworks of equilibrium problems for dynamic networks, as defined by evolutionary variational inequalities (EVI) and DLD.

Equilibrium Problems: Variational Inequalities and Projected Differential Equations

Since their introduction in the 60s [18, 19], VI problems have been extensively used in the study of Wardrop, Nash, Walras, Cournot and mathematical physics equilibrium problems [2, 9, 12, 13]. As such, we consider a VI on a Euclidean space of arbitrary dimension X, with a non-empty, closed, and convex set $K \subset X$, and a mapping $F : K \to X$ is given by:

Definition 1 Variational inequality problem [22].

$$\text{find } x^* \in K \text{ so that } \langle F(x^*), y - x^* \rangle \geq 0, \ \forall y \in K$$

The set of points $x^* \in K$ satisfying the inequality above is called the solution set of the VI, which we denote by $SOLVI(F, K)$.

There is an important connection between VI and PrDE [1, 7, 21], where a PrDE on a non-empty, closed, and convex set $K \subset X$, with a Lipschitz continuous mapping $F : K \to X$ is defined as:

Definition 2 Projected differential equation [7].

$$\frac{dx(\tau)}{d\tau} = P_{T_K(x(\tau))}(-F(x(\tau))), \quad x(0) \in K, \tag{1}$$

where the set $T_K(x) = \overline{\bigcup_{\delta>0} \frac{1}{\delta}(K - x)}$ represents the tangent cone at the point x to K and the mapping $P_{T_K(x)}(\cdot)$ is the closest element mapping from X to the set $T_K(x) \in X$.

The important connection between a VI defined by Definition 1 and a PrDE defined by Definition 2 is the correspondence between solutions $x^* \in SOLVI(F, K)$ and the critical points $P_{T_K(x^*)}(-F(x^*)) = 0$. As such, when some mild conditions are satisfied [5, 9], it follows that

$$x^* \in SOLVI(F, K) \text{ if and only if } P_{T_K(x^*)}(-F(x^*)) = 0.$$

Dynamic Equilibrium Problems: Evolutionary Variational Inequality and Double Layered Dynamics Problems

Akin to the relationship between VI and PrDEs, there is a similar connection between an EVI [3, 14, 15] and DLD. In essence, an EVI represents a dynamic network, or an equilibrium problem that evolves with time, and can be viewed as an infinite dimensional VI. A similar view can also be taken with the connection between PrDE and DLD.

Formally, we take an EVI to be defined on a Hilbert space of arbitrary dimension $X := L^2([0, T], \mathbb{R}^q)$ to be given by:

Definition 3 Evolutionary Variational Inequality [4, 6].

$$\text{find } x^* \in \mathbb{K} \text{ so that } \int_0^T \langle F(x^*(t)), v(t) - x^*(t)\rangle dt \geq 0, \;\; \forall v \in \mathbb{K}, \tag{2}$$

where $\ll \phi, x \gg := \int_0^T \langle \phi(x)(t), x(t)\rangle dt$ is the Hilbert space inner-product, with ϕ and $x \in X$ and $F : \mathbb{K} \to X$ is a Lipschitz continuous mapping. The set of points $x^* \in \mathbb{K}$ that satisfy the EVI is called the solution set of the inequality, which we denote as $SOLEVI(F, \mathbb{K})$.

For simplicity, the constraint (feasible) set $\mathbb{K} \subset X$ of an EVI is taken to be

$$\mathbb{K} = \Big\{x \in X \mid \lambda(t) \leq x(t) \leq \mu(t),\, A(t)x(t) = \rho(t), \text{ for a.a.} t \in [0, T]\Big\}, \tag{3}$$

where $\lambda, \mu \in L^2([0,T], \mathbb{R}^q)$, $A \in L^2([0,T], \mathbb{R}^{l\times q})$ and $\rho \in L^2([0,T], \mathbb{R}^l)$. Such a set is typically assumed to be closed, convex, and bounded in $L^2([0,T], \mathbb{R}^q)$, and therefore appropriate convexity conditions on the functions λ, μ, and ρ are required.

Definition 4 Double layered dynamic [5, 9]. Let $F : \mathbb{K} \to X$ be a Lipschitz continuous mapping, where $\mathbb{K}$ is given by (3). Then a double layered dynamic is given by:

$$\frac{dx(\cdot,\tau)}{d\tau} = P_{T_{\mathbb{K}}(x(\cdot,\tau))}(-F(x(\cdot,\tau))), \text{with } x(\cdot,0) \in \mathbb{K} \subset X, \tag{4}$$

where $x \in AC([0,\infty), \mathbb{K})$.

Similar to a VI and PrDE, there is a connection between the solution set of an EVI defined by Definition 3 and the critical points of a DLD defined by Definition 4. That is [8, 9],

$$x^*(\cdot) \in SOLEVI(F, \mathbb{K})\text{if and only if } P_{T_{\mathbb{K}}(x^*(\cdot))}(-F(x^*(\cdot))) = 0.$$

Hybrid Systems

A hybrid system is a dynamical system composed of continuous and discrete dynamics [23]. Often, hybrid systems combine multiple systems of differential equations (the continuous dynamic) through a series of jump rules (the discrete dynamic), which take place at time instances called event-times [23]. To construct the trajectory of a hybrid system from the continuous and discrete dynamics, one starts from an initial point and continuously evolves in accordance with a system of differential equations until an event-time occurs. At the first event-time the continuous evolution of the hybrid system temporarily pauses, and the model states and parameters are updated according to the specified jump rule. After updating the model states and parameters, the hybrid system starts to continuously evolve again according to the (potentially new) system of differential equations, and proceeds until the next event-time occurs. This process repeats itself until a desired time is reached.

Any differential equation can be used to describe the continuous dynamic of a hybrid system. For a hybrid system composed of states, $(x_1, x_2, \ldots, x_n) := x$, parameters $(\theta_1, \theta_2, \ldots, \theta_m) := \theta$, and a function $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, the continuous dynamic can be stated as:

Definition 5 The continuous dynamic.

$$\frac{dx}{dt} = F(t, x; \theta), \text{with } x_0 = x(t_0) \in \mathbb{R}^n. \tag{5}$$

The discrete dynamic of a hybrid system is a jump rule [23] that updates the model's states, and parameter values and functional structure upon the occurrence of an event-time. Consequently, at event-times t_j, in accordance with the jump rules $G_j : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ (for the model's states), and $H_j : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ (for the model's parameter values), we determine new states and parameter values of the model. Thus the discrete dynamic can be stated as:

Definition 6 The discrete dynamic.

$$G_j(t_j^-, x(t_j^-), \theta) \to x(t_j^+), \tag{6}$$

and

$$H_j(t_j^-, x(t_j^-), \theta^j) \to \theta^{j+1}. \tag{7}$$

Note the $-$ and $+$ superscripts are used to distinguish between model states at pre and post event-times.

Given the continuous dynamic (5) and discrete dynamic (6)–(7) we now construct the hybrid system. To construct the hybrid system, as with a standard system of differential equations, we require an initial (observed) point of information $x(0)$ as well as initial parameter values θ^1 and a time interval $[0, T]$. From the initial conditions, parameter values, and corresponding continuous dynamic, we proceed to compute the evolution of model states through

$$\frac{dx}{dt} = F_1(t, x; \theta^1), x_0 = x(t_0) \in \mathbb{R}^n, t \in [t, t_1^-],$$

up until the first event-time t_1. At the first event-time we stop the continuous evolution of the model. The model, in accordance with the jump rule (6)–(7) undergoes a change in state and an update of parameters:

$$G_1(t_1^-, x(t_1^-), \theta^1) \to x(t_1^+),$$

and

$$H_1(t_1^-, x(t_1^-), \theta^1) \to \theta^2.$$

With the state, parameter values, and functional structure updated, the evolution of the continuous dynamic starts again. The continuous dynamic

$$\frac{dx}{dt} = F_2(t, x; \theta^2), x_0 = x(t_1^+) \in \mathbb{R}^n, t \in [t_1^+, t_2^-]$$

is followed until the next event-time t_2, where we then once again undergo an update in model states G_2, parameters H_2, and functional structure F_3. This procedure is repeated until the end of the time interval $[0, T]$ is reached. Formally, we have that a hybrid system is given by the following:

Definition 7 Hybrid systems. For a given uniform partition of $[0, T]$ into segments $[t_j, t_{j+1}]$, we have that for $t \in [t_j^+, t_{j+1}^-)$ we evolve according too

$$\frac{dx}{dt} = F_j(t, x; \theta^j), \text{with } x_0 = x(t_j^+) \in K_{j+1}, \tag{8}$$

and when $t = t_j^-$,

$$x(t_j^+) : +G_j(t_j^-, x(t_j^-), \theta^j), \tag{9}$$

and

$$\theta^{j+1} := H_j(t_j^-, x(t_j^-), \theta^j) \tag{10}$$

where, once again the $-$ and $+$ superscripts are used to distinguish between model states at pre and post event-times.

Non-equilibrium Solutions of Dynamic Networks

The difference in approach between DLD and hybrid system non-equilibrium solutions of dynamic networks is evident under the context of a dynamic traffic network problem:

1. A DLD non-equilibrium solution can be seen as an external view of the entire traffic network, where observed information on the evolution of the entire traffic flow across all links can be provided.
2. A hybrid system non-equilibrium solution can be seen as an internal view of the traffic network, with knowledge of the network structure (links, nodes and equilibrium), but only current information about immediately viewable traffic (point observations) can be provided.

Definition 8 DLD non-equilibrium solutions. From the association between $SOLEVI(F, \mathbb{K})$ and the critical points of a DLD, a DLD non-equilibrium solution is given by

$$\frac{dx(\cdot, \tau)}{d\tau} = P_{T_{\mathbb{K}}(x(\cdot,\tau))}(-F(x(\cdot, \tau))), \text{with } x(\cdot, 0) \in \mathbb{K} \subset L^2([0, T], \mathbb{R}^q), \tag{11}$$

where the mapping F and constraint set $\mathbb{K}$ are taken as in Definition 3 and Eq. (3), respectively, and $x(\cdot, \tau) \neq x^*$.

Definition 9 Hybrid system non-equilibrium solutions. To construct a non-equilibrium solution from a hybrid system, we define the continuous and discrete dynamic related to the dynamic equilibrium problem. We consider a hybrid system non-equilibrium solution to be composed of a series of jump rules that connect a series of projected differential equation. Each projected differential equation is defined on the set K_j, where

$$K_j = \mathbb{K}|_{t_j} \text{ for each event-time } t_j. \tag{12}$$

Thus, the continuous dynamic for $t \in [t_{j-1}, t_j^-)$ of the hybrid system is

$$\frac{dx}{dt} = F(t, x; \theta) := P_{T_{K_{j+1}}(x)}(-F(x)), , x(t_j^-) \in K_j,$$

and the discrete dynamic is

$$x(t_{j-1}^+) = G_j(t_{j-1}^-, x(t_{j-1}^-)) \text{ where } G_j : \mathbb{R} \times K_{j-1} \to K_j.$$

To analyze the stability properties of hybrid system non-equilibrium solutions we define the following:

Definition 10 Hybrid system trajectory. For a uniform division Δ of $[0, T]$, a hybrid system trajectory $HS_\delta : [0, T] \to \mathbb{R}$ is defined as:

$$HS_\delta(t) = \begin{cases} x_0(0^-) & t = 0^-, \\ x_0(0^+) = G_1(0^-, x_0(0^-)) & t = 0^+, \\ x(t) & t \in [t_0^+, t_1^-), \\ x(t_1^+) = G_2(t_1^-, x(t_1^-)) & t = t_1^+, \\ \vdots & \vdots \\ x(t) & t \in [t_{N-1}^+, T). \end{cases}$$

where $N = \frac{T}{\delta}$.

Definition 11 The sequence of hybrid system trajectories. For all $t \in [0, T]$ we denote $\{HS_{\delta_m}\}_m$ as the sequence of hybrid trajectories with uniform division Δ_m that consist of m divisions of length $\frac{T}{\delta_m}$.

Definition 12 The feasible sets of hybrid system trajectories. For all $t \in [0, T]$ and uniform divisions Δ, we denote the feasible set of hybrid trajectories as:

$$\mathbb{K}_\delta(t) = \begin{cases} K_0 & t = 0^-, \\ K_1 & t \in [0^+, t_1^-), \\ \vdots & \vdots \\ K_N & t \in [t_{N-1}^+, T]. \end{cases}$$

where $N = \frac{T}{\delta}$.

From these definitions, for almost all $t \in [0, T]$, it follows that

$$\mathbb{K}_\delta(t) = \{\lambda_\delta(t) \le x(t) \le \mu_\delta, A_\delta(t)x(t) = \rho_\delta(t)\},$$

where

$$\lambda_\delta(t) = \begin{cases} \lambda|_{0^-} & t = 0^- \\ \lambda|_{t_1} & t \in [t_0^+, t_1^-) \\ \vdots & \vdots \\ \lambda|_T & t \in [t_{N-1}^+, T] \end{cases}, \mu_\delta = \begin{cases} \mu|_{0^-} & t = 0^- \\ \mu|_{t_1} & t \in [t_0^+, t_1^-) \\ \vdots & \vdots \\ \mu|_T & t \in [t_{N-1}^+, T] \end{cases},$$

$$A_\delta = \begin{cases} A|_{0^-} & t = 0^- \\ A|_{t_1} & t \in [t_0^+, t_1^-) \\ \vdots & \vdots \\ A|_T & t \in [t_{N-1}^+, T] \end{cases}, \quad \text{and} \quad \rho_\delta = \begin{cases} \rho|_{0^-} & t = 0^- \\ \rho|_{t_1} & t \in [t_0^+, t_1^-) \\ \vdots & \vdots \\ \rho|_T & t \in [t_{N-1}^+, T] \end{cases}.$$

While there are many possible jump rules, imposing some physical properties on the jump rule will ensure that the discrete dynamic does not destabilize the system. As such, we consider jump rules that satisfy the following properties:

1. Jump rules G_j map equilibrium points to equilibrium points,

$$G_j(t_j^-, x^*(t_j)) = x^*(t_{j+1}), \tag{13}$$

   and
2. jump rules do not increase the distance from equilibrium points,

$$\|x(t_j^-) - x^*(t_j)\| \ge \|G_j(t_j^-, x(t_j^-)) - G_j(t_j^-, x^*(t_j))\| = \|x(t_j^+) - x^*(t_{j+1})\|. \tag{14}$$

From the definition of the feasible set of a hybrid trajectory, it follows that on any sub-interval $[t_j^+, t_{j+1}^-)$ that the critical point of the continuous dynamic $x^* \in SOLVI(F, K_j)$. Thus, with the conditions (13)–(14), we can approximate the equilibrium curve with

$$x^*_{\delta_m} = \begin{cases} x * (0)|_{0^-} & t = 0^-, \\ x * (t_1)|_{t_1} & t \in [0^+, t_1^-), \\ \vdots & \vdots \\ x * (t_N) & t \in [t_{N-1}^+, T]. \end{cases}$$

Given the definitions on hybrid system non-equilibrium solutions, we can now state the following stability result:

Theorem 1 *Stability of the hybrid systems non-equilibrium solution. If the mapping F from a hybrid system non-equilibrium solution is strongly pseudomonotone of degree $\alpha < 2$ with constant η (Appendix 1), and the jump rules are given by*

$$G_i(t_i^-, x_\delta(t_{j+1}^-)) = P_{K_{j+1}}(x_\delta(t_{j+1}^-) - x_\delta^*(t_j^-) + x_\delta^*(t_{j+1}^-)),$$

then δ can be selected so that for some time $t^* \in [0, T]$,

$$\|x_\delta - x_\delta^*\|_{L^2([t^*,T],\mathbb{R}^q)} < \epsilon \text{ for any } \epsilon > 0.$$

Furthermore, δ can be selected so that the hybrid system trajectory will converge to a curve arbitrarily close to the equilibrium curve after time t^*, where

$$t^* \geq \frac{2}{(2-\alpha)\eta)} \|x_\delta(0^-) - x_\delta^*(0)\|^{2-\alpha}.$$

For the proof see Appendix 3.

A Dynamic Traffic Network Example

To illustrate the DLD and hybrid system non-equilibrium solutions we consider a dynamic traffic network consisting of a single O/D pair of nodes with two direct links. The feasible set of the dynamic traffic network is taken to be

$$\mathbb{K} = \{x \in L^2([0, 110], \mathbb{R}^2) | 0 \leq x_1 \leq 120, 0 \leq x_2 \leq 120, x_1 + x_2 = \rho\},$$

and the travel demand function,

$$\rho = \begin{cases} 61 - \frac{1}{4}t(15-t) & 0 \leq t \leq 15, \\ 61 & 15 < t \leq 20, \\ 3t + 1 & 20 < t \leq 40, \\ 121 & 40 < t \leq 91, \\ 212 - t & 91 < t \leq 110. \end{cases}$$

The user cost function on the links is taken as

$$F(x) = (2\sqrt{x_1 - x_1^*} + x_2 - x_2^*, x_2 - x_2^*)^T,$$

where the equilibrium flows are given by

$$x_1^*(t) = \begin{cases} \frac{1}{2}(\rho - 91) \text{ if} 0 \leq \frac{1}{2}(\rho - 91) \text{ and } \frac{1}{2}(\rho + 91) \leq 100, \\ \\ 0 \quad \text{o.w.} \end{cases},$$

and

$$x_2^*(t) = \begin{cases} \frac{1}{2}(\rho + 91) \text{ if} 0 \leq \frac{1}{2}(\rho - 91) \text{ and } \frac{1}{2}(\rho + 91) \leq 100, \\ \\ \rho \qquad \text{o.w.} \end{cases}.$$

DLD Non-equilibrium Solution

With initial distributions of travel demand across the links given by

$$x_1(t, 0) = \frac{4}{5}\rho, \text{ and } x_2(t, 0) = \frac{1}{5}\rho,$$

user cost function F and constraint set $\mathbb{K}$, the DLD non-equilibrium solutions gradually converge to the equilibrium (Fig. 1). Importantly, the convergence to the equilibrium is a result of the mapping F being strongly pseudomonotone of degree $\alpha = \frac{3}{2}$ with constant $\eta = 2^{1/4}$ (Appendix 2). This property of F ensures the stability of the dynamic traffic network, as it implies that any disturbance eventually dampens out by time [10],

$$\tau^* \leq \frac{\|x(\cdot, 0) - x^*\|^{2-\alpha}}{(2-\alpha)\eta} \approx 17.3.$$

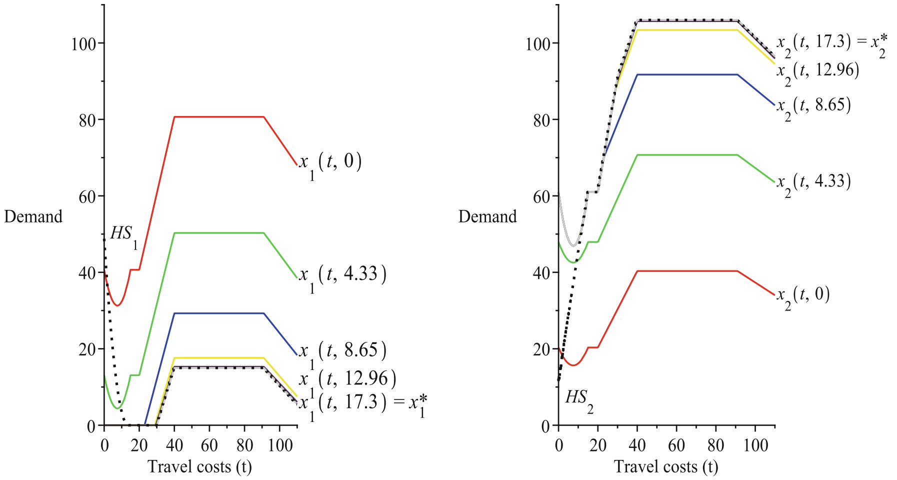


Fig. 1 Non-equilibrium solutions adjusting back to the traffic network equilibrium in a finite duration of time. For the demand on each link, the DLD non-equilibrium solutions consist of the *red, green, blue, yellow and violet curves*, the hybrid system non-equilibrium solution are the *black dotted curves*, and equilibrium solutions are the *grey curves*

HS Non-equilibrium Solution

For hybrid system non-equilibrium solutions we consider the initial conditions $x_1(0) = \frac{4}{5}\rho(0) = 48.8, x_2(0) = \frac{1}{5}\rho(0) = 12.2$, together with a uniform partition of the interval $[0, T]$ into segments $[t_j, t_{j+1})$ such that $|t_{j+1} - t_j| = \delta$. In addition, we require the definition of jump rules G_j, which are taken as

$$G_i(t_i^-, x_\delta(t_{j+1}^-)) = P_{K_{j+1}}(x_\delta(t_{j+1}^-) - x_\delta^*(t_j^-) + x_\delta^*(t_{j+1}^-)), \tag{15}$$

where the constraint sets K_j are defined by (12).

From this construction, an interpretation of hybrid system non-equilibrium solution is that of a non-equilibrium solution that crosses (or travels along) the DLD non-equilibrium solutions (Fig. 1). With this in mind, it is possible to show that both frameworks share desirable properties.

As such, the hybrid system non-equilibrium solution can converge to x_δ^* by $t^* \approx 13.9$ due to the strongly pseudomonotone of degree $\frac{3}{2}$ of F (Theorem 1). Consequently, it follows that one can select δ sufficiently small so that the hybrid system non-equilibrium solution converges arbitrarily close to the equilibrium of the dynamic traffic network on $[t^*, T]$ (Figs. 2 and 3). In other words, hybrid system non-equilibrium solutions, like their DLD non-equilibrium solution counterparts, will eventually dampen out any disturbance.

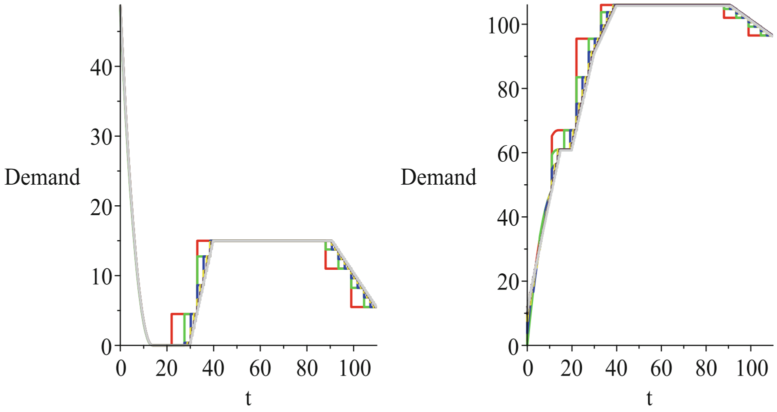


Fig. 2 HS non-equilibrium solutions with 10 jumps (*red*), 20 jumps (*blue*), 40 jumps (*green*), 80 jumps (*yellow*), 160 jumps (*violet*), and 320 jumps (*grey*) for link 1 (*left*) and link 2 (*right*)

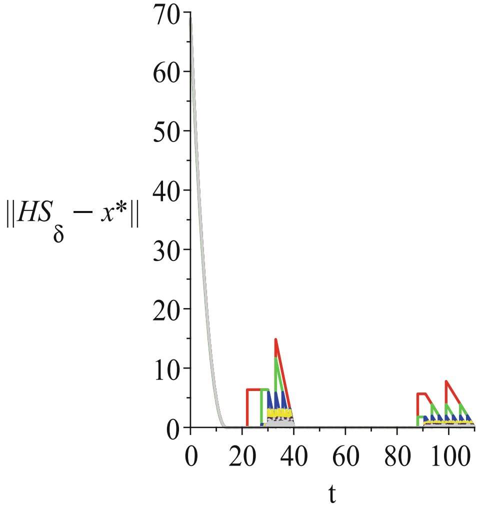


Fig. 3 Euclidean distance of HS non-equilibrium solution from the equilibrium solution with 10 jumps (*red*), 20 jumps (*blue*), 40 jumps (*green*), 80 jumps (*yellow*), 160 jumps (*violet*), and 320 jumps (*grey*)

Discussion

We developed a hybrid system framework for modeling non-equilibrium solutions of dynamic networks. Our framework provides an alternative approach to model the adjustment of dynamic networks in response to disturbances. The primary advantage of hybrid system non-equilibrium solutions, in comparison to DLD non-equilibrium solutions, is the reduction of the requirement to track distributions of information across the entire network to that of point observations.

We illustrated the validity of our approach by comparing it to DLD non-equilibrium solutions of a dynamic traffic network. In particular, we show that if the cost function F is strongly pseudomonotone of degree $\alpha < 2$, then there are similar stability behaviors in both the hybrid system framework and the DLD framework. More specifically, if F is strongly pseudomonotone of degree $\alpha < 2$, then disturbances completely dampen out in a finite amount of time for both frameworks.

While there are numerous benefits to using hybrid system non-equilibrium solutions, there is a cost in reducing the requirement of tracking entire distributions of information. Namely, a pre-defined jump rule is required. Fortunately, such a rule could be based on previously known information as supposed to the current information requirement of DLD non-equilibrium solutions.

A particularly interesting avenue for future investigation is combining DLD and HS frameworks to model non-equilibrium solutions of dynamic networks that have partial point information and partial distribution information. Such a merger of frameworks would provide an interesting way to maximize the use of all possible information in modeling non-equilibrium solutions of dynamic networks. In addition, incorporating network delays, or stochastic disturbances would also further strengthen the applicability of the HS framework to model non-equilibrium solutions of dynamic networks.

Overall, the hybrid system non-equilibrium solutions are directly applicable to many dynamic networks, including traffic networks, oligopolistic market problems and noncooperative Nash games. Advantageously, the hybrid system framework can be applied to any dynamic network modeled by DLD non-equilibrium solutions, starting from any point on a DLD non-equilibrium solution.

Acknowledgements Scott Greenhalgh and Monica-Gabriela Cojocaru would like to thank the referees comments which led to a clearer presentation of this work. Monica-Gabriela Cojocaru graciously acknowledges the support received from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix 1: Common Definitions and Theorems for VI, EVI, and PrDE

Definition 13 Some classifications of the mapping F [17]. Given that X is a Hilbert space of arbitrary dimension, $\mathbb{K} \subset X$ is a non-empty, closed, and convex set, then a mapping $F : \mathbb{K} \to X$ is said to be

1. Pseudomonotone on $\mathbb{K}$ if

$$\langle F(x), y - x\rangle \geq 0 \Rightarrow \langle F(y), y - x\rangle \geq 0 \ \forall x, y \in \mathbb{K}$$

2. Strictly pseudomonotone on $\mathbb{K}$ if

$$\langle F(x), y - x\rangle \geq 0 \Rightarrow \langle F(y), y - x\rangle > 0 \ \forall x \neq y \in \mathbb{K}$$

3. Strongly pseudomonotone of degree α on $\mathbb{K}$ if for some $\eta > 0$,

$$\langle F(x), y - x\rangle \geq 0 \Rightarrow \langle F(y), y - x\rangle \geq \eta\|x - y\|^{\alpha} \ \forall x, y \in \mathbb{K}$$

Definition 14 Monotone attractor. Let X be a Hilbert space of arbitrary dimension, $\mathbb{K} \subset X$ be a non-empty, closed, and convex set, and $F : \mathbb{K} \to X$ a Lipschitz continuous mapping. Then

1. A point $x^* \in \mathbb{K}$ is a local monotone attractor for a PrDE if there exists a neighborhood V of x^* such that the function $\phi(\tau) := \|x(\tau) - x^*\|_X$ is non-increasing with respect to τ for any solution $x(\tau)$ of a PrDE starting in V.
2. A point $x^* \in \mathbb{K}$ is a global monotone attractor for a PrDE if condition X is satisfied for any $x(\tau) \in \mathbb{K}$.

Definition 15 Stability of equilibria. Let X be a Hilbert space of arbitrary dimension, $\mathbb{K} \subset X$ be a non-empty, closed, and convex set, and $F : \mathbb{K} \to X$ a Lipschitz continuous mapping. If $x^* \subset \mathbb{K}$ is an equilibrium of a PrDE, $B(x, r)$ is a ball of radius r centered on $x : \mathbb{R}^+ \to \mathbb{K}$ (a non-equilibrium solution to a PrDE), then

1. The point x^* is exponentially stable if there exists $\epsilon > 0$ and $\mu > 0$ such that $\forall x \in B(x^*, \epsilon)$ and $\forall \tau \geq 0$, we have that $\|x(\tau) - x^*\|_X \leq \|x(0) - x^*\|_X exp(-\mu\tau)$.
2. The point x^* is a finite-time attractor if there exists $\epsilon > 0$ such that $\forall x \in B(x^*, \epsilon)$ and $\forall \tau \geq 0$, there exists $T := T(x) < \infty$, where $x(\tau) = x^*$ for all $\tau \geq T$.
3. The point x^* is globally exponentially stable, or a global finite-time attractor if X, or respectively Y hold for any $x \in \mathbb{K}$.

Theorem 2 *Let $\mathbb{K} \subset X$ be a non-empty, closed, and convex set, $F : \mathbb{K} \to X$ a Lipschitz continuous mapping, and x^* an equilibrium of a PrDE.*

1. *If F is locally (strictly) pseudomonotone around x^*, then x^* is a local (strictly) monotone attractor.*
2. *If F is (strictly) pseudomonotone on $\mathbb{K}$, then x^* is a global (strictly) monotone attractor.*

Theorem 3 *Let $\mathbb{K} \subset X$ be a non-empty, closed, and convex set, $F : \mathbb{K} \to X$ a Lipschitz continuous mapping, and x^* an equilibrium of a PrDE.*

1. *If F is strongly pseudomonotone around x^*, then x^* is a locally exponentially stable.*
2. *If F is strongly pseudomonotone with degree $\alpha < 2$ around x^*, then x^* is a local finite-time attractor.*
3. *If F is strongly pseudomonotone on $\mathbb{K}$, then x^* is a globally exponentially stable.*
4. *If F is strongly pseudomonotone with degree $\alpha < 2$ on $\mathbb{K}$,around x^*, then x^* is a global finite-time attractor.*

Appendix 2: Strongly Pseudomonotone of Degree $\alpha < 2$

Here we show that the mapping F from the example in section "A Dynamic Traffic Network Example" is strongly pseudomonotone of degree $\frac{3}{2}$.

Proof To begin, recall that

$$F(x) = (2\sqrt{x_1 - x_1^*} + x_2 - x_2^*, x_2 - x_2^*)^T$$

with the constraint set,

$$\mathbb{K} = \{x \in L^2([0, 110], \mathbb{R}^2) | 0 \leq x_i \leq 100, x_1 + x_2 = \rho\}.$$

To show F is strongly pseudomonotone of degree $\frac{3}{2}$, we use the following identity:

$$x_1 - y_1 = -(x_2 - y_2) \text{ for all } x, y \in \mathbb{K}.$$

It follows that

$$\langle F(x)-F(y), x-y\rangle = (2\sqrt{x_1-x_1^*}+x_2-2\sqrt{y_1-x_1^*}-y_2)(x_1-y_1)+(x_2-y_2)(x_2-y_2).$$

Equivalently, replacing $x_2 - y_2$ through the identity above, we have that

$$\langle F(x)-F(y), x-y\rangle = (2\sqrt{x_1-x_1^*}-2\sqrt{y_1-x_1^*}-(x_1-y_1))(x_1-y_1)+(x_1-y_1)(x_1-y_1),$$

and

$$\langle F(x)-F(y), x-y\rangle = (2\sqrt{x_1-x_1^*}-2\sqrt{y_1-x_1^*})(x_1-y_1).$$

Because the square root function is subadditive, it follows that

$$\langle F(x)-F(y), x-y\rangle \geq (2\sqrt{x_1-y_1})(x_1-y_1) = 2(x_1-y_1)^{\frac{3}{2}}.$$

Finally, the proof is complete upon noting that

$$\eta\|x-y\|^{\alpha} = \eta\sqrt{(x_1-y_1)^2-(x_2-y_2)^2}^{\alpha} = \eta\sqrt{2}^{\alpha}(x_1-y_1)^{\alpha}.$$

Thus, we have that F is strongly pseudomonotone of degree $\alpha = \frac{3}{2}$ with $\eta = \sqrt{2}^{2-\alpha}$.

Appendix 3: Stability of a Hybrid System Non-equilibrium Solution

To demonstrate the stability properties of a hybrid system non-equilibrium solution, consider a mapping F that is strongly pseudomonotone of degree $\alpha < 2$ with constant η, and the jump rules:

$$G_j(t_j^-, x_\delta(t_j^-)) = P_{K_{j+1}}(x_\delta(t_j^-) - x_\delta^*(t_j^-) + x_\delta^*(t_j^-))$$

and

$$H_j(\theta_j) = \theta^j.$$

It follows that δ can be selected sufficiently small so that for some $t^* \in [0, T]$,

$$\|x_\delta - x^*\|_{L^2([t^*,T],\mathbb{R}^q)} < \epsilon \text{ for any } \epsilon > 0.$$

Proof The proof here follows the same approach for showing finite time attraction to an equilibrium of a projected differential equation [9, 21]. To begin, let $\Delta := \Delta_m$

be a uniform division of $[0, T]$ for some fixed m, with division points t_j, so that $|t_{j+1} - t_j| = \delta$. Taking $t > t_j$ we have that

$$\|x_\delta(t) - x^*(t_{j+1})\|_{\mathbb{R}^q}^{2-\alpha} \leq \|x_\delta(t_j) - x^*(t_{j+1})\|_{\mathbb{R}^q}^{2-\alpha} - (2-\alpha)\tfrac{\eta}{2}(t - t_j). \tag{16}$$

From the jump rule defined by (15), we have that

$$\|x_\delta(t_j) - x^*(t_{j+1})\|_{\mathbb{R}^q} \leq \|x_\delta(t_j^-) - x^*(t_j)\|_{\mathbb{R}^q}. \tag{17}$$

Since $2 - \alpha > 0$ and the power function is increasing we get

$$\begin{aligned} \|x_\delta(t) - x^*(t_{j+1})\|_{\mathbb{R}^q}^{2-\alpha} &\leq \|x_\delta(t_j^-) - x^*(t_j)\|_{\mathbb{R}^q}^{2-\alpha} - (2-\alpha)\tfrac{\eta}{2}(t - t_j), \\ &\leq \|x_\delta(t_{j-1}) - x^*(t_j)\|_{\mathbb{R}^q}^{2-\alpha} - (2-\alpha)\tfrac{\eta}{2}(t - t_{j-1}). \end{aligned} \tag{18}$$

Continuing in this fashion, we finally arrive at

$$\|x_\delta(t) - x^*(t_{j+1})\|_{\mathbb{R}^q} \leq \left(\|x_\delta(0^-) - x^*(0)\|_{\mathbb{R}^q}^{2-\alpha} - (2-\alpha)\tfrac{\eta}{2}t\right)^{\frac{1}{2-\alpha}}. \tag{19}$$

Thus t^* is taken such that:

$$t^* \geq \frac{1}{(2-\alpha)\frac{\eta}{2}}\|x_\delta(0^-) - x^*(0)\|_{\mathbb{R}^q}^{2-\alpha}. \tag{20}$$

Thus on the subinterval $[t_k, t_{k+1}]$ that contains t^*, we have necessarily that

$$\|x_\delta(t) - x^*(t_{k+1})\|_{\mathbb{R}^q} = 0 \text{ for } t \geq t^*. \tag{21}$$

Furthermore, since the jump rule maps $x^*(t_j) \to x^*(t_{j+1})$ for all j,

$$\|x_\delta(t) - x^*(t_{i+1})\|_{\mathbb{R}^q} = 0, \tag{22}$$

for all $t \geq t^*$ on each interval $[t_i, t_{i+1}] \quad \forall i > k$. Thus by Lebesgue's dominated convergence theorem, we have that δ can be selected so that

$$\|x_\delta - x^*\|_{L^2([t^*,T],\mathbb{R}^q)} \leq \epsilon, \tag{23}$$

for any $\epsilon > 0$.

References

1. J.P. Aubin, A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer (1984)
2. J.P. Aubin, A. Cellina, Differential inclusions. J. Appl. Math. Mech. **67**(2), 100 (1987)

3. A. Barbagallo, M.G. Cojocaru, Dynamic equilibrium formulation of the oligopolistic market problem. Math. Comput. Model. **49**, 966–976 (2009)
4. B. Brogliato, A. Daniilidis, C. Lemaréchal, V. Acary, On the equivalence between complementarity systems, projected systems and differential inclusions. Syst. Control Lett. **55**, 45–51 (2006)
5. M.-G. Cojocaru, *Double-Layer Dynamics Theory and Human Migration After Catastrophic Events* (Bergamo University Press, Bergamo, 2007)
6. M.G. Cojocaru, Piecewise solutions of evolutionary variational inequalities. Consequences for the doublelayer dynamics modelling of equilibrium problems. J. Inequal. Pure Appl. Math. **8**(2), 17 (2007)
7. M.-G. Cojocaru, L.B. Jonker, Existence of solutions to projected differential equations in Hilbert spaces. Proc. Am. Math. Soc. **132**, 183–193 (2004)
8. M.G. Cojocaru, P. Daniele, A. Nagurney, Projected dynamical systems and evolutionary variational inequalities via Hilbert spaces with applications. J. Optim. Theory Appl. **127**, 549–563 (2005)
9. M.G. Cojocaru, P. Daniele, A. Nagurney, Double-layered dynamics: a unified theory of projected dynamical systems and evolutionary variational inequalities. Eur. J. Oper. Res. **175**, 494–507 (2006)
10. M.G. Cojocaru, C.T. Bauch, M.D. Johnston, Dynamics of vaccination strategies via projected dynamical systems. Bull. Math. Biol. **69**, 1453–1476 (2007)
11. M. Cojocaru, P. Daniele, A. Nagurney, Projected dynamical systems, evolutionary variational inequalities, applications, and a computational procedure, in *Pareto Optimality, Game Theory* . . . , Springer (2008), pp. 387–406
12. S. Dafermos, Traffic equilibrium and variational inequalities. Transp. Sci. **14**(1), 42–54 (1980).
13. S. Dafermos, Congested transportation networks and variational inequalities, in *Flow Control of Cogested Networks* (Springer, Berlin, 1987)
14. P. Daniele, *Dynamic Networks and Evolutionary Variational Inequalities* (Edward Elgar, Cheltenham, Northampton, MA, 2006)
15. P. Daniele, A. Maugeri, W. Oettli, Time-dependent traffic equilibria. J. Optim. Theory Appl. **103**(3), 543–555 (1999)
16. P.T. Harker, A variational inequality approach for the determination of oligopolistic market equilibrium. Math. Program. **30**, 105–111 (1984)
17. S. Karamardian, S. Schaible, Seven kinds of monotone maps. J. Optim. Theory Appl. **66**, 37–46 (1990)
18. D. Kinderlehrer, G. Stampacchia, *An Introduction to Variational Inequalities* (SIAM, Philadelphia, 2000)
19. J. Lions, G. Stampacchia, Variational inequalities. Commun. Pure Appl. Math. **20**(3), 493–519 (1967)
20. A. Nagurney, *Network Economics: A Variational Inequality Approach* (Springer, Berlin, 1993)
21. A. Nagurney, D. Zhang, *Projected Dynamical Systems and Variational Inequalities with Applications* (Kluwer Academic, Dordrecht, 1996)
22. G. Stampacchia, Variational inequalities, in *Theory and Applications of Monotone Operators* Proceedings of a Nato Advance Study Inst. Vienice, Italy (1969), pp. 101–192
23. A. van der Schaft, H. Schumacher, *An Introduction to Hybrid Dynamical Systems* (Springer, Berlin, 2000)

Measuring Ballistic Dispersion for the Purpose of Ammunition Quality Assurance

W.J. Hurley, Jack Brimberg, and Andrey Pavlov

Abstract There are a variety of measures of ballistic dispersion. We examine which is best in the context of quality assurance for ammunition. It has been shown that the measure of ballistic dispersion currently used by the US and Canadian militaries for the purpose of ammunition quality assurance is not the most powerful in the case where the fall of shot follows a circular normal distribution. Here we consider the more general case of a general bivariate normal with unequal component variances.

Keywords Ballistic dispersion • Ammunition • Quality assurance testing • Bivariate normal

Introduction

Measuring ballistic dispersion is fundamental to military operations research. Whether it's for constructing artillery range tables, or comparing ballistic weapons effects, or for ammunition quality assurance testing, dispersion must be measured properly. The difficulty is that, in practise and theory, a variety of measures have been suggested and used. This begs the question which is best. In this paper, we tackle that question in the context of ammunition quality assurance.

While it is true that modern militaries, particularly the US military, have begun to include "smart" weapon systems in their force structures, ballistic systems continue to play a significant role and are expected to do so for the foreseeable future. One of the primary concerns for ballistic systems is that variation in where rounds fall be as low as possible. This principle applies to all ballistic rounds ranging from artillery systems that have ranges measured in the tens of kilometers to direct-fire personal weapons that are employed at close range. In the case of ballistic ammunition, each lot of ammunition is subjected to a battery of quality assurance tests. One of these is a test for ballistic dispersion.

W.J. Hurley • J. Brimberg (✉) • A. Pavlov
Department of Mathematics and Computer Science, Royal Military College of Canada,
Kingston, ON, Canada
e-mail: William.Hurley@rmc.ca

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_14

Empirical tests over many years have found that the fall of shot for most ballistic rounds is well described by a bivariate normal distribution. Let us suppose that the parameters characterizing variation for this family of distributions are σ_x and σ_y. For artillery rounds, the range error, σ_x, is larger than the deflection error, σ_y. But for other rounds, such as 25mm rounds, $\sigma_x = \sigma_y$. For these rounds, the distribution is circular normal, a special case of the bivariate normal.

In practise, tests for dispersion require that a random sample of rounds from a lot be fired and the locations recorded. Using this sample data, a sample mean point of impact and component sample standard deviations, s_x and s_y, are calculated. We will give the details of this calculation in a subsequent section. Both the Canadian Forces (CF) and US Army Ordnance Corp (AOC) use these component standard deviations when testing for dispersion. The critical region for these tests takes the form

$$\max(s_x, s_y) \geq \kappa_0. \tag{1}$$

That is, if one of the component standard deviations is too high, the lot is rejected. But in the case where the fall of shot follows a circular normal distribution, [1] has shown that a uniformly most powerful test results in a critical region of the form

$$\frac{s_x^2 + s_y^2}{2} \geq \tau_0. \tag{2}$$

That is, the lot is rejected if the *average* of the component sample variances is high enough. Equivalently, the lot is rejected if the *sum* of the component sample variances is sufficiently high:

$$s_x^2 + s_y^2 \geq \tau_0'. \tag{3}$$

So, an obvious and important question arises: In Canada and the USA, is the current practise of using the statistic $\max(s_x, s_y)$ the best in the case of the general bivariate normal distribution when $\sigma_x \neq \sigma_y$, or is there a better one? This is the problem we study in this paper.

Our main result is that a uniformly most powerful test for the general case $\sigma_x \neq \sigma_y$ does not exist. However, in considering tests where a fixed value of dispersion is specified in the alternative hypothesis, we are able to partition the (s_x, s_y) space into regions where the test statistics given in (1) or (2) are best. More particularly if σ_x and σ_y are close enough, then (2) is best; otherwise (1) is best. In the case where a significant covariance component occurs, we propose a third statistic that works best. These results clearly demonstrate that there is not a single correct measure of dispersion when it comes to ammunition quality assurance.

The literature on acceptance sampling is large; for example, see the text by [3] or just about any operations management or statistics textbook. However, these sources do not usually concern themselves with specialized tests as occurs in the case of ballistic dispersion. Our intention is to help fill this particular gap.

The paper is in three sections. In the next section, we discuss approaches to the measurement of ballistic dispersion. Following that and for completeness, we review the argument for a uniformly most powerful test of dispersion in the case of a circular normal distribution as presented in [1]. In a final section we develop testing procedures for the general bivariate normal distribution.

Measuring Ballistic Dispersion

By way of example, consider a lot of 25mm rounds fired from the main primary armament of the Canadian Forces' LAV III armoured vehicle. The Department of National Defence buys these rounds in lots of 5,000 and 10,000 units. To test for dispersion, a random sample of rounds is first drawn and then fired from a fixed-mount Mann barrel at a target 300 meters away. The purpose for using a fixed-mount barrel is to remove, as much as possible, the various sources of error that a gun would be subjected to in an operational setting. The locations of these rounds on the target are recorded electronically. Figure 1 shows a sample "cloud." Note that two axes have been drawn by first inserting a y-axis through the point furthest to the left of the cloud and then an orthogonal x-axis through the point furthest down.

Let the set of point locations be

$$L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}.$$

The *Sample Mean Point of Impact* (*sMPI*) is calculated by taking the averages of the x and y components of the round locations:

$$\bar{x} = \frac{1}{n}\sum x_i \, , \bar{y} = \frac{1}{n}\sum y_i. \tag{4}$$

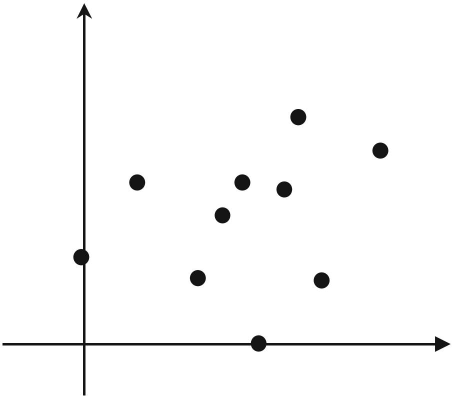


Fig. 1 Ballistic footprint with x and y axes

There are other possibilities to estimate the "center" of the cloud. For example, we could use the *median point* (x^*, y^*) that solves

$$\min_{x,y} \sum_{i=1}^{n} \sqrt{(x_i - x)^2 + (y_i - y)^2}. \tag{5}$$

In this paper, we have chosen to use the sMPI for the following reason. We are going to model the fall of shot as a random drawing from a probability distribution. For instance, 25mm rounds follow a circular normal distribution

$$f_C(x, y) = \frac{1}{2\pi\theta} \exp\left[-\frac{(x - \mu_x)^2 + (y - \mu_y)^2}{2\theta} \right], \tag{6}$$

where μ_x and μ_y are location parameters and θ is a dispersion parameter. Given a sample fall of shot, a rational approach to estimate the location parameters would be with their maximum likelihood estimators, which in this case, are the sample means, $\overline{x}$ and $\overline{y}$. Hence, the sample MPI will be used here to estimate μ_x and μ_y, the coordinates of some hypothetical point of aim. We will argue what the estimator of θ should be below.

The CF and AOC measures of dispersion are

$$s_x = \sqrt{\frac{1}{n} \sum (x_i - \overline{x})^2} \text{ and } s_y = \sqrt{\frac{1}{n} \sum (y_i - \overline{y})^2}. \tag{7}$$

These are just the sample standard deviations in each component direction. Two useful properties are noted:

Property 1 s_x and s_y are invariant to a translation of the axes.

To see this, note that, for the transformation $x_i' = b + x_i, i = 1, \ldots, n$, we have that $\overline{x}' = b + \overline{x}$ and

$$s_{x'} = \sqrt{\frac{1}{n} \sum (b + x_i - (b + \overline{x}))^2} = s_x \tag{8}$$

Similarly, s_y is invariant to a translation of the axes.

Property 2 Consider the dataset L and suppose we use the translation

$$\begin{matrix} x_i' = x_i - \overline{x} \\ y_i' = y_i - \overline{y} \end{matrix} \quad \text{for } i = 1, 2, \ldots, n. \tag{9}$$

Then the sMPI is $(0, 0)$ and the sample dispersions are $s_{x'} = s_x$ and $s_{y'} = s_y$.

To see this property, note that

$$\overline{x}' = \frac{1}{n}\sum(x_i - \overline{x}) = \frac{1}{n}\sum x_i - \frac{1}{n}\sum \overline{x} = \overline{x} - \overline{x} = 0 \tag{10}$$

and, by the same argument, $\overline{y}' = 0$. That the component sample standard deviations are unchanged follows directly from Property 1.

Quality Assurance in the Case of a Circular Normal Distribution

Suppose the fall of shot is consistent with a circular normal distribution as shown in (6) where the point of aim is (μ_x, μ_y) and the variance is θ. Clearly dispersion is governed by the parameter θ. Suppose that we have a random sample of n rounds from a lot resulting in the locations given in L, and this dataset yields the sample standard deviations s_x and s_y.

We are interested in testing the null hypothesis

$$H_0\colon \theta = \theta^* \tag{11}$$

against the alternative

$$H_1\colon \theta > \theta^*, \tag{12}$$

where θ^* is some maximum allowable dispersion identified by defence scientists.

Assuming that the sMPI, $(\overline{x}, \overline{y})$, may replace the unknown point of aim (μ_x, μ_y) without significant error, as in the case of a large sample, and applying Property 2, we set $\overline{x} = \overline{y} = 0$ by an appropriate translation of the axes so that the underlying distribution becomes:

$$f_{C_0}(x,y) = \frac{1}{2\pi\theta}\exp\left[-\frac{x^2+y^2}{2\theta}\right]. \tag{13}$$

The joint distribution of the sample, then, is

$$J(\theta;(x_1,y_1),\ldots,(x_n,y_n)) = \left(\frac{1}{2\pi\theta}\right)^n \exp\left[-\frac{1}{2\theta}\left(\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2\right)\right]. \tag{14}$$

Now consider testing

$$H_0\colon \theta = \theta^* \tag{15}$$

against the simple alternative

$$H_1 \colon \theta = \theta^{**} \tag{16}$$

where $\theta^{**} > \theta^{*}$. By the Neyman-Pearson lemma, a best critical region for this test is obtained by solving

$$\frac{J(\theta^{*};(x_1, y_1), \ldots, (x_n, y_n))}{J(\theta^{**};(x_1, y_1), \ldots, (x_n, y_n))} \leq a \tag{17}$$

for some $a > 0$. The ratio on the left-hand side simplifies to

$$\frac{J(\theta^{*})}{J(\theta^{**})} = \left(\frac{\theta^{**}}{\theta^{*}}\right)^{n} \exp\left[-\frac{1}{2}\left(\frac{\theta^{**} - \theta^{*}}{\theta^{**}\theta^{*}}\right)\left(\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2\right)\right]. \tag{18}$$

Taking the ln of both sides of $J(\theta^{*})/J(\theta^{**}) \leq a$ gives the critical region

$$\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 \geq b, \tag{19}$$

where

$$b = \frac{2\theta^{**}\theta^{*}}{\theta^{**} - \theta^{*}}\left[n \ln\left(\frac{\theta^{**}}{\theta^{*}}\right) - \ln(a)\right]. \tag{20}$$

But given that the sMPI is the origin of the coordinate system, we can rewrite (19) as

$$s_x^2 + s_y^2 \geq b/n. \tag{21}$$

Under the assumption of the circular normal distribution in (13), it is well known that the marginal densities in the coordinate directions are normal (e.g., [2, pp. 158–159]). That is, the density in the x direction is

$$f(x) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right) \tag{22}$$

and the density in the y direction is

$$g(y) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{y^2}{2\theta}\right). \tag{23}$$

Using this result, we conclude that both $\sum_{i=1}^{n} X_i^2/\theta$ and $\sum_{i=1}^{n} Y_i^2/\theta$ are chi-square random variables with n degrees of freedom. Hence,

$$\frac{n}{\theta}\left(S_x^2+S_y^2\right)=\frac{1}{\theta}\left(\sum_{i=1}^{n}X_i^2+\sum_{i=1}^{n}Y_i^2\right) \tag{24}$$

has a chi-square distribution with $n+n=2n$ degrees of freedom. For a given level of significance, α, we can calculate the value of b/θ using

$$\alpha=\Pr\left(\frac{1}{\theta}\left(\sum_{i=1}^{n}X_i^2+\sum_{i=1}^{n}Y_i^2\right)\geq\frac{b}{\theta};H_0\right) \tag{25}$$

and then get b which we denote b_α. Hence the critical region of size α is

$$s_x^2+s_y^2\geq b_\alpha/n. \tag{26}$$

Continuing the argument, there is nothing special about the value θ^{**}. The argument above is true for any value of $\theta^{**}>\theta^{*}$. Consequently, we conclude that (26) is a uniformly most powerful critical region of size α for testing H_0: $\theta=\theta^{*}$ against the alternative H_1: $\theta>\theta^{*}$.

Therefore, the appropriate statistic for a uniformly most powerful test of dispersion in the case of the circular normal distribution is not $\max(s_x,s_y)$ but rather the sum of the component sample variances, or equivalently, the average of the component variances.

In the next section, we examine the more general case of a bivariate normal distribution with unequal variances.

Testing Dispersion for a Bivariate Normal Distribution

Let $Z^T=(X,Y)$ be a random vector having a bivariate normal distribution with mean at the origin (i.e., $(\mu_x,\mu_y)=(0,0)$) and a non-degenerate but otherwise arbitrary covariance matrix

$$\Sigma=\begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix} \tag{27}$$

where σ_x and σ_y are the standard deviations along the x and y axes, and ρ is the correlation coefficient. We consider a sample drawn from this distribution. Generally speaking, the data is mostly spread along the large axis of the ellipse. More precisely, we define a dispersion measure D to be the largest variance observed in all possible directions:

$$D=\sup_{\|h\|=1}Var(h^TZ). \tag{28}$$

Here, the supremum is taken over all two-dimensional vectors h of Euclidean length 1.

To calculate the right-hand side, recall that the covariance matrix can be factored into the product of a symmetric square-root matrix with itself:

$$\Sigma = \Sigma^{1/2}\Sigma^{1/2}. \tag{29}$$

This enables us to derive that

$$\begin{aligned} Var(h^T Z) &= E(h^T Z)(h^T Z)^T \\ &= h^T E(ZZ^T)h \\ &= h^T \Sigma h \\ &= (h^T \Sigma^{1/2})(h^T \Sigma^{1/2})^T \\ &= \left\| h^T \Sigma^{1/2} \right\|^2 \end{aligned} \tag{30}$$

which is the squared length of the vector $h^T \Sigma^{1/2}$. Hence D can be expressed as

$$D = \sup_{\|h\|=1} \left\| h^T \Sigma^{1/2} \right\|^2. \tag{31}$$

It is well known that the Euclidean matrix norm coincides with the spectral norm which, in turn, is equal to the largest eigenvalue of $\Sigma^{1/2}\Sigma^{1/2}$ or just the covariance matrix, Σ. To find the largest eigenvalue, we solve the characteristic equation $|\Sigma - \lambda I| = 0$, or

$$\begin{vmatrix} \sigma_x^2 - \lambda & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 - \lambda \end{vmatrix} = 0. \tag{32}$$

This equation has two real solutions, the largest yielding D:

$$D = \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 + \sqrt{\left(\sigma_x^2 - \sigma_y^2\right)^2 + 4\rho^2\sigma_x^2\sigma_y^2}\right). \tag{33}$$

Result 1 $D \geq \max(\sigma_x^2, \sigma_y^2)$ with equality occurring only if the correlation coefficient $\rho = 0$.

Proof Clearly

$$\begin{aligned} D &\geq \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 + \sqrt{\left(\sigma_x^2 - \sigma_y^2\right)^2}\right) \\ &= \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 + \left|\sigma_x^2 - \sigma_y^2\right|\right) \\ &= \frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 + \max(\sigma_x^2, \sigma_y^2) - \min(\sigma_x^2, \sigma_y^2)\right) \\ &= \max(\sigma_x^2, \sigma_y^2) \end{aligned}$$

with equality in the first line only if $\rho = 0$.

So the dispersion measure, D, of a tilted ellipse is at least as large as each of the component variances.

Our primary interest lies in testing whether D exceeds a certain critical value, say D^*, given a sample of n independent draws. Hence the null hypothesis is given by

$$H_0\colon D \leq D^*. \tag{34}$$

We now argue that this is equivalent to testing

$$H_0\colon D \leq 1 \tag{35}$$

with a suitable rescaling of the data. Consider the transformation

$$\begin{aligned} u_i &= x_i/\sqrt{a}, i = 1, 2, \ldots, n, \\ v_i &= y_i/\sqrt{a}, i = 1, 2, \ldots, n, \end{aligned} \tag{36}$$

where $a > 0$. Note that

$$\sigma_u^2 = \sigma_x^2/a \tag{37}$$

and

$$\sigma_v^2 = \sigma_y^2/a. \tag{38}$$

We now consider the calculation of D in uv-space:

$$\begin{aligned} D_{uv} &= \frac{1}{2}\left(\sigma_u^2 + \sigma_v^2 + \sqrt{\left(\sigma_u^2 - \sigma_v^2\right)^2 + 4\rho^2\sigma_u^2\sigma_v^2}\right) \\ &= \frac{1}{2}\left(\frac{1}{a}\sigma_x^2 + \frac{1}{a}\sigma_y^2 + \sqrt{\left(\frac{1}{a}\sigma_x^2 - \frac{1}{a}\sigma_y^2\right)^2 + 4\rho^2\frac{1}{a}\sigma_x^2\frac{1}{a}\sigma_y^2}\right) \\ &= \frac{1}{a}\left[\frac{1}{2}\left(\sigma_x^2 + \sigma_y^2 + \sqrt{\left(\sigma_x^2 - \sigma_y^2\right)^2 + 4\rho^2\sigma_x^2\sigma_y^2}\right)\right] \\ &= D/a. \end{aligned} \tag{39}$$

Considering the hypothesis

$$D \leq D^*, \tag{40}$$

we note that it is equivalent to

$$\frac{1}{D^*}D \leq 1 \tag{41}$$

or, using (39) with $a = D^*$,

$$D_{uv} \leq 1. \tag{42}$$

Hence a rescaling of the data will always give the null, H_0: $D \leq 1$.

So without loss of generality, we assume that the critical value is 1, and consider the null

$$H_0 : D \leq 1 \tag{43}$$

versus the alternative H_1: $D > 1$.

In terms of the model parameters, H_0 defines a closed set in the 3-dimensional space $\left(\sigma_x^2, \sigma_y^2, \rho^2\right)$. We also know that this set is a subset of a unit cube for $\rho \geq 0$ or ≤ 0 through the inequality in Result 1 and the bound on correlation. A few of its ρ-sections are shown in Fig. 2. These range from the full square for $\rho = 0$ to the unit quarter circle for $\rho = \pm 1$.

We are going to examine three tests of H_0 based on the sample quantities s_x, s_y, and

$$s_{xy} = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}):$$

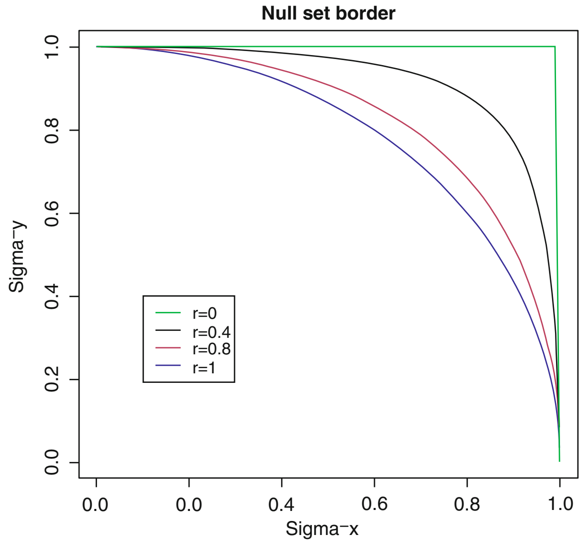


Fig. 2 ρ-sections of the parameter space defined by the null hypothesis

$T1$. Reject H_0 if $t_1 = (s_x^2 + s_y^2)/2 > B_1(\alpha)$
$T2$. Reject H_0 if $t_2 = \max(s_x, s_y) > B_2(\alpha)$
$T3$. Reject H_0 if $t_3 = \left(s_x^2 + s_y^2 + \sqrt{\left(s_x^2 - s_y^2\right)^2 + 4s_{xy}^2}\right)/2 > B_3(\alpha)$

where the $B_i(\alpha)$ values are chosen to maintain the probability of Type I error:

$$\sup_{H_0} P(T_i > B_i(\alpha)) = \alpha. \tag{44}$$

Our calculations are simplified due to a well-known fact that the supremum generally occurs on the boundary of H_0. This is explained intuitively in our case by noting that the set $D = 1$ is a collection of the "worst" points, the points that are most difficult to distinguish from those with $D > 1$. We give the result formally without proof.

Result 2 The supremum of $P(T_i > B_i(\alpha))$ over H_0 satisfies $D = 1$.

That the supremum occurs on the boundary is important because we can now execute a constrained two-dimensional search to find it. The following algorithm can be used:

Step 1. Simulate two independent sequences, X_{kj}^* and Y_{kj}^*, of standard normal random variates where $k = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, N$. N is the number of Monte Carlo iterations and should be large.

Step 2. Let $\left(\sigma_x^2, \rho^2\right)$ be chosen, and σ_y^2 determined so that $D = 1$. Then calculate the standard transformation to get bivariate normal drawings:

$$\begin{aligned} X_{kj} &= \sigma_x X_{kj}^* \\ Y_{kj} &= \sigma_y \left(\rho X_{kj}^* + \sqrt{1-\rho^2} Y_{kj}^*\right) \\ Z_{kj} &= (X_{kj}, Y_{kj}) \end{aligned} \tag{45}$$

Step 3. Compute statistics $T_{i,j}$ from the sample $Z_{1j,}Z_{2j}, \ldots, Z_{nj}$ for $i = 1, 2, 3$ and $j = 1, 2, \ldots, N$.

Step 4. Determine $B_i(\alpha)$ as the $(1-\alpha)$-th quantile. That is, take the $[N(1-\alpha)]$-th order statistic among $T_{i,j}, j = 1, 2, \ldots, N$.

Step 5. Repeat Steps 2–4 to maximize $B_i(\alpha)$ with respect to $\left(\sigma_x^2, \rho^2\right)$. (We assume the algorithm terminates when a standard stopping criterion has been satisfied.)

Thus, critical values can be determined for any sample size n by applying steps 1–4 of the algorithm. For $n = 20$, the following critical values have been obtained:

$$\begin{aligned} B_1(0.05) &= 1.33 \\ B_2(0.05) &= 1.64 \\ B_3(0.05) &= 1.78 \end{aligned} \tag{46}$$

The three tests are compared by studying their powers for various alternatives. Any alternative can be specified by a subset in the space $(\sigma_x^2, \sigma_y^2, \rho^2)$. From a practical point of view, a reasonable alternative would be formulated in terms of D. For example, it could be, say, "D exceeds the critical value by 50%." With our convention, this reduces to $D = 1.5$. This set is a two-dimensional manifold and is best described by the parameters ρ^2 and $\gamma = \sigma_x/\sigma_y$. Indeed, due to the scaling property described above, all tests should be equally sensitive to a simultaneous increase or decrease of σ_x and σ_y by the same factor. It is therefore their ratio, γ, that matters when comparing power. Furthermore, X and Y can be relabelled without changing any of the three statistics. Thus it suffices to consider $\gamma \geq 1$.

To determine power, we use an algorithm similar to what we used above for determining critical values:

Step 1. Simulate two independent sequences, X_{kj}^* and Y_{kj}^*, of standard normal random variates where $k = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, N$. N is the number of Monte Carlo iterations and should be large.

Step 2. Let $\left(\gamma, \rho^2\right)$ be chosen and σ_y^2 determined so that $D = 1.5$ and $\sigma_x^2 = \gamma^2\sigma_y^2$. Then put

$$\begin{aligned} X_{kj} &= \sigma_x X_{kj}^* \\ Y_{kj} &= \sigma_y \left(\rho X_{kj}^* + \sqrt{1-\rho^2} Y_{kj}^*\right) \\ Z_{kj} &= (X_{kj}, Y_{kj}) \end{aligned} \tag{47}$$

Step 3. Compute statistics $T_{i,j}$ from the sample $Z_{1j,} Z_{2j}, \ldots, Z_{nj}$ for $i = 1, 2, 3$ and $j = 1, 2, \ldots, N$.

Step 4. Determine power $\beta_i(\alpha)$ as the proportion of statistics $T_{i,j}, j = 1, 2, \ldots, N$, which exceed the critical level $B_i(\alpha)$.

For given values of γ and ρ, there is an ordering of the three tests by power. In Fig. 3 we have plotted the three regions where each of $T1$, $T2$, and $T3$ is best. Consider first the case where $\rho = 0$ (i.e., along the horizontal axis). For $\gamma = 1, T1$ is best which is consistent with what [1] derived. As γ increases, note that there is a crossover point where $T2$ is best. We judge this crossover point to be approximately $\gamma = 1.19$. Keep in mind these results are for the case $n = 20$ and the alternative hypothesis $D = 1.5$. Obviously the crossover point will depend on the specific values used for these parameters. In Table 1 we show the power of the three tests for various values of γ. The conclusion is that for ammunition types where ρ happens to be 0 and $\gamma > 1$, one must choose the correct test statistic carefully. The Monte Carlo algorithms specified above can be used to calculate critical values and the power for given values of n and specific alternative hypotheses.

In the case where ρ is non-zero (the cloud exhibits correlation), care must be taken. With ballistic rounds, correlation in the cloud usually is an indication of a more fundamental problem with the ammunition. For instance, rounds leaving the barrel with significant yaw can lead to correlation in the cloud. So a lot exhibiting

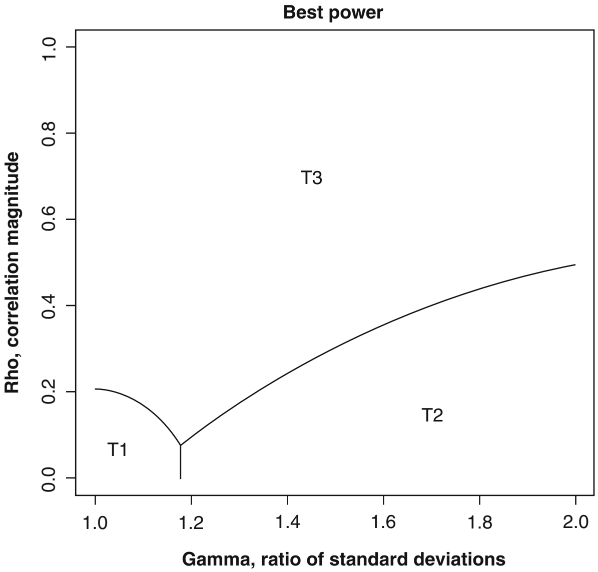


Fig. 3 Regions in (γ, ρ) space where each of T1, T2, and T3 is optimal

Table 1 Power for each of the tests for various values of γ when $\rho = 0$, $n = 20$ and the alternative hypothesis is $D = 1.5$

γ	$T1$	$T2$	$T3$
1.0	0.582	0.500	0.504
1.1	0.426	0.375	0.377
1.5	0.119	0.293	0.237
2.0	0.049	0.293	0.219

such correlation is unlikely to pass the complete bank of tests that an ammunition lot is put through. Nonetheless, to assess dispersion when ρ is non-zero, the test T3 is robust to correlation and should be used if the tilt in the cloud is sufficiently high.

Conclusions

Various statistical tests are examined for determining whether the ballistic dispersion of an ammunition lot is sufficiently low in the case where the fall of shot follows a general bivariate normal distribution. Monte Carlo procedures are derived to calculate the critical values and the powers for three tests. For the case of zero correlation, and the bivariate normal distribution is circular or almost circular, the statistic $(S_x^2 + S_y^2)/2$ (or equivalently $S_x^2 + S_y^2$) should be used; otherwise $\max(S_x, S_y)$ is best. The precise indifference point will depend on the number of rounds sampled and the specific alternative hypothesis. A third statistic is recommended when the correlation is sufficiently large.

What is perhaps most clear is that there is not a single measure of ballistic dispersion that fits all purposes. Indeed, the choice of ballistic measure for ammunition quality assurance depends primarily on the relative sizes of the component variances and the covariance. Future efforts will focus on finding a clear set of guidelines through an intensive simulation study.

References

1. W.J. Hurley, Acceptance sampling procedures for ballistic dispersion. J. Def. Model. Simul. **5**(1), 21–32 (2008)
2. C.R. Rao, *Linear Statistical Inference and Its Applications*, 2nd edn. (Wiley, New York, 1973)
3. K.S. Stephens, *The Handbook of Applied Acceptance Sampling: Plans, Procedures and Principles* (ASQ Quality Press, Milwaukee, WI, 2001)

Region-Based Watermarking for Images

Konstantinos A. Raftopoulos, Nikolaos Papadakis, Klimis S. Ntalianis, Paraskevi Tzouveli, Georgios Goudelis, and Stefanos D. Kollias

Abstract Plain rotation, scaling, and/or translation (RST) of an image can lead to loss of watermark synchronization and thus authentication failure with standard techniques. The block-based approaches in particular, albeit strong against frequency and cropping attacks, are sensitive to geometric distortions due to the need for repositioning the blocks' rectangular grid of reference. In this paper, we propose a block-based approach for watermarking *image objects* in a way that is invariant to RST distortions. With the term *image object* we refer to semantically contiguous parts of images that have a specific contour boundary. The proposed approach is based on shape information since the watermark is embedded in image blocks, the location and orientation of which are defined by *Eulerian* tours that are appropriately arranged in layers, around the object's *robust skeleton*. The object's robust skeleton is derived by its boundary after applying an extraction technique and not only is invariant to RST transformations but also to cropping, clipping, and other common deformation attacks, difficult to defend with current methods. Experiments using standard benchmark datasets demonstrate the advantages of the proposed scheme in comparison to alternative state-of-the-art methods.

K.A. Raftopoulos (✉)
National Technical University of Athens, Zografou, Greece

The American College of Greece, Agia Paraskevi, Greece
e-mail: craftopoulos@acg.edu; raftop@image.ntua.gr

N. Papadakis
Hellenic Military Academy, Vari Attikis, Greece
e-mail: npapadakis@sse.gr

K.S. Ntalianis
Technical Educational Institute of Athens, Egaleo, Greece
e-mail: kntal@teiath.gr

P. Tzouveli • G. Goudelis • S.D. Kollias
National Technical University of Athens, Zografou, Greece
e-mail: tpar@image.ntua.gr; gg@image.ntua.gr; stefanos@cs.ntua.gr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_15

Keywords Image watermarking • Intellectual property protection • Invariant watermarking • Region-based watermarking

AMS Subject Classification Numbers: 65D18, 68U10, 94A08, 68P25

Introduction to Image Watermarking and the Problem of Geometric Distortions

Even though several watermarking techniques have been proposed in the literature [1–4], most of them are not designed for image objects and they are not resistant to the geometric attacks of rotation, scaling, and translation (RST). Watermark synchronization may thus be lost and in such a case, copyright cannot be automatically claimed. The early papers (e.g., [3]) assumed that the undergone distortion was known at the detection phase, thus only checking whether the embedded message could still be recovered after a distort-restore process was enough. In other schemes, called non-blind or non-oblivious, watermarking relies on information about the original content. In order to recover synchronization, one uses the original undistorted content to establish correspondence between both signals. Among the proposed registration techniques are both area-based [5] and feature-based methods [6].

Another group of approaches is based on exhaustive search. These techniques consider all possible deformations, performing the respective inverse transformations for detection. In these cases, the embedded watermark is extracted by choosing the best detection confidence value, which is above an appropriate threshold. Besides computational constraints, one must carefully study false positive probability as it increases with the size of the search [7]. Strong hypotheses can be made on the relative continuity or smoothness of the deformation, further reducing the search space [8]. Another means to reduce the search space is to use periodically structured watermarks [9] so that search for synchronization is limited over one repetition period. Additionally, training (pilot) signals, which are known and easy-to-detect features, are used in digital communications to identify channel characteristics. In the watermarking domain, pilot signals should survive distortions. Distortion estimation requires an exhaustive search to register detected features with reference pilots and several such approaches exist [10, 11]. Others design a reference signal (template) with strong spectral characteristics, and embed it in a Fourier related or other domains [12, 13]. Associated reference marks can be detected in the Fourier magnitude spectrum [14] or by using the signal autocorrelation function.

Another set of methods considers transformed domains such as log-polar and Fourier-Mellin. These approaches rely on the properties of the Fourier transform to be insensitive to translation at its magnitude spectrum, scaling produces inverse scaling and image rotation yields identical rotation on the spectrum. Performing log-polar mapping of the Fourier domain, rotation and scaling of the image reduce to plain translations. By combining both transforms one can design the Fourier-Mellin domain, which is insensitive to rigid RST deformations. Several authors focus on these transforms [15, 16]. Performing log–log mapping of the Fourier domain provides a representation insensitive to cropping, scaling, and modification of aspect ratio [17]. However, discrete implementation of such transforms is not straightforward, the most recent approaches introduce significant complexity, while confronting accuracy problems.

Another deficiency of the majority of existing techniques is that they are frame-based and thus semantic regions such as humans, buildings, and cars are not considered explicitly. These regions may need better protection or depending on the specific application, can be the only regions that need protection (e.g., versus the background). Even though a limited number of region watermarking schemes have also been proposed [18–20], the literature still lacks efficient algorithms for content authentication especially in the case of copy-paste attack. In this kind of attack, a copyrighted image is cropped and a part or parts of the image (e.g., semantically meaningful objects) appear inside a different content (another image) after some usually minor modifications. The new content is then distributed as a new creation.

Few methods can confront such a treatment. The proposed system protect against this by embedding the watermark around the object's *robust skeleton*. Robust skeleton is defined as a certain MAT transform that is insensitive to high frequency perturbations at the object's boundary. Compared to alternative RST block-based approaches, the proposed method exhibits performance advantages. The position and orientation of the watermarked blocks are readily inferred from appropriate *Eulerian tours* around the robust skeleton. In comparison, the Fourier-Mellin transform relies to tedious transformations that require re-sampling in the frequency domain or/and expensive searches for possible rotations of rectangular block grids.

In the proposed method, the skeleton of each host image object is initially extracted. Starting from a specific point on the skeleton (see section "The Proposed Approach: Invariant Watermarking of Image Regions" for details on choosing this point), the pseudo-random watermark sequence is embedded in the DCT domain of non-overlapping blocks along the skeleton's Eulerian tour. The Eulerian tour is extended outwards, in consecutive layers towards the object's boundary, until the watermark sequence is spread to the whole image object. The length of the watermark is calculated based on the length of the extended Eulerian tour, so that it is statistically undetectable. During watermark detection, initially the skeleton of the candidate object is extracted and the potentially watermarked blocks that are located along the extended Eulerian tour are matched against the respective blocks of the initial skeleton. The respective extended Eulerian tour is traversed until either the object is authenticated or the object boundary is reached. For each Eulerian tour, the extracted sequence is correlated to the embedded pseudo-random sequence. Experimental results with objects from the Caltech and Kimia datasets show that the proposed approach re-gains synchronization in cases of RST attacks, mixed attacks, and the copy-paste attack. The method has also been presented in [21] but new experiments on the robustness of the skeleton using the benchmark datasets, comparisons with state of the art, and new attack/retrieval scenarios have been added here. The rest of this paper is organized as follows: in Sect. *The Proposed Approach: Invariant Watermarking of Image Regions* the proposed approach is presented while in Sect. *Experimental Results: Proposed Method vs State of the Art* experiments are performed to indicate the superiority of the proposed system. A conclusion in Sect. *Synopsis: A Geometric Invariant Watermarking Scheme for Image Regions* closes this contribution.

The Proposed Approach: Invariant Watermarking of Image Regions

Watermarking approaches using spread spectrum modulation of pseudo-random signals, even though successfully improve watermark resistance to various attacks, cannot withstand rotation because the rectangular grid arrangement is changed and thus synchronization is lost. A reference point for placing the rectangular block grid for watermark recovery is not easy to obtain after image/object rotation, if no a priori registration information is available. The situation is illustrated in [22] where elaborate methods of high complexity are proposed to alleviate this problem. The main contribution in this paper is that it incorporates shape information during the block-based watermark embedding process in such a way that the watermarked blocks' location is readily identifiable at the retrieval phase, even after RST transformations. The Medial Axis Transform (MAT) possesses significant advantages with respect to object boundary distortion and general object deformation, since a significant portion of the MAT skeleton is invariant to most of the typical watermarking deformation attacks on objects. Even though the MAT is sensitive to boundary details, it is possible to extract a *main medial axis* which represents the most salient features of the underlying shape [23]. Such regularization methods establish a robust skeletal representation of a shape, the most important advantage of which is that it can capture the main characteristics of shapes with a significantly distorted, jagged boundary. Local boundary deformations affect only the local skeleton structure whereas global deformation, boundary perturbation, cropping, or articulation attacks have no significant or at most local effect on the object's skeleton.

The key idea in this paper is that, instead of using a rectangular block grid with no obvious connection to the objects shape, we let the Eulerian tour [24] (see Fig. 1) around the object's skeleton to define the position, rotation, and sequencing of the watermarked blocks. The proposed method receives an image object as input and computes its skeleton. From a starting point on the skeleton and by traversing the extended Eulerian tour clockwise, we embed the watermark in the blocks defined along this tour, using a standard block-based DCT approach [1]. The blocks are chosen to be adjacent but not overlapping. The watermark is created using a pseudo-random number generator (PRNG) [25]. The PRNG starts from an arbitrary state using a seed state, which in our method is the authorization key *k*. Using a specific key as input to the PRNG, the same sequence N is always produced in the output of the PRNG.

The maximum length of the sequence is determined by the size of the key k and it is measured in bits. This sequence N is the watermark, which is embedded to the computed DC coefficient of each block along the corresponding Eulerian tours. After a complete Eulerian tour around the skeleton (Fig. 2), returning to the

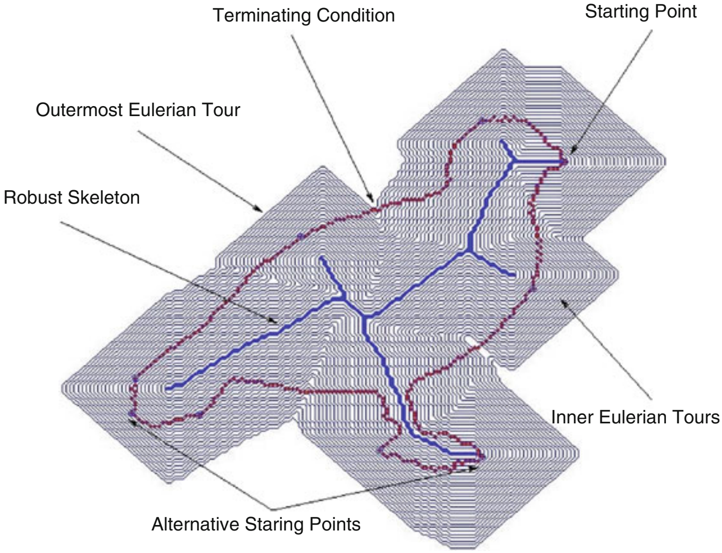


Fig. 1 The proposed block-based watermarking method along the extended Eulerian tour

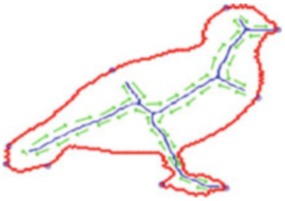


Fig. 2 Eulerian tour traverse

starting point, we continue with another complete Eulerian tour that extends one layer outwards, towards the object's boundary, (see Fig. 1) and we keep extending the Eulerian tour outwards until we reach the boundary. The process ends when the whole interior of the object has been covered with blocks this way. It is possible to meet the boundary at various points before the whole shape has been covered with watermarked blocks (e.g., at a narrow bottleneck type of boundary formation). In such a case we don't break the path, we just continue until we are in the interior again but we watermark only the blocks in the interior of the shape as in the bird's feet in Fig. 1.

In Fig. 1 we can see how the Eulerian tours are formed beginning from around the skeleton and extending to the outermost Eulerian tour, marked as such in the figure, where the terminating condition of covering the whole shape has been met. The watermark is embedded redundantly several times along the Eulerian tours, the length of the key being a tradeoff between the strength against key recovery attacks and the watermark resistance to cropping attacks, thus the length of the key is chosen as a parameter depending on the total length of the Eulerian tours. Finally, by applying the IDCT to each interior block, the watermarked object is produced. The extraction module receives a candidate object at its input, with the purpose to define this object's intellectual rights source. The algorithm, similarly to the embedding phase, initially computes the object's skeleton and corresponding Eulerian tours and then defines the non-overlapping blocks, for which the DCT is computed.

Since the same procedure is followed for the original object, the DC values of both the original and test objects are available to the authentication mechanism. Object's originality is thus decided based on the correlation between the original and retrieved watermarks. The proposed approach has many advantages, mainly due to the robust skeleton and the uniqueness of the extended Eulerian tours. With regard to initialization, the choice of the starting point on the Eulerian tour is critical for the method's performance, since a random choice would demand an exhaustive search of all the possible matching combinations at the retrieval phase.

Fortunately, the robust skeleton provides uniquely identifiable points that can be used to restrict the exhaustive search; these are the end points of the skeleton, which are marked with big dots on the boundary (see Fig. 1). By choosing one of these points as a starting point during watermark embedding, we restrict the possible matches of starting points at the retrieval phase to the number of the robust skeleton end points. In the next section the performance of the proposed scheme against traditional attacks is exhibited.

Experimental Results: Proposed Method vs State of the Art

The first step of the proposed method is the extraction of the object's skeleton. Research in this area can be found at [26, 27] leading to popular methods like *Fast Marching* [28]. The skeletonization algorithm presented here has to be robust against reasonable deformations on the boundary and of course invariant to RST transformations. To assess the proposed skeletonization approach against noise and deformation effects we conduct experiments using the KIMIA benchmark dataset. Results are presented in Figs. 3 and 4 and discussed in subsequent sections.

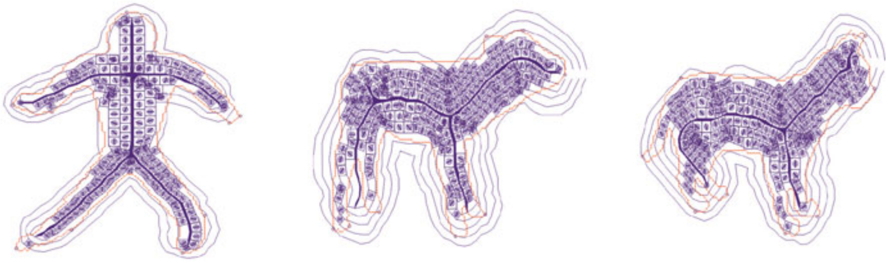


Fig. 3 Different samples from Kimia-99 dataset having skeletons extracted and circles/lines sequence embedded

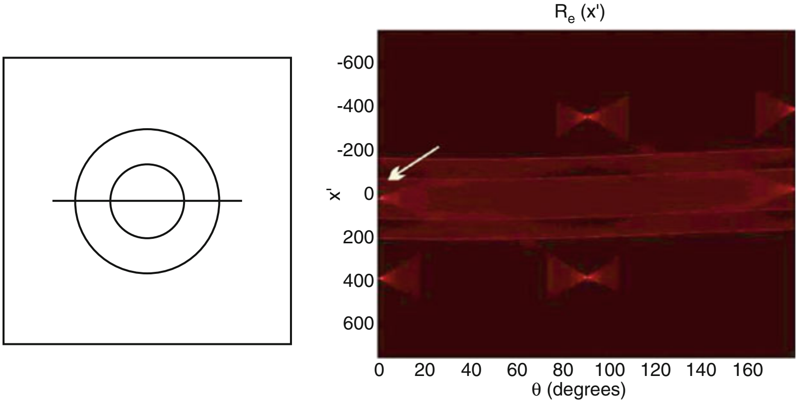


Fig. 4 Sample line and circles embedded in a single square (*left*) and its Radon transform (*right*). The arrow indicates the internal line position. In this example, a line is found at zero degrees

Proposed Approach: Testing Blocks' Location After Geometric Transformations

In this experiment, the robustness of the skeletal approach, in localizing the rectangular blocks, when different deformation scenarios are applied, is evaluated. For this purpose a series of random rigid transforms (scale/rotation) was applied to the Kimia-99 [29] database (Fig. 3). The Kimia-99 dataset consists of 99 shapes grouped in 9 classes with 11 shapes in each class. The database has a fair amount of visual transformations (occlusion articulation and deformation of parts). To assess the ability of the proposed method to perform efficiently under the above scenarios the following experimental setup is adopted; after the skeleton of each image object was extracted, a series of circles and lines with special properties are embedded along the skeleton's Eulerian tour.

More specifically, each block contained two concentric circles and one line passing through their center. The specific line rotates from square to square by an angle of 25 degrees (Figs. 4 and 5). The aim was to retrieve the same proportion of square side versus circle radius, as well as the same angle of the embedded line for the whole helical sequence of compared squares (when comparing original versus transformed). To retrieve the embedded circles and lines, we used two well known for their efficiency methods, namely the Radon [30] and the Hough transform [31]. We randomly rotated, shrunk, or enlarged the samples and thereafter attempted to retrieve the embedded sequences. As mentioned above, the purpose of this experiment is to show that the same sequence of rotated lines and circles (having the same square side/radius proportion) is preserved in the transformed image. Circles are detected by the Hough transform while lines inside the circles are detected by the Radon transform (Fig. 5). Examples of Hough transform for a small sequence

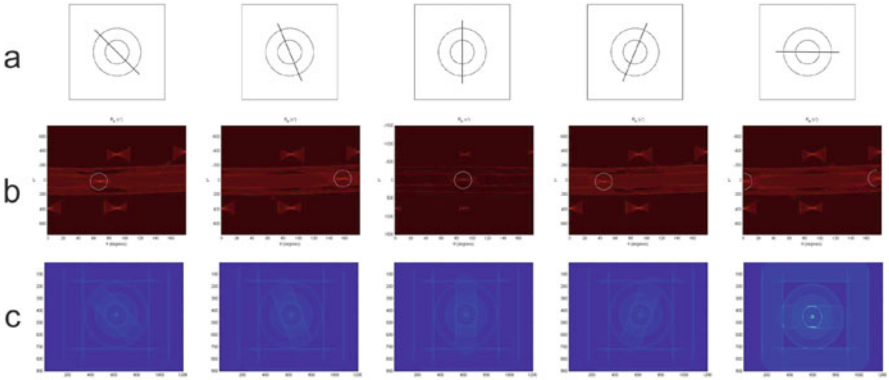


Fig. 5 Extracted Radon (**b**) and Hough (**c**) transforms for a short sequence of squares (**a**). For illustration purposes, the embedded line retrieved was marked with a circle. The degrees of the line found by the Radon transform, as well as the circles found by Hough transform (**c**) are shown

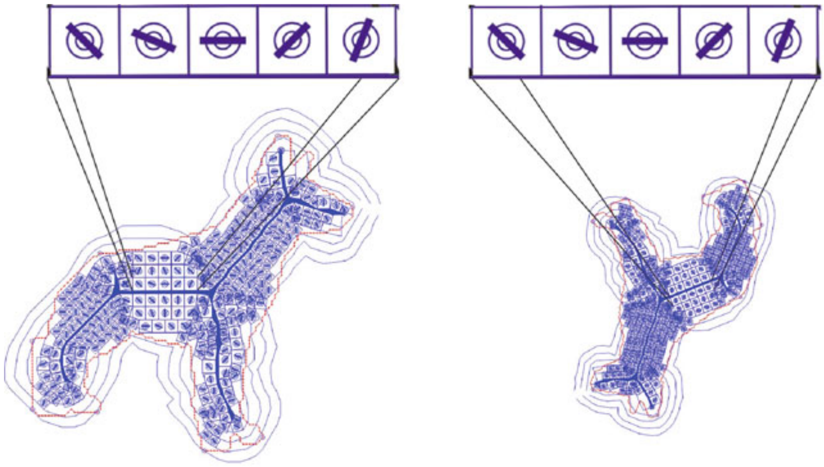


Fig. 6 A normal (*left*) and a deformed (*right, rotated and scaled*) sample from Kimia-99 dataset. The magnification illustrates how a random sequence is preserved in the attacked image

are shown in Fig. 5c. The results show a very robust behavior (Fig. 6). Comparing the sequences retrieved from all of Kimia-99 images resulted in a total of 26,730 squares, where 99.7% were matched.

Proposed Approach: Testing Watermark Retrieval Under Common Attacks

In this section, the robustness of the proposed method is investigated under typical attacks. The object database which is used for this purpose is the Caltech database, which contains 101 categories of objects and where object boundaries are readily available. Even though the proposed method is better suited for an authentication/integrity scenario, the case of tracking can also be considered with the difference that the object's boundary may need to be extracted before the MAT transformation can be applied. In such a case, techniques like Graph Cuts [32] or Snakes [33] can be utilized with great success. The steps of the watermark embedding procedure are:

1. Find the robust skeleton from the boundary and save it as a binary image.
2. Find the Eulerian tour around the skeleton as the contour of the skeleton image.
3. Re-sample the Eulerian tour to the desired equally spaced number of points.
4. For every pair of points on the Eulerian tour calculate the rotated rectangular block that has lower left and right corners on these points.
5. Calculate the next outward Eulerian tour from the upper left and right corners of the blocks, calculated in the previous step.
6. Keep calculating new Eulerian tours by extending outwards towards the object's boundary, until the whole object's region is covered.
7. Embed the watermark in the positions of the above calculated blocks, starting from a skeleton end point and following the clockwise/outwards direction of the combined Euler tours.

Since blocks' locations and sequencing are defined in step 4 of the above procedure, the watermark values are embedded in the DC coefficients of these blocks. The important is that under our approach the location of the watermarked blocks is now scale and rotation invariant.

During watermark extraction, the authentication module extracts a sequence of values from the DC coefficients of appropriate blocks. In this sequence the watermark is repeated more than once. Then, during the correlation phase this sequence is split into non-repetitive parts that correspond to the initial watermark. Having partitioned the sequence into non-repetitive parts, the original watermark is correlated to each part and the object is authenticated even for a single correlation value over the threshold.

In the experiments that follow, the proposed approach is tested under various attacks and different combinations. For this purpose, in each sample of all Caltech's categories, we apply geometric distortions to the watermarked objects, such as rotation ($30°, 45°, 60°, 90°$) and scaling (40%, 60%, 80%, 120%), filtering (Gaussian filtering with mean value (m=0) and standard deviations σ (0.05, 0.50, 1) as well as median filtering with windows (3x3, 5x5, 7x7)) and compression (JPEG compression (30%, 60%,100%). To evaluate the overall system performance, we compare to the *Fourier-Mellin* approach, one of the few and most prominent

Table 1 Precision-recall comparisons after different geometric, frequency, and domain related attacks

Category		Airplanes (801 samples)				Motorbikes (798 samples)			
		Proposed		Fourier-Mellin		Proposed		Fourier-Mellin	
Method		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Rotation	30°	0.99	0.99	0.99	0.98	1	0.99	0.99	0.97
	45°	0.99	0.99	0.98	0.97	0.99	0.99	0.98	0.97
	60°	0.99	0.98	0.97	0.96	0.99	0.98	0.98	0.96
	90°	0.99	0.97	0.97	0.95	0.99	0.98	0.97	0.96
Scaling	40%	0.99	0.98	0.98	0.97	0.99	0.98	0.98	0.97
	60%	0.99	0.98	0.98	0.97	0.99	0.98	0.98	0.97
	80%	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98
	120%	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.98
Gaussian filtering m=0	$\sigma = 0.05$	0.98	0.97	0.97	0.95	0.98	0.96	0.96	0.94
	$\sigma = 0.50$	0.96	0.94	0.95	0.91	0.97	0.95	0.93	0.90
	$\sigma = 1.00$	0.93	0.88	0.89	0.81	0.94	0.92	0.90	0.88
Median filtering	3 × 3	0.98	0.96	0.97	0.94	0.98	0.95	0.97	0.92
	5 × 5	0.97	0.94	0.94	0.89	0.96	0.93	0.95	0.90
	7 × 7	0.95	0.92	0.92	0.84	0.94	0.90	0.93	0.89
JPEG compression	30%	0.91	0.87	0.85	0.79	0.92	0.89	0.87	0.82
	60%	0.95	0.92	0.87	0.81	0.96	0.94	0.89	0.85
	100%	0.96	0.95	0.92	0.86	0.96	0.95	0.93	0.89

The proposed method is evaluated against the Fourier-Mellin technique, using two typical object classes from the Caltech database

methods in the literature that provides RST invariance to block-based approaches. The comparison is performed by means of precision and recall rates. The recall is the ratio of the correctly extracted to the total objects in the class. The precision is the ratio of the correctly extracted to the total that have been tested. In Table 1 the results of the comparison are reported. Our implementation of the Fourier-Mellin algorithm was as follows:

1. Apply a Fourier transform (FT), to the extracted object of the initial image, this provides the translation invariance.
2. Perform a conversion of FT to log-polar coordinates. This converts the scale and rotation differences to vertical and horizontal offsets which can be measured.
3. Perform a second FFT, called the Mellin transform (MT). This gives a transformed object that is invariant to translation, rotation, and scale.
4. Insert the watermark at the values of this transform-space object.
5. Perform the inverse procedure to produce the watermarked object.

The comparison between the two methods of Table 1 leads to the following conclusions: (a) as far as rotation and scaling transformations are concerned, the FMT could be robust enough during watermark extraction and (b) the proposed method is more robust not only to RST transformation but also to filtering and

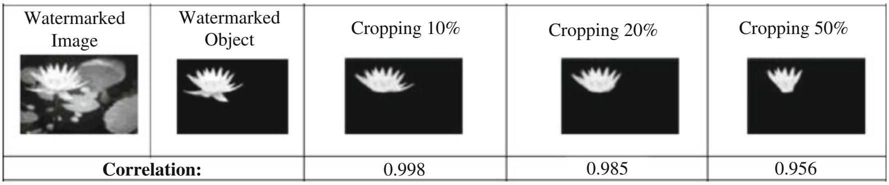


Fig. 7 Correlation measurements under cropping attacks

compression attacks, where the FMT fails. As far as the performance of the embedding method is concerned, the FMT method's response is 15 s on a typical PC (Core 2 CPU in 2.13 GHz, 2 GB RAM) while the proposed method's response on the same PC is 8 s. The proposed method efficiently confronts the copy-paste attack, since each candidate object passes through the watermark extraction procedure which is robust to cropping inaccuracies. As already mentioned, the watermark is redundantly embedded along the extended Eulerian tour. A single non-repetitive part is adequate to provide object authentication. Results of cropping inaccuracies are given in Fig. 7, using a watermarked water lily flower from the Caltech collection.

We compute correlation values for the watermark sequences extracted from the cropped and original objects. Even for 50% cropping the proposed system successfully authenticates the watermarked object. However, for cropping values over 70%, the authentication module fails. Nevertheless, in these cases the distortion is so severe that the resulting objects may not qualify as originals and thus protection is not necessary.

Synopsis: A Geometric Invariant Watermarking Scheme for Image Regions

A robust watermarking scheme for image objects has been proposed. The motivation for this approach lies in the inability of current frame-oriented schemes to protect semantic content. A watermark sequence is produced by a pseudo-random number generator and it is redundantly embedded into the object along Eulerian tours around a robust skeleton. During extraction, object authentication is achieved by splitting the extracted sequence into non-repetitive parts that correspond to the initial watermark. The original watermark is correlated to each part and the object is authenticated by means of correlation values over the threshold. The success of the proposed method stems from the robustness of the specific skeleton approach in relation to both RST and cropping attacks. Experimental results, involving benchmark datasets, illustrate the advantages of the proposed method over the state of the art with respect to various signal distortions, mixed processing, and copy-paste attacks.

References

1. J. Cox, M.L. Miller, J.A. Bloom, *Digital Watermarking* (Morgan Kaufmann, San Mateo, 2001)
2. N. Bi, Q. Sun, D. Huang, Z. Yang, J. Huang, Robust image watermarking based on multiband wavelets and empirical mode decomposition. IEEE Trans. Image Process. **16**(8), 1956–1966 (2007)
3. I.J. Cox, J. Kilian, F.T. Leighton, T. Shamoon, Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Process. **6**(12), 1673–1687 (1997)
4. M. Alghoniemy, A.H. Tewfik, Geometric invariance in image watermarking. IEEE Trans. Image Process. **13**(2), 145–153 (2004)
5. P. Dong, J. Brankov, N. Galatsanos, Y. Yang, Geometric robust watermarking based on a new mesh model correction approach. in *IEEE International Conference on Image Processing ICIP* (Rochester, New York, 2002)
6. S. Kay, E. Izquierdo, Robust content based image watermarking. in *WIAMIS 2001 - Workshop on Image Analysis for Multimedia Interactive Services, Tampere, Finland*, 16–17 May 2001
7. J. Lichtenauer, I. Setyawan, T. Kalker, R. Lagendijk, Exhaustive geometrical search and false positive watermark detection probability, in *SPIE Electronic Imaging 2002, Security and Watermarking of Multimedia Contents V, Santa Clara*, January 2003
8. S. Baudry, P. Nguyen, H. Maître, Estimation of geometric distortions in digital watermarking, in *IEEE- ICIP 2002, Rochester*, September 2002
9. D. Delannay, B. Macq, 2-D Periodic patterns for image watermarking, in *Advanced Sciences and Technologies for Security Applications*, vol. 1 (Springer, New York, 2005), pp. 297–18
10. M. Alghoniemy, A.H. Tewfik, Progressive quantized projection watermarking scheme, in *Proceedings of the 7th ACM International Multimedia Conference, Orlando, FL* (1999), pp. 295–298
11. J. Wang, G. Liu, Y. Dai, J. Sun, Z. Wang, S. Lian, Locally optimum detection for Barni's multiplicative watermarking in DWT domain. Signal Process. **88**(1), 117–130 (2008)
12. W. Lu, H. Lu, F.-L. Chung, Feature based watermarking using watermark template match. Appl. Math. Comput. **177**, 377–386 (2006). Elsevier
13. C. Serdean, M. Ambroze, M. Tomlinson, G. Wade, Dwt based video watermarking for copyright protection invariant to geometrical attacks, in *International Symposium on Communication Systems, Networks and Digital Signal Processing, Staffordshire University, UK*, 15–17 July 2002
14. F. Deguillaume, S. Voloshynovskiy, T. Pun, A method for the estimation and recovering from general affine transforms in digital watermarking applications, in *SPIE Electronic Imaging 2002, Security and Watermarking of Multimedia Contents IV, San Jose*, February 2002
15. J.J.K.O. Ruanaidh, T. Pun, Rotation, scale and translation invariant spread spectrum digital image watermarking. Signal Process. **66**, 303–318 (1998)
16. B.-S. Kim, J.-G. Choi, K.-H. Park, *RST-Resistant Image Watermarking Using Invariant Centroid and Reordered Fourier-Mellin Transform*. Lecture Notes in Computer Science, vol. 2939 (Springer, New York, 2004), pp. 370–381
17. C.-Y. Lin, Public watermarking surviving general scaling and cropping: an application for print-and-scan process, in *Multimedia and Security Workshop at ACM Multimedia 99, Orlando, FL*, October 1999
18. H. Liu, M. Steinebach, Non-ubiquitous watermarking for image authentication by region of interest masking, in *Proceedings of the Picture Coding Symposium, Portugal* (2007)
19. K. Zebbiche, F. Khelifi, Region-based watermarking of biometric images: case study in fingerprint images. Int. J. Digital Multimed. Broadcast. **2008**, 1–13 (2008)
20. X. Guo, T.-G. Zhuang, A region-based lossless watermarking scheme for enhancing security of medical data. J. Digit. Imaging **22**(1), 53–64 (2009). Springer

21. K. Raftopoulos, K. Ntalianis, P. Tzouveli, N. Tsapatsoulis, A. Parker-Wood, M. Ferecatu, Watermark resynchronization: an efficient approach based on Eulerian tours around a robust skeleton, in *Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany*, 25–26 September 2013
22. C.Y. Lin, M. Wu, J.A. Bloom, I.J. Cox, M.L. Miller, Y.M. Lui, Rotation, scale, and translation resilient watermarking for images. IEEE Trans. Image Process. **10**(5), 767–82 (2001)
23. R.L. Ogniewicz, *Discrete Voronoi Skeletons* (Swiss Federal Institute of Technology, Zurich, 1992)
24. T. Sebastian, P. Klein, B. Kimia, Recognition of shapes by editing their Shock graphs. IEEE Pattern Anal. Mach. Intell. **26**, 551–57 (2004)
25. J. Viega, Practical random number generation in software, in *Proceedings of the 19th Annual Computer Security Applications Conference*, December 2003
26. J.A. Sethian, A fast marching level set method for monotonically advancing fronts. Proc. Natl. Acad. Sci. **93**(4), 1591–1595 (1996)
27. J.A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd edn. (Cambridge University Press, Cambridge, 1999)
28. P. Kapsalas, S. Kollias, Affine morphological shape stable boundary regions (SSBR) for image representation, in *Proceedings of International Conference on Image Processing (ICIP) 2011, Brussels*, September 2011
29. T.B. Sebastian, P. Klein, B.B. Kimia, Recognition of shapes by editing Shock graphs. Proc. Int. Conf. Comput. Vis. **1**, 755–762 (2001)
30. S.R. Deans, *The Radon Transform and Some of Its Applications* (Krieger Publishing Company, Malabar, 1983)
31. R.O. Duda, P.E. Hart, Use of the hough transformation to detect lines and curves in pictures. Commun. Assoc. Comput. Mach. **15**(1), 11–15 (1972)
32. V. Kolmogorov, R. Zabih, What energy functions can be minimized via graph cuts? IEEE Trans. Pattern Anal. Mach. Intell. **26**(2), 147–159 (2004)
33. M. Kass, A. Witkin, D. Terzopoulos, Snakes: active contour models. Int. J. Comput. Vis. **1**(4), 321–331 (1988). Springer

Optimal Inventory Policies for Finite Horizon Inventory Models with Time Varying Demand: A Unified Presentation

Konstantina Skouri, Lakdere Benkherouf, and Ioannis Konstantaras

Abstract This paper aims to put forward a general framework for derivation of optimal control policies for inventory systems with time varying demand over a finite planning horizon. This permits the treatment of a large number of known inventory problems in a unified manner. As decision variables are considered the number of cycles and the times that each cycle starts and ends, where the term cycle can be used to represent various operational activities in inventory control. If the objective function, for a fixed number of cycles, passes successfully a couple of tests, then existence and uniqueness of a solution of the corresponding optimization problem is guaranteed. In this case, the search for the optimal solution reduces to a univariate search problem on a bounded interval. This together with a convexity (like) property leads to the optimal inventory policy.

Keywords Inventory • Optimization • Finite horizon • Time varying demand

2010 Mathematics Subject Classification: 90B05; 90C30

K. Skouri (✉)
Department of Mathematics, University of Ioannina, Ioannina, Greece
e-mail: kskouri@uoi.gr

L. Benkherouf
Faculty of Science, Department of Statistics and Operations Research, Kuwait University, Safat, Kuwait
e-mail: lakdere.benkherouf@ku.edu.kw

I. Konstantaras
Department of Business Administration, School of Business Administration, University of Macedonia, Thessaloniki, Greece
e-mail: ikonst@uom.gr

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_16

Introduction

It is well known that inventory models with time varying demand encompass a broad range of practical situations including multiechelon assembly operations, production to contract, products with seasonal demand, etc. (Silver et al. [1]). The optimal inventory policy for such models consists of determining the number of orders and the ordering times.

The finite horizon deterministic lot size problem with constant demand is found in Carr and Howe [2]. The linearly time varying demand rate is treated in Resh et al. [3] and Donaldson [4]. Barbosa and Friedman [5] examined inventory models with power form of time demand function. The general time varying demand lot sizing model was examined in Henery [6] where he showed that for a given number of orders, an optimal unique timing of the ordering exists and is unique. Friedman [7] showed, under the assumption of uniqueness of the ordering times, the corresponding optimal cost function is convex in the number of orders. It turns out that the convexity result can be obtained under very general conditions on the cost function generated between orders, see Denardo et al. [8]. A fundamental general result which appears to have attract little attention in the inventory literature. Yao and Klein [9] extended the result of Denardo et al. [8] to economic lot size models with backlogging.

Models with finite replenishment rates were considered in Friedman [10]. A model with finite replenishment rate is a generalization of the EPQ model to finite horizon and it is also called batching model. Hill [11] examined the issue of finding batching policies for linear increasing demand. Hill et al. [12] suggested a dynamic programming formulation for the model. Omar and Smith [13] studied the model to allow for the integration of raw materials with that of production in certain manufacturing systems. Rau and Ou Yang [14] derived the optimal inventory policy for Omar and Smith's model. This was later extended in Al-Khamis et al. [15] to the case when demand is log-concave.

The present chapter aims to provide a tool in order the optimization problems, which follows from the inventory models mentioned above (and others models, which will be presented later on), to be faced in a unified way, using a theory developed in Benkherouf and Gilding [16]. The result in [16] ensures the existence of a unique optimal replenishment schedule, that specifies the number of cycles (number of orders/setups) and the times that a cycle starts and ends (timing of order/setup). As a matter of fact, the results in [16] were motivated by lot sizing type inventory models. Applications of those results to EPQ models and others are found a later time. The results presented in this paper are stated without proof. Interested readers may consult [16] for the initial main results and the corresponding references for their required modifications in order to be applicable in more complex inventory problems.

The next section contains a specific optimization problem together with conditions that ensures the existence and the uniqueness of its solution. This optimization problem is used as prototype problem for the search for optimal inventory policies for a number of inventory systems with finite planning horizon. Then, these inventory systems are reviewed. The conclusions with some general remarks are found in last section.

A Specific Optimization Problem

The optimization problem, which is considered below, will serve as a prototype formulation for all inventory models that are examined in this paper. Let

$$P_1 : \min_{(t_1,t_2,\ldots,t_n,n)} \quad v_n c_0 + \sum_{i=1}^{n} R_i(t_{i-1}, t_i)$$

subject to

$$\begin{aligned} &0 = t_0 \le t_1 \le \ldots \le t_n = H, \\ &n \ \text{ integer,} \end{aligned}$$

where $c_0 > 0, H > 0$ and known. Also, v_n is a function of n and for $1 \le i \le n$, R_i is a function defined on some subset $\Omega \subset \mathbb{R}$ with

$$\Omega = \{(x, y) : 0 \le x < y \le H\}.$$

For the solution of the above class of problems a general theory for the existence of unique optimal solution has been proposed by Benkherouf and Gilding [16] under two hypotheses:

Let

∂_x : The partial derivative of a bivariate function with respect to the first variable.
∂_y : The partial derivative of a bivariate function with respect to the second variable.
∂_x^2 : The second derivative of a bivariate function with respect to the first variable.
∂_y^2 : The second derivative of a bivariate function with respect to the second variable.
$\partial_x \partial_y$: The cross partial derivatives of a bivariate function.

Hypothesis 1 For every $i \ge 1$ the function $R_i \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ is such that

(1) $R_i > 0$
(2) $R_i = 0$, on $\Omega \backslash \bar{\Omega}$
(3) $\partial_x R_i < 0 < \partial_y R_i$ in Ω
(4) $\partial_x \partial_y R_i < 0$ in Ω

Hypothesis 2 For all $1 \leq i \leq n - 1$ there hold

(1) $\partial_y R_i + \partial_x R_{i+1}$ on $\Omega \backslash \bar{\Omega} = 0$
(2) there exits a function $f \in C(0, H)$ such that $\mathrm{L}_x R_{i+1} \geq 0$ and $\mathrm{L}_y R_i \geq 0$ in Ω where

$$\mathrm{L}_x z := \partial_x^2 z + \partial_x \partial_y z + f(x) \partial_x z$$

$$\mathrm{L}_y z := \partial_y^2 z + \partial_x \partial_y z + f(y) \partial_x z.$$

It is worth noting that part (4) in Hypothesis 1 is key in the analysis of finite horizon models. It is equivalent for the requirement that R_i is submodular in Ω for functions that are twice differentiable. This condition has been used successfully in [8] and [9] to show convexity of some value function with respect to the number of orders. A useful property for deriving the optimal frequency of orders.

The main results of the theory found in [16] are briefly outlined below:

Let

$$S_n(t_1, t_2, \ldots, t_n) = \sum_{i=1}^{n} R_i(t_{i-1}, t_i) \tag{1}$$

Theorem 1 *For fixed n and under Hypotheses 1 and 2 the function S_n has a unique minimum satisfying $0 = t_0 \leq t_1 \leq \ldots \leq t_n = H$. The solution is found by setting*

$$\nabla S_n = 0.$$

Theorem 2 *Let s_n denotes the minimum value of S_n with respect to t_0, t_1,..., t_n satisfying $0 = t_0 \leq t_1 \leq \ldots \leq t_n = H$.*

(i) Then s_n is a strictly decreasing function of $n \geq 1$.
(ii) If there exists an integer $p \geq 1$ such that $R_{j+p} = R_j$ for all $j \geq 1$, then $s_n - s_{n+p}$ is a strictly decreasing function of $n \geq 1$.

Note that for $p = 1$ the Theorem 2, (ii) is equivalent to the convexity of s_n in n. The cases $p = 1$ and $p = 2$ can be found in [8] and [9], respectively, without the need for uniqueness result of Theorem 1. Moreover, strictly speaking existence of the optimal solution and Theorem 2 apply under Hypothesis 1. Hypothesis 2 guarantees uniqueness of the optimal solution. It also turns out that in general it is the hardest hypothesis to check. We shall comment on the applicability of these hypothesis in the next section when specific models are discussed.

Inventory Models with Time Varying Demand Over Finite Horizon

In this section, inventory models with time varying demand over finite horizon will be examined and they will be treated using Problem P_1. Models with infinite replenishment rate (ELS models) are discussed first followed by finite replenishment rate models (EPQ models).

Assumptions

The common assumptions behind existing models are the following:

(1) The planning horizon of the system is finite and is taken as H time units, $H > 0$. The initial and the final inventory levels are both zero.
(2) The demand rate at time t is given by a continuous function D, $D(t) : [0, H] \rightarrow (0, \infty)$.
(3) The cost structure is (a) a fixed order/setup cost per order/setup, c_0, (b) a holding cost per unit in stock per unit of time c_h, (c) a purchasing cost per unit c_p.
(4) The lead time is zero.

Inventory Models with Infinite Replenishment Rate

In this section inventory models with infinite replenishment rate will be presented. In Fig. 1 a possible representation of the inventory level during the planning horizon is given. The symbol n denotes the number of cycles (orders) and t_i the time at which the inventory level, $I(t)$, reaches zero for cycle i, say, with $t_0 = 0$ and $t_n = H$, $i = 1, \ldots n$.

During the time interval $[t_{i-1}, t_i]$ (cycle i), the level of stock is described by the following equation:

$$\frac{dI(t)}{dt} = -D(t), \; t_{i-1} \leq t \leq t_i, I(t_i) = 0 \tag{2}$$

The problem of finding the optimal replenishment schedule reduces to solving the mixed integer non-linear program P_1, where

$v_n = n$ (number for cycles) and

$$R_i(t_{i-1}, t_i) := R_h(t_{i-1}, t_i) = c_p \int_{t_{i-1}}^{t_i} D(t)dt + c_h \int_{t_{i-1}}^{t_i} (t - t_{i-1})D(t)dt \tag{3}$$

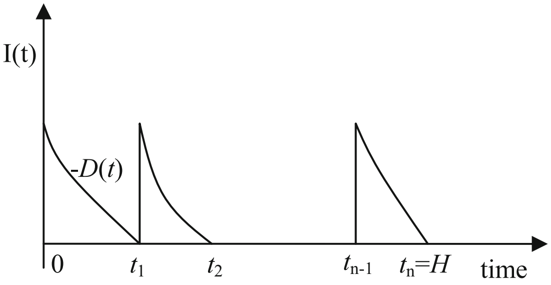


Fig. 1 A finite horizon inventory model with infinite replenishment rate

Here R_i represents the inventory costs (purchase+holding) in period i. It is also clear that the subscript i in the function R_i in (2) can be dropped from the notation.

Note that depending on function of the demand rate $D(t)$ in (3), various existing models can be recovered: the model of Donaldson [4] if $D(t)$ is linear in t, the model of Henery [6] if the demand rate is log-concave in t.

Problem P_1 and the corresponding theoretical results can also be used if shortages are allowed, by setting $v_n = n/2$ and modifying appropriately the functions R_i. Specifically, the functions R_i is equal to (3) for odd i and

$$R_i(t_{i-1}, t_i) = R_s(t_{i-1}, t_i) = c_s \int_{t_{i-1}}^{t_i} (t_i - t)D(t)dt \tag{4}$$

for even i, where c_s refers to the shortages cost per unit of unsatisfied demand per unit of time. If the demand rate in (3) and (4) is linear in t, then the model of Goyal et al. [17] and Bhounia and Maiti [18] is obtained.

We shall next comment briefly on the extent to which Hypotheses 1 and 2 are satisfied by the general model with shortages.

It is easy to check that Hypothesis 1 is satisfied. If i is odd then

$$R_i(x, y) = c_p \int_x^y D(t)dt + c_h \int_x^y (t - x)D(t)dt.$$

Indeed, on Ω, $R_i(x, y) > 0$, $\partial_x R_i(x, y) = -c_p D(x) - c_h \int_x^y D(t)dt < 0$, $\partial_y R_i(x, y) = c_p D(x) + c_h(y - x)D(y) > 0$, $\partial_x \partial_y R_i = -c_h D(y) < 0$. Also, $R_i(x, x) = 0$. Likewise, the case i even can be checked in a similar fashion. Therefore, Hypothesis 1 is satisfied. This means that at this stage we can safely assert that an optimal inventory policy exists. Also, Theorem 2 applies.

Part (1) of Hypothesis 1 can easily be checked. Let $f(t) = -D'(t)/D(t)$, then it can be shown that *for i* odd $\mathrm{L}_y R_i(x, y) = 0$, and

$$\mathrm{L}_x R_i(x, y) = c_h \left[\frac{D'(x)}{D(x)} \int_x^y D(t)dt - \{D(y) - D(x)\} \right].$$

The expression $\mathrm{L}_x R_i(x, y)$ can be shown to be ≥ 0 if D is log-concave. For more details see: [6] and [16]. Hence Theorems 1 and 2 apply.

The models with infinite replenishment rate below (and their extensions) have objective functions, corresponding to the optimal inventory policy, which satisfy the submodularity condition of Hypothesis 1. Therefore, an optimal inventory policy exists and Theorem 2 holds.

Relations (3) and (4) can be modified to cater for various existing inventory systems to be considered. For example, in order to model systems with time varying deterioration rate and partial backlogging of unsatisfied demand, set $v_n = n/2$ and let $R_i(t_{i-1}, t_i)$ be defined by:

$$R_i(t_{i-1},t_i) = R_h(t_{i-1},t_i) = c_p \int_{t_{i-1}}^{t_i} e^{\delta(u)-\delta(t_{i-1})} D(t)dt$$
$$+ c_h \int_{t_{i-1}}^{t_i} \int_{t}^{t_i} e^{\delta(u)-\delta(t)} D(u)dudt \tag{5}$$

for odd i, and where

$$\delta(t) = \int_0^t \theta(u)du,$$

and $\theta(u)$ the deterioration rate. Also,

$$R_i(t_{i-1},t_i) = R_s(t_{i-1},t_i) = c_p \int_{t_{i-1}}^{t_i} \beta(t_i - t)D(t)dt$$
$$+ c_s \int_{t_{i-1}}^{t_i} (t_i - t)\beta(t_i - t)D(t)dt + c_l \int_{t_{i-1}}^{t_i} (1 - \beta(t_i - t))D(t)dt \tag{6}$$

for even i and $v_n = n/2$, where c_l is the lost sales cost per unit time

Note that relations (5) and (6) can be interchanged for odd and even i leading to different classes of policies (see Skouri and Papachristos [19] and [16]). Now, by assuming specific functions for demand, deterioration and backlogging rates, several inventory models can be obtained as those found, for example, in: Barbosa and Friedman [20], Dave [21], Teng [22], Teng [23], Chakrabarti and Chaudhuri [24], Chakrabarti et al. [25], Chang and Dye [26], Teng et al. [27], Wee and Mercan [28], Papachristos and Skouri [29], Teng et al. [30], Skouri and Papachristos [31].

It is worth noting that the extension of the above model by considering inflation has a drastic effect on the solution of the optimization problem as Gilding [32] point out. Gilding [32] proved that under equitable hypotheses when the number of replenishment cycles is sufficiently large, the optimal solution involves the placement of token orders at the end of the planning period. Then by assuming specific function for the involved parameters of the systems various existing models can be obtained (i.e. Bose et al. [33], Chandra and Bahner [34], Chern et al. [35], Chung et al. [36], Datta and Pal [37], Dye and Hsieh [38], Hariga [39], Hsieh and Dye [40], Moon et al. [41], Yang et al. [42], Yang et al. [43]).

Problem P_1 can also be used to model vendor-buyer coordination problems. Assuming a single production run for the vendor and multiple replenishment for buyer, the functions $R_i(t_{i-1},t_i)$ are modified as (see Benkherouf and Omar [44]):

$$R_1(t_{i-1},t_i) = \int_0^{t_1} (h_2 t + h_1(t_1 - t)D(t)dt \tag{7}$$

and

$$R_i(t_{i-1}, t_i) = \int_{t_{i-1}}^{t_i} (h_2 - h_1)(t - t_{i-1})D(t)dt \tag{8}$$

for $i \geq 2$,
where h_1 the holding cost for the vendor and h_2 the holding cost for the buyer. By assuming $D(t)$ to be linear decreasing in t then the model of Omar [45] is obtained.

Inventory Models with Finite Replenishment Rate

In this section the problem P_1 is shown that it can also be used for modeling inventory models with finite replenishment. In Fig. 2 a possible representation of the inventory level during the planning horizon is given.

In a cycle i, say, i.e. during the time interval $[t_{i-1}, t_i]$, the level of stock is described by the following differential equations:

$$\frac{dI(t)}{dt} = p - D(t),\ t_{i-1} \leq t \leq t_i^p, \quad \text{with } I(t_{i-1}) = 0, \tag{9}$$

$$\frac{dI(t)}{dt} = -D(t),\ t_i^p \leq t \leq t_i, \quad \text{with } I(t_i) = 0 \tag{10}$$

Then, using the continuity of the inventory level at t_i^p, the function R_i becomes:

$$\begin{aligned} R_i(t_{i-1}, t_i) &:= R_p(t_{i-1}, t_i) \\ &= c_p \left\{ \int_{t_{i-1}}^{t_i} D(t)dt \right\}^2 + c_h \left[\int_{t_{i-1}}^{t_i} (t - t_{i-1})D(t)dt - \frac{1}{2p} \left\{ \int_{t_{i-1}}^{t_i} D(t)dt \right\}^2 \right] \end{aligned} \tag{11}$$

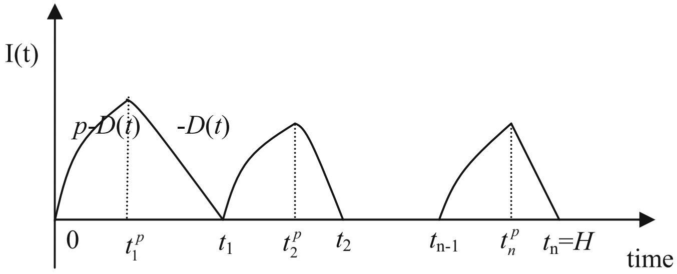


Fig. 2 A finite horizon inventory model with finite replenishment rate

It is an easy exercise to see that the function R_i defined in (11) passes the tests for Hypothesis 1. Hypothesis 2 requires slightly more effort but is doable under some technical conditions which includes the log-concavity of the demand rate function (see Al-Khamis at al. [15]). The case when $D(t)$ is linear in t, the model of Hill [11] is recovered.

Production–inventory models with time varying demand and finite planning horizon have been developed for deteriorating items. Specifically, Balkhi [46] studied a production-inventory system assuming that production, demand and deteriorating rates are continuous function of time. Sana et al. [47] presented a model with constant production and deterioration rates and linear time dependent demand rate. Yang [48] extended the model of [46] assuming partial backlogging. Benkherouf and Boushehri [49] considered a model with constant deterioration rate with time varying demand rate.

The corresponding objective of P_1 for models with deteriorating rates is complex. Hypothesis 1 is easily checked. However, for Hypothesis 2 to be satisfied more elaborate technical conditions (other than the usual log-concavity condition) need to be imposed: see [49]. Again, existence of an optimal inventory policy is easily established, and Theorem 2 applies.

The above basic production–inventory model can be used as a basis for handling more complex inventory production problems under rework or remanufacturing options. Benkherouf et al. [50] studied a recovery inventory system. The remanufacturing process brings used products up to quality standards that are as rigorous as those of new products. The demand is satisfied either by new produced items or by the used items that are repaired back to an "as new" condition, before being sold again. At the beginning of the planning horizon, the demand is satisfied by new manufacturing lots. When the production stops, the demand is satisfied by remanufacturing lots. The solution of this problem requires the determination of the newly produced and remanufactured quantities that minimize the total cost over the planning horizon. To this end, two sub-problems should be solved and the optimal time point, τ_0, of switching from a manufacturing activity to a remanufacturing one should be determined. The two sub-problems, which evidently conform to the above problem P_1, are:

Sub-problem 1

$$\min \; nc_{0,1} + h_1 \sum_{i=1}^{n} \left[\int_{t_{i-1}}^{t_i} (t - t_{i-1}) D(t)dt - \frac{1}{2p} \left(\int_{t_{i-1}}^{t_i} D(t)dt \right)^2 \right]$$

$$+ h_2 \int_0^{\tau_0} (\tau_0 - t)\phi D(t)dt \tag{12}$$

subject to

$$0 = t_0 < t_1 < \ldots . < t_n = \tau_0,$$

with $n \geq 1$ and integer.

Sub-problem 2

$$\min \;\; (m-1)c_{0,2} \\ + (h_1 - h_2)\sum_{j=1}^{m-1}\left[\int_{\tau_{j-1}}^{\tau_j}(t-\tau_{j-1})D(t)dt - \frac{1}{2r}\left(\int_{\tau_{j-1}}^{\tau_j}D(t)dt\right)^2\right] \qquad (13)$$

subject to

$$\tau_0 < \tau_1 < \ldots . < \tau_m = H,$$

with $m \geq 2$, and integer. Here

r is the remanufacturing rate,
h_1 is the holding cost for the serviceable items,
h_2 is the holding cost for the used (recoverable) items,
$c_{0,1}$ is the production setup cost,
$c_{0,2}$ is the remanufacturing setup cost,
n is the number of production runs,
m is the number of remanufacturing runs,
t_{i-1} is the production starting time, $i = 1, \ldots, n$ with $t_0 = 0$,
τ_{j-1} is the remanufacturing starting time, $j = 1, \ldots, m$ with $\tau_0 = t_n$.

It is not surprising that both Sub-problems 1 and 2 conform to Hypothesis 1 without no extra assumptions on the model. Hypothesis 2 needs some extra effort: see [50].

Benkherouf et al. [51] studied an inventory system with production, remanufacturing and refurbishing activities. Used products are returned by customers and after inspection they can be classified either as "remanufacturable" or as "refurbishable" items. The remanufacturing process brings "remanufacturable" items up to quality standards that are as rigorous as those of new items. The refurbished items are sold to a secondary market at a reduced price. In order to control the system, two types of policies are considered, namely $P(1, n_2)$ and $P(n_1, 1)$. According to $P(1, n_2)$ policy, one remanufacturing batch and n_2 orders for new items are considered. According to $P(n_1, 1)$ policy it is assumed n_1 batches of remanufacturing and one batch of new items that ordered at time 0. For these two policies, the order and remanufacturing quantities and the inventory level of returned (used) items at the start of the inspection and recovery processes, which minimize the total cost, should be determined.

In order this problem to be solved the following optimization problems should be solved. The optimization problem that corresponds to $P(1, n_2)$ is:

$$\min \left[n_2 S + h_2 \sum_{i=1}^{n_2}\int_{t_{i-1}}^{t_i}(t - t_{i-1})D(t)dt\right]$$

subject to

$$0 = t_0 \leq \cdots \leq t_{n_2} = \tau_0,$$

where $n_2 \geq 0$, and integer.

The optimization problem that corresponds to $P(n_1, 1)$ is:

$$\min \left[(R+W)(n_1-1) + \varphi^2 h_1 \frac{(1-q)q}{2p} \sum_{i=1}^{n_1-1} \left\{ \int_{t_{i-1}}^{t_i} D(t)dt \right\}^2 \right]$$

$$+ \varphi(h_1 - (1-q)h_2) \sum_{i=1}^{n_1-1} \left[\int_{t_{i-1}}^{t_i} (t_i - t)D(t)dt - \varphi \frac{1-q}{2p} \left\{ \int_{t_{i-1}}^{t_i} D(t)dt \right\}^2 \right]$$

subject to

$$0 = t_0 \leq t_1 \leq \ldots t_{n_1-1} = \eta,$$

where $n_1 \geq 0$ and integer, and

φ is the return factor
p: is the remanufacturing rate
R: is remanufacturing setup cost
S: is the ordering cost of a batch of new items
W: is the fixed inspection and sorting charge
h_1: is the holding cost of used/refurbished items
h_2: is the holding cost of serviceable items
q: is the percentage of used items classified as refurbished

The same remarks made about the models treated in [50] apply here.

Conclusions

The objective of the present paper is to provide a unified treatment for the search for optimal inventory policies in deterministic inventory problems with time varying demand in a finite planning horizon. Keys in the analysis are two hypotheses called Hypotheses 1 and 2. The crucial element in Hypothesis 1 is a submodularity requirement. The treatment was illustrated with inventory models with finite and infinite replenishment rates where the submodularity requirement is satisfied with minimal assumptions on the models. As a matter of fact, Hypothesis 1 alone ensures existence of the optimal inventory policy together with the crucial property: Theorem 2. Hypothesis 2 checks the uniqueness of the optimal inventory policy. This hypothesis reduces for the basic Economic Lot Size model to requiring that the demand rate for the item be log-concave.

As a future research direction, it would be of interest to find new inventory models with optimal policies which can be cast in the optimization form of Problem P_1. Also, we believe that checking Hypothesis 2 in certain models may be technical. This could possibly be handicap for its applicability. May be a search for an alternative simpler requirement is worth pursuing.

References

1. E.A. Silver, D.F. Pyke, R. Peterson, *Inventory Management and Production Planning and Scheduling* (Wiley, New-York, 1998)
2. C.R. Carr, C.W. Howe, Optimal service policies and finite time horizons. Manag. Sci. **9**, 126–140 (1962)
3. M. Resh, M. Friedman, L.C. Barbosa, On a general solution of the deterministic lot size problem with time-proportional demand. Oper. Res. **24**, 718–725 (1976)
4. W.A. Donaldson, Inventory replenishment policy for a linear trend in demand–An analytical solution. Oper. Res. Q. **28**, 663–670 (1977)
5. L.C. Barbosa, M. Friedman, Optimal policies for inventory models with some specified markets and finite time horizon. Eur. J. Oper. Res. **8**, 175–183 (1981)
6. R.J. Henery, Inventory replenishment policy for increasing demand. J. Oper. Res. Soc. **30**, 611–617 (1979)
7. M.F. Friedman, Inventory lot–size models with general time–dependent demand and carrying cost functions. INFOR **20**, 157–167 (1982)
8. E.V. Denardo, G. Huberman, U.G. Rothblum, Optimal locations on a line are interleaved. Oper. Res. **30**, 745–759 (1982)
9. D. Yao, M. Klein, Lot sizes under continuous demand: the backorder case. Nav. Res. Logist. **36**, 615–624 (1989)
10. M.F. Friedman, On some optimality conditions and their equivalence in time dependent inventory models. Comput. Oper. Res. **11**, 245–251 (1984)
11. R.M. Hill, Batching policies for linearly increasing demand with a finite input rate. Int. J. Prod. Econ. **43**, 149–154 (1996)
12. R.M. Hill, M. Omar, D.K. Smith, Stock replenishment policy for deterministic linearly increasing demand with a finite input rate. J. Sains **8**, 977–986 (2000)
13. M. Omar, D.K. Smith, An optimal batch size for a production system under linearly increasing-time varying demand process. Comput. Ind. Eng. **42**, 35–42 (2002)
14. H. Rau, B.C. Ou Yang, A general and optimal approach for three inventory models with a linear trend in demand. Comput. Ind. Eng. **52**, 521–532 (2007)
15. T.M. Al-Khamis, L. Benkherouf, M.A. Omar, Optimal policies for a finite–horizon batching inventory model. Int. J. Syst. Sci. **45**, 2196–2202 (2014)
16. L. Benkherouf, B.H. Gilding, On a class of optimization problems for finite time horizon inventory models. SIAM J. Control. Optim. **48**, 993–1030 (2009)
17. S.K. Goyal, D. Morrin, F. Nebebe, The finite horizon trended inventory replenishment problem with shortages. J. Oper. Res. Soc. **43**, 1173–1178 (1992)
18. A.K. Bhunia, M. Maiti, An inventory model of deteriorating items with lot-size dependent replenishment cost and a linear trend in demand. Appl. Math. Model. **23**, 301–308 (1999)
19. K. Skouri, S. Papachristos, A continuous review inventory model, with deteriorating items, time-varying demand, linear replenishment cost, partially time-varying backlogging. Appl. Math. Model. **26**, 603–617 (2002)
20. L.C. Barbosa, M. Friedman, Deterministic inventory lot size models–a general root law. Manag. Sci. **24**, 819–826 (1978)

21. U. Dave, On a heuristic inventory replenishment rule for items with linearly increasing demand incorporating shortages. J. Oper. Res. Soc. **40**, 827–830 (1989)
22. J.-T. Teng, A note on inventory replenishment policy for increasing demand. J. Oper. Res. Soc. **45**, 1335–1337 (1994)
23. J.-T. Teng, A deterministic inventory replenishment model with a linear trend in demand. Oper. Res. Lett. **19**, 33–41 (1996)
24. T. Chakrabarti, K.S. Chaudhuri, An EOQ model for deteriorating items with a linear trend in demand and shortages in all cycles. Int. J. Prod. Econ. **49**, 205–213 (1997)
25. T. Chakrabarty, B.C. Giri, K.S. Chaudhuri, An EOQ model for items with Weibull distribution deterioration, shortages and trended demand: an extension of Philip's model. Comput. Oper. Res. **25**, 649–657 (1998)
26. H.-J. Chang, C.-Y. Dye, An EOQ model for deteriorating items with time varying demand and partial backlogging. J. Oper. Res. Soc. **50**, 1176–1182 (1999)
27. J.-T. Teng, M.-S. Chern, H.-L. Yang, Y.J. Wang, Deterministic lot-size inventory models with shortages and deterioration for fluctuating demand. Oper. Res. Lett. **24**, 65–72 (1999)
28. H.M. Wee, H.M. Mercan, Exponentially decaying inventory with partial back–ordering. Optim. Control Appl. Meth. **20**, 43–50 (1999)
29. S. Papachristos, K. Skouri, An optimal replenishment policy for deteriorating items with time varying demand and partial–exponential type– backlogging. Oper. Res. Lett. **27**, 175–184 (2000)
30. J.-T. Teng, H.J. Chang, C.Y. Dye, C.H. Hung, An optimal replenishment policy for deteriorating items with time-varying demand and partial backlogging. Oper. Res. Lett. **30**, 387–393 (2002)
31. K. Skouri, S. Papachristos, Optimal stopping and restarting production times for an EOQ model with deteriorating items and time dependent partial backlogging. Int. J. Prod. Econ. **81–82**, 525–531 (2003)
32. B.H. Gilding, Inflation and the optimal inventory replenishment schedule within a finite planning horizon. Eur. J. Oper. Res. **234**, 683–693 (2014)
33. S. Bose, A. Goswami, K.S. Chaudhuri, An EOQ model for deteriorating items with linear time-dependent demand rate and shortages under inflation and time discounting. J. Oper. Res. Soc. **46**, 771–782 (1995)
34. M.J. Chandra, M.L. Bahner, The effects of inflation and the time value of money on some inventory systems. Int. J. Prod. Res. **23**, 723–730 (1985)
35. M.-S. Chern, H.-L. Yang, J.-T. Teng, S. Papachristos, Partial backlogging inventory lot–size models for deteriorating items with fluctuating demand under inflation. Eur. J. Oper. Res. **191**, 127–141 (2008)
36. K.-J. Chung, J. Liu, S.-F. Tsai, Inventory systems for deteriorating items taking account of time value. Eng. Optim. **27**, 303–320 (1997)
37. T.K. Datta, A.K. Pal, Effects of inflation and time-value of money on an inventory model with linear time-dependent demand rate and shortages. Eur. J. Oper. Res. **52**, 326–333 (1991)
38. C.-Y. Dye, T.-P. Hsieh, Deterministic ordering policy with price- and stock dependent demand under fluctuating cost and limited capacity. Expert Syst. Appl. **38**, 14976–14983 (2011)
39. M.A. Hariga, Effects of inflation and time-value of money on an inventory model with time-dependent demand rate and shortages. Eur. J. Oper. Res. **81**, 512–520 (1995)
40. T.-P. Hsieh, C.-Y. Dye, Pricing and lot-sizing policies for deteriorating items with partial backlogging under inflation. Expert Syst. Appl. **37**, 7234–7242 (2010)
41. I. Moon, B.C. Giri, B. Ko, Economic order quantity models for ameliorating/deteriorating items under inflation and time discounting. Eur. J. Oper. Res. **162**, 773–785 (2005)
42. H.-L. Yang, J.-T. Teng, M.-S. Chern, Deterministic inventory lot-size models under inflation with shortages and deterioration for fluctuating demand. Nav. Res. Logist. **48**, 144–158 (2001)
43. H.-L. Yang, J.-T. Teng, M.-S. Chern, An inventory model under inflation for deteriorating items with stock-dependent consumption rate and partial backlogging shortages. Int. J. Prod. Econ. **123**, 8–19 (2010)

44. L. Benkherouf, M. Omar, Optimal integrated policies for a single-vendor single-buyer time-varying demand model. Comput. Math. Appl. **60**, 2066–2077 (2010)
45. M. Omar, An integrated equal-lots policy for shipping a vendor's final production batch to a single buyer under linear decreasing demand. Int. J. Prod. Econ. **118**, 185–188 (2009)
46. Z.T. Balkhi, On a finite production lot size inventory model for deteriorating items: an optimal solution. Eur. J. Oper. Res. **132**, 210–223 (2001)
47. S. Sana, S.K. Goyal, K.S. Chaudhuri, Production–inventory model for a deteriorating item with trended demand and shortages. Eur. J. Oper. Res. **157**, 357–371 (2004)
48. H.-L. Yang, A partial backlogging production–inventory lot–size model for deteriorating items with time–varying production and demand rate over a finite time horizon. Int. J. Syst. Sci. **42**, 1397–1407 (2011)
49. L. Benkherouf, D. Boushehri, Optimal policies for a finite–horizon production inventory model. Adv. Oper. Res. Article ID 768929, 16 pages. DOI:10.1155/2012/768929 (2012)
50. L. Benkherouf, K. Skouri, I. Konstantaras, Optimal lot sizing for a production–recovery system with time–varying demand over a finite planning horizon. IMA J. Manag. Math. **25**, 403–420 (2014)
51. L. Benkherouf, K. Skouri, I. Konstantaras, Optimal control of production, remanufacturing and refurbishing activities in a finite planning horizon inventory system. J. Optim. Theory Appl. **168**, 677–698 (2016)

Metrical Pareto Efficiency and Monotone EVP

Mihai Turinici

Abstract A uniform version is established for normed Pareto efficient point results in Isac [Comb. Global Optim., pp. 133–144, World Sci. Publ., 2002]. Its basic tool is the monotone variant of Ekeland's variational principle obtained in Turinici [An. Şt. UAIC Iaşi, 36 (1990), 329–352].

Keywords Inf-proper lsc function • Dependent Choice • Monotone variational principle • Generalized metric/uniform space • Pareto efficiency • Super-additive function

Introduction

Let (M, d) be a metric space; and $\varphi \in \mathscr{F}(M, R \cup \{\infty\})$ be some function with

(a01) φ is inf-proper ($\mathrm{Dom}(\varphi) \neq \emptyset$ and $\inf[\varphi(M)] > -\infty$)
(a02) φ is d-lsc: $\liminf_n \varphi(x_n) \geq \varphi(x)$, whenever $x_n \xrightarrow{d} x$;
or, equivalently:
$[\varphi \leq t] := \{x \in M; \varphi(x) \leq t\}$ is closed (modulo d), for each $t \in R$.

(Here, for each couple A, B of nonempty sets, $\mathscr{F}(A, B)$ stands for the class of all functions from A to B; in particular, if $A = B$, we write $\mathscr{F}(A)$ in place of $\mathscr{F}(A, A)$).

The following 1979 statement in Ekeland [16] (referred to as Ekeland's variational principle; in short: EVP) is our starting point.

Theorem 1 *Let the precise conditions hold. In addition, suppose that (M, d) is complete. Then, for each $u \in \mathrm{Dom}(\varphi)$ there exists $v = v(u) \in \mathrm{Dom}(\varphi)$ with*

*(**11-a**) $d(u, v) \leq \varphi(u) - \varphi(v)$ (hence $\varphi(u) \geq \varphi(v)$)*
*(**11-b**) $x \in M$, $d(v, x) \leq \varphi(v) - \varphi(x)$ imply $v = x$.*

M. Turinici (✉)
"A. Myller" Mathematical Seminar, "A. I. Cuza" University, Iaşi 700506, Romania
e-mail: mturi@uaic.ro

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_17

This principle found some basic applications to control and optimization, generalized differential calculus, critical point theory, and global analysis; we refer to the quoted (survey) paper for details. So, it cannot be surprising that, soon after its formulation, many extensions of (EVP) were proposed. For example, the *dimensional* way of extension refers to the ambient space (R) of $\varphi(M)$ being substituted by a (topological or not) vector space. An account of the results in this area is to be found in the 2003 monograph by Goepfert et al. [21, Chap. 3]. Further, the *(pseudo) metrical* one consists in conditions imposed to the ambient metric over M being relaxed. The basic result in this direction was obtained in 1992 by Tataru [42], via Ekeland type techniques; subsequent extensions of it may be found in the 1996 paper by Kada et al. [30]. Finally, another way of extending (EVP) is the *quasi-order* one, proposed in the 1990 paper by Turinici [45] (cf. section "Monotone EVP"). As we shall see, it is a handy tool for the study of Pareto efficiency over metric and uniform spaces. The metric/normed case will be discussed in section "Metrical Efficiency"; note that, by our precise methods, one gets a completion of the result in Isac [27] established for normed structures. The uniform version of it is given in section "Uniform Pareto Efficiency" (after some preliminaries in section "Fang Spaces"); precisely, we show that a related statement in Isac [27] is reducible to such metrical methods. Finally (after some preliminaries in section "Semigroup Anti-Measures") the semigroup versions of the metric/normed developments above are treated in section "Semigroup Pareto Efficiency"; in particular, this yields the result in Goepfert and Tammer [19], proved via different techniques. Finally, section "Preliminaries" has an introductory character. Further aspects will be discussed elsewhere.

Preliminaries

Throughout this exposition, the working axiomatic system is Zermelo-Fraenkel's (abbreviated: ZF), as described in Cohen [11, Chap. 2, Sect. 3]. The notations and basic facts about these are standard. Some important ones are given below.

(A) Let X be a nonempty set. By a *relation* over it, we mean any (nonempty) part $\mathscr{R}$ of $X \times X$; in this case, $(X, \mathscr{R})$ is called a *relational structure*. As usual, we may regard $\mathscr{R}$ as a mapping from X to 2^X (=the class of all subsets in X). Precisely, for each $x \in X$, denote

$X(x, \mathscr{R}) = \{y \in X; x\mathscr{R}y\}$ (the *section* of $\mathscr{R}$ through x);

then, the mapping in question is

$\mathscr{R}(x) = X(x, \mathscr{R}), x \in X.$

Call $\mathscr{R}$, *proper* when

$\mathscr{R}(x)$ is nonempty, for each $x \in X$;

note that, in such a case, $\mathscr{R}$ appears as a mapping between X and $(2)^X$ (=the class of all nonempty parts in X). This will be also referred to as: $(X, \mathscr{R})$ is a *proper* relational structure.

By a *sequence* in X, we mean any mapping $x : N \to X$; where $N := \{0, 1, \ldots\}$ is the set of *natural* numbers. 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 *divergent*: $i(n) \to \infty$ 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 all $Y \in (2)^X$.

Sometimes, when the ambient set X is endowed with countable type structures, it will suffice using—in our choice procedures—a weaker form of (AC), called: *Dependent Choice Principle* (in short: DC). Some preliminaries are needed. Let X be a nonempty set. For each natural number $k \geq 1$, call the map $F : N(k, >) \to X$, a *k-sequence*; if $k \geq 1$ is generic, we talk about a *finite* sequence. The following local result is available in the *strongly reduced* Zermelo-Fraenkel system (ZF-AC). Given $a \in X$, call the k-sequence $F : N(k, >) \to X$ (where $k \geq 2$), $(a, \mathscr{R})$*-iterative* provided it fulfills

$F(0) = a$ and $F(i) \mathscr{R} F(i + 1)$, for all $i \in N(k - 1, >)$.

Proposition 1 *Let the relational structure $(X, \mathscr{R})$ be proper. Then, for each $k \geq 2$, the following property holds:*

$(\pi(k))$ for each $a \in X$, there exists an $(a, \mathscr{R})$-iterative k-sequence.

Proof Clearly, $(\pi(2))$ is true; just take $b \in \mathscr{R}(a)$ and define $F : N(2, >) \to X$ as: $F(0) = a$, $F(1) = b$. Assume that $(\pi(k))$ is valid, for some $k \geq 2$; we claim that $(\pi(k + 1))$ is true as well. In fact, let $F : N(k, >) \to X$ be an $(a, \mathscr{R})$-iterative k-sequence, assured by hypothesis. As $\mathscr{R}$ is proper, $\mathscr{R}(F(k - 1))$ is nonempty; let u be some element of it. The map $G : N(k + 1, >) \to X$ introduced as

$G(i) = F(i), i \in N(k, >)$; $G(k) = u$

is an $(a, \mathscr{R})$-iterative $(k + 1)$-sequence; and then, we are done.

By definition, for each $k \geq 2$, the local property $(\pi(k))$ we just described is called the *k-finite Dependent Choice property* (in short: (DC-k)); and, if $k \geq 2$ is generic, the obtained global property $(\pi(k); k \geq 2)$ will be referred to as the *Finite Dependent Choice property* (in short: (DC-fin)). Now, it is natural to see what happens when k "tends to infinity" in the property $(\pi(k))$ of (DC-k). A formal result of this process is the so-called *Dependent Choice Principle* (in short: DC). Given $a \in X$, call the sequence $(x_n; n \geq 0)$ in X, $(a; \mathscr{R})$-*iterative* provided

$$x_0 = a; x_{n+1} \in \mathscr{R}(x_n), \forall n.$$

Proposition 2 *Let the relational structure $(X, \mathscr{R})$ be proper. Then, for each $a \in X$ there is an $(a, \mathscr{R})$-iterative sequence in X.*

At a first glance, (DC) seems to be obtainable in (ZF-AC) from such a "limit" process upon $((DC - k); k \geq 2)$. However, this is just an illusion; because—from a technical perspective—the limit process in question does not work in (ZF-AC); whence, (DC) is not obtainable from the axioms of our strongly reduced system. On the other hand, this principle—proposed, independently, by Bernays [3] and Tarski [41]—is deductible from (AC), but not conversely; cf. Wolk [50]. Moreover, by the developments in Blair [4], Goldblatt [22], Moskhovakis [35, Chap. 8], and Schechter [39, Chap. 6], the *reduced system* (ZF-AC+DC) is large enough so as to cover the "usual" mathematics; see also Moore [34, Appendix 2, Table 4].

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 for technical purposes.

Proposition 3 *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 our statement; we do not give details.

(C) A basic maximal statement in (ZF-AC+DC) to be used further is an asymptotic version of the 1976 Brezis-Browder ordering principle [6] (in short: BB). Let M be a nonempty set. Take a *quasi-order* $(\leq)$ (i.e.: reflexive and transitive relation) over it; and a function $x \mapsto \rho(x)$ from M to $R_+ \cup \{\infty\} = [0, \infty]$. Define the $(\leq, \rho)$-*maximal* property of some $z \in M$ as

$$z \leq w \in M \text{ implies } \rho(z) = \rho(w); \text{ i.e.: } \rho(M(z, \leq)) = \{\rho(z)\}.$$

The following "asymptotic functional" variant of (BB) (denoted as: (BB-af)) is our starting point.

Theorem 2 *Assume that the conditions below hold:*

(b01) $(M, \leq)$ is sequentially inductive (modulo ρ): each ascending sequence (x_n) with $\rho(x_n) \to 0$ has an upper bound (modulo $(\leq)$)

(b02) $(M, \leq)$ *is almost regular (modulo* ρ*):*
$\forall x \in M$, $\forall \varepsilon > 0$, $\exists y = y(x, \varepsilon) \geq x$, *such that* $\rho(y) \leq \varepsilon$
(b03) $\rho(.)$ *is* $(\leq)$*-decreasing (*$x \leq y$ *implies* $\rho(x) \geq \rho(y)$*).*

Then, for each $u \in M$ *there exists* $v \in M$ *with*

(21-a) $u \leq v$ *(or, equivalently:* $v \in M(u, \leq)$*)*
(21-b) $\rho(v) = 0$ *(hence* v *is* $(\leq, \rho)$*-maximal).*

Proof For each $\varepsilon > 0$, let $\mathscr{R}(\varepsilon)$ denote the relation (over M):

$x\mathscr{R}(\varepsilon)y$ iff $x \leq y$ and $\rho(y) \leq \varepsilon$;

note that, from the almost regular property, we have

$$M(c, \mathscr{R}(\varepsilon)) \neq \emptyset, \text{ for all } c \in M \text{ and all } \varepsilon > 0.$$

Now, let (ε_n) be a strictly descending sequence in $]0, \infty[$, with $\varepsilon_n \to 0$ as $n \to \infty$. From the above obtained fact, the sequence $(\mathscr{R}_n := \mathscr{R}(\varepsilon_n); n \geq 0)$ consists of proper relations. So, by (DDC), there must be a sequence (u_n) in M with $u_0 = u$ and

$$u_n \leq u_{n+1}, \rho(u_{n+1}) \leq \varepsilon_n, \text{ for all } n.$$

This sequence is therefore $(\leq)$-ascending and $\rho(u_n) \to 0$ as $n \to \infty$; wherefrom, by the sequentially inductive (modulo ρ) condition, (u_n) has an upper bound (modulo $(\leq)$):

$u_n \leq v$, for all $n \geq 0$ and some $v \in M$.

Combining with (ρ=decreasing) yields ($\rho(v) \leq \varepsilon_n$, $\forall n$); so that, $\rho(v) = 0$.

A basic particular case of these facts corresponds to the construction below. By a (generalized) *pseudometric* over M we shall mean any map $d : M \times M \to R_+ \cup \{\infty\}$, with the property

d is *reflexive*: $d(x, x) = 0, \forall x \in M$.

Call $z \in M$, $(\leq, d)$-*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*: $d(x, y) = 0 \Longrightarrow x = y$,

this maximal property becomes

$z \leq w \in M \Longrightarrow z = w$ (called: z is *strongly* $(\leq)$-*maximal*).

So, existence results involving such points are "metrical" versions of the Zorn-Bourbaki maximal principle (cf. Moore [34, Chap. 4, Sect. 4]). Returning to the general case, we stress that, in terms of the associated function (from M to $R_+ \cup \{\infty\}$)

$\rho_d(x) = \sup\{d(u, v); x \leq u \leq v\}, \ x \in M,$

this property may be characterized as: $\rho_d(z) = 0$. So, a basic source for determining such elements is the maximal principle (BB-af) above, applied to ρ_d. In this direction, we note that ρ_d is $(\leq)$-decreasing. Concerning the remaining properties, some conventions are needed. Given the (ascending) sequence (x_n), define the property

(x_n) is *strongly* $(\leq, d)$*-Cauchy*: $\rho_d(x_n) \to 0$ as $n \to \infty$;

and then, let us consider the associated condition

(si-d) $(M, \leq; d)$ is sequentially inductive (modulo d):
each ascending strongly $(\leq, d)$-Cauchy sequence has an upper bound;

it is nothing else than

$(M, \leq)$ is sequentially inductive (modulo ρ_d).

On the other hand, let us consider the condition

(reg-w) $(M, \leq; d)$ is weakly regular (modulo d):
$\forall x \in M, \forall \varepsilon > 0, \exists y = y(x, \varepsilon) \geq x$: $y \leq u \leq v \Longrightarrow d(u, v) \leq \varepsilon$;

clearly, it is nothing else than (see above)

$(M, \leq)$ is almost regular (modulo ρ_d).

Putting these together, it results (via (BB-af)) the following maximality statement (referred to as the "asymptotic pseudometric" variant of the Brezis-Browder ordering principle; in short: (BB-ap)).

Theorem 3 *Assume that $(M, \leq; d)$ is sequentially inductive (modulo d) and weakly regular (modulo d). Then, for each $u \in M$ there exists a $(\leq, d)$-maximal $v \in M$ with the property $u \leq v$.*

To discuss the former of these conditions, define the *d-Cauchy* property of an (ascending) sequence (x_n) in X as

for each $\varepsilon > 0$, there exists $n(\varepsilon)$, such that
$n(\varepsilon) \leq p \leq q \Longrightarrow d(x_p, x_q) \leq \varepsilon$.

Clearly, the following (generic) property holds

(for each (ascending) sequence in X):
strongly $(\leq, d)$-Cauchy $\Longrightarrow$ d-Cauchy.

Hence, the sequentially inductive (modulo d) condition we just imposed holds under

(si-c) $(M, \leq; d)$ is d-Cauchy sequentially inductive:
each ascending d-Cauchy sequence has an upper bound.

To discuss concrete circumstances under which this condition is to be assured, two main strategies may be followed.

Strategy 1 The former of these directions is based on sequential convergence methods. Precisely, let $(\xrightarrow{d})$ stand for the property:

$x_n \xrightarrow{d} x$ iff $d(x_n, x) \to 0$ as $n \to \infty$.

This will be also referred to as: x is the *d-limit* of (x_n); if such elements x exist, we shall say that (x_n) is *d-convergent*. Concerning the basic properties of this object, we have (by the reflexivity of d)

(conv-1) $(\xrightarrow{d})$ is *reflexive*:
$(x_n = u; n \geq 0)$ implies $x_n \xrightarrow{d} u$;

as well as (by definition)

(conv-2) $(\xrightarrow{d})$ is hereditary:
$x_n \xrightarrow{d} x$ implies $y_n \xrightarrow{d} x$, for each subsequence (y_n) of (x_n).

In other words, $(\xrightarrow{d})$ is a *convergence structure* on M, under Kasahara's terminology [32]. Note that, by the arbitrary character of d, no connection is to be deduced between the convergence property of an (ascending) sequence and the d-Cauchy of the same.

Having these precise, let us consider the couple of conditions

(o-com) d is $(\leq)$-complete:
each ascending d-Cauchy sequence is d-convergent
(self-c) $(\leq)$ is self-closed:
the d-limit of each ascending sequence is an upper bound of it.

It is now clear that

$$(\mathrm{o-com}) + (\mathrm{self-c}) \Longrightarrow (\mathrm{si-c}) \Longrightarrow (\mathrm{si-d}).$$

From (BB-ap), we then have the following Granas-Horvath version of (BB) (in short: (BB-GH)).

Theorem 4 *Let* $(M, \leq; d)$ *be weakly regular (modulo d); and one of conditions below be admitted;*

(i) $(M, \leq; d)$ *is d-Cauchy sequentially inductive*
(ii) d is $(\leq)$*-complete and* $(\leq)$ *is self-closed.*

Then, for each $u \in M$ *there exists a* $(\leq, d)$*-maximal element* $v \in M$ *with the property* $u \leq v$.

Now, evidently, (o-com) holds under

(com) d is complete (each d-Cauchy sequence is convergent);

and (self-c) is valid under the restrictive condition (due to Nachbin [36, Appendix])

(semi-c) $(\leq)$ is semi-closed ($M(x, \leq)$ is closed, $\forall x \in M$).

Note that, the corresponding variants of (BB-GH) include a lot of results in Granas and Horvath [23]; this, in particular, motivates the notational convention for our principle. Further aspects may be found in Altman [1]; see also Kang and Park [31].

Strategy 2 The second of these directions is being founded on (metric) regularity methods.

Let (x_n) be an ascending sequence in M. Remember that the *d-Cauchy* property for it is introduced as:

$\forall \varepsilon > 0, \exists n(\varepsilon)$, such that $n(\varepsilon) \leq p \leq q \Longrightarrow d(x_p, x_q) \leq \varepsilon$.

Also, call this sequence *d-asymptotic*, when

$d(x_n, x_{n+1}) \to 0$, as $n \to \infty$.

Let us now attach them the global conditions

(reg-C) each ascending sequence is d-Cauchy
(reg-A) each ascending sequence is d-asymptotic.

Proposition 4 *Under these conventions, we have*

(reg-A) $\Longrightarrow$ *(reg-C); hence, (reg-A)* $\Longleftrightarrow$ *(reg-C).*

Proof Suppose by absurd that (reg-A) holds; but (x_n) is not entitled with the d-Cauchy property:

$\exists \varepsilon > 0$, such that: $\forall n, \exists (p, q)$, with $n \leq p \leq q, d(x_p, x_q) > \varepsilon$.

As a consequence, we have that, for each n, the subset $\mathscr{R}_n \subseteq N \times N$ introduced as

$\mathscr{R}_n = \{(p, q) \in N \times N; n \leq p \leq q, d(x_p, x_q) > \varepsilon\}$

is a nonempty relation over N. Denote, for each n

$p(n) = \min(\mathrm{Dom}(\mathscr{R}_n)), q(n) = \min(\mathscr{R}_n(p(n)))$;

clearly, by the reflexivity of d, we must have

$$n \leq p(n) < q(n), \text{ for all } n.$$

Now, fix some rank $i(0)$. By the above working assumption, there may be determined [in a precise way—not related to choice procedures] a couple $(i(1), i(2)) = (p(i(0)), q(i(0)))$, with

$i(0) \leq i(1) < i(2), d(x_{i(1)}, x_{i(2)}) > \varepsilon$.

Further, given the rank $i(2)$, there may be determined [again in a precise way—not related to choice procedures] a couple $(i(3), i(4)) = (p(i(2)), q(i(2)))$, with

$$i(2) \le i(3) < i(4), d(x_{i(3)}, x_{i(4)}) > \varepsilon.$$

By induction, we therefore get a subsequence $(y_n = x_{i(n)})$ of (x_n), with

$$d(y_{2n+1}, y_{2n+2}) > \varepsilon, \text{ for all } n.$$

This, however, contradicts (reg-A); hence, the underlying working assumption cannot be true; and the claim follows.

By definition, either of these conditions (reg-C) and (reg-A) will be referred to as $(M, \le; d)$ is *regular* (modulo d).

Concerning the relationship with our weakly regular concept, one has

Proposition 5 *The generic inclusion is valid, over the class of pseudometric quasi-ordered structures* $(M, \le; d)$*:*

(DC) and regular (modulo d) $\Longrightarrow$ *weakly regular (modulo d), in (ZF-AC);*
or, equivalently,
regular (modulo d) $\Longrightarrow$ *weakly regular (modulo d), in (ZF-AC+DC).*

Proof Suppose—under (DC)—that $(M, \le; d)$ is regular (modulo d); i.e. (see above) one of the (equivalent) global conditions (reg-C) and (reg-A) is holding. We have to establish that $(M, \le; d)$ is weakly regular (modulo d); i.e.,

(reg-w) $\forall x \in M$, $\forall \varepsilon > 0$, $\exists y = y(x, \varepsilon) \ge x$, such that
$y \le u \le v \Longrightarrow d(u, v) \le \varepsilon$.

Suppose that this property fails; i.e., for some $c \in M$, $\varepsilon > 0$,

for each $y \ge c$, there exist $u, v \in X$ such that $y \le u \le v$, $d(u, v) > \varepsilon$.

Let $\mathrm{gr}(\le; c) := \{(a, b) \in X \times X; c \le a \le b\}$ stand for the c-section of the *graph* attached to $(\le)$. We introduce a relation $\mathscr{R}$ over $\mathrm{gr}(\le; c)$, according to

$(x, y)\mathscr{R}(u, v)$ iff $y \le u$, $d(u, v) > \varepsilon$.

Clearly, $(\mathrm{gr}(\le; c), \mathscr{R})$ is a proper relational structure. Hence, from the Dependent Choice Principle (DC), it follows that, given $(x_0, y_0) \in \mathrm{gr}(\le; c)$, there must be a sequence $((x_n, y_n); n \ge 0)$ in $\mathrm{gr}(\le; c)$, with

$$(x_n, y_n)\mathscr{R}(x_{n+1}, y_{n+1}), \ \forall n;$$

or, equivalently (by definition)

$$y_n \le x_{n+1}, d(x_{n+1}, y_{n+1}) > \varepsilon, \ \forall n.$$

The sequence $(z_n; n \geq 0)$ in M introduced as $(z_{2n} = x_n, z_{2n+1} = y_n; n \geq 0)$ is ascending and

$$d(z_{2n+2}, z_{2n+3}) > \varepsilon, \ \forall n;$$

hence, (z_n) is not d-asymptotic. This contradicts the regularity (modulo d) of so that, the working hypothesis above cannot hold. The proof is complete.

Further, note that the condition

(si) $(M, \leq)$ is *sequentially inductive*:
each ascending sequence has an upper bound (modulo $(\leq)$)

is a sufficient one for

(si-c) $(M, \leq; d)$ is d-Cauchy sequentially inductive:
each ascending d-Cauchy sequence has an upper bound.

The reciprocal holds as well, in a regular setting; i.e.,

$(M, \leq; d)$ is regular (modulo d) and $(M, \leq; d)$ is d-Cauchy sequentially inductive imply $(M, \leq)$ is sequentially inductive.

Combining with our preceding developments, the following Conserva-Rizzo version of (BB) (denoted as: (BB-CR)) is available over (ZF-AC+DC):

Theorem 5 *Let the pseudometric quasi-ordered structure $(M, \leq; d)$ be such that $(M, \leq; d)$ is regular (modulo d), and one of the conditions below is holding*

(j) $(M, \leq; d)$ is d-Cauchy sequentially inductive
(jj) $(M, \leq)$ is sequentially inductive
(jjj) d is $(\leq)$-complete and $(\leq)$ is self-closed.

Then, for each $u \in M$ there exists a $(\leq, d)$-maximal $v \in M$ with $u \leq v$.

Proof (Sketch) By the former condition, $(M, \leq; d)$ is weakly regular (modulo d). Hence, (BB-GH) applies to these data; wherefrom, all is clear.

Finally, note that the sequential inductive part of our maximal principle above is nothing else than the related statement in Conserva and Rizzo [12]; this, among others, motivates our convention. Further aspects of the problem were discussed in the paper by Turinici [44].

(C) A basic maximal statement in (ZF-AC+DC) deductible via these developments is the 1976 Brezis-Browder ordering principle [6] (in short: BB). Let M be a nonempty set. Take a *quasi-order* $(\leq)$ (i.e.: reflexive and transitive relation) over it; and a function $x \mapsto \psi(x)$ from M to $R_+ := [0, \infty[$. Call $z \in M$, *$(\leq, \psi)$-maximal* when:

$z \leq w \in M$ implies $\psi(z) = \psi(w)$.

Theorem 6 *Suppose that*

(b04) $(M, \leq)$ *is sequentially inductive:*
each ascending sequence has an upper bound (modulo $(\leq)$*)*
(b05) 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 includes (as we shall see) Ekeland's Variational Principle [16] (in short: EVP); and found some useful applications to convex and non-convex analysis. So, it was the subject of many extensions; see, for instance, Hyers et al. [26, Chap. 5]. These are interesting from a technical perspective; but, in all concrete situations when a variational principle of this type [(VP), say] is to be applied, a substitution by the Brezis-Browder's is always possible. This raises the question as to what extent are the logical inclusions (VP) $\Longrightarrow$ (BB) $\Longrightarrow$ (EVP) effective. As a result of the developments below, the former inclusion is sometimes reversible; i.e., many statements (VP) including (BB) are but logical equivalents of (BB). Concerning the latter inclusion, we show that (BB) is deductible from (DC). So, to close the circle between these, it will suffice proving that (EVP) includes (DC). An early result of this type was provided in 1987 by Brunner [9]; a refinement of it was provided in the 2011 paper in Turinici [48]. Summing up, (BB) and (EVP) are both equivalent with (DC); and, as such, mutually equivalent. This tells us that all variational statements (VP) with

(DC) $\Longrightarrow$ (VP) $\Longrightarrow$ (BB) and/or (DC) $\Longrightarrow$ (VP) $\Longrightarrow$ (EVP)

are equivalent to each other.

Proof (BB) For technical reasons, we shall provide an argument for (BB) above being reducible to either (BB-GH) or (BB-CR).

Define the function $\beta : M \to R_+$ as:

$\beta(v) := \inf[\psi(M(v, \leq))]$, $v \in M$;

clearly, β is increasing and

$$\psi(v) \geq \beta(v), \text{ for all } v \in M.$$

Further, the decreasing property of ψ gives a characterization like

$$v \text{ is } (\leq, \psi)\text{−maximal iff } \psi(v) = \beta(v).$$

Now, assume by contradiction that the conclusion in this statement is false; i.e. [in combination with the above], there must be some $u \in M$ such that:

for each $v \in M_u := M(u, \leq)$, one has $\psi(v) > \beta(v)$.

We intend to show that a contradiction is to be reached with respect to either (BB-GH) or (BB-CR). To this end, let us introduce the mapping (on M)

$$d(x,y) = |\psi(x) - \psi(y)|, x, y \in M.$$

Clearly, d is reflexive; hence, it is a pseudometric on M. In addition,

d is *triangular*: $d(x,z) \leq d(x,y) + d(y,z)$, $\forall x, y, z \in M$
d is *symmetric*: $d(x,y) = d(y,x)$, $\forall x, y \in M$;

hence, d is a *semimetric* on M.

I) Let (x_n) be an ascending sequence in M_u. By the decreasing property, $(\psi(u_n))$ is descending in R_+; hence, a Cauchy sequence. This tells us that (x_n) is d-Cauchy; wherefrom (by the arbitrariness of this object) $(M_u, \leq; d)$ is regular (modulo d). On the other hand, by the sequential inductivity, (x_n) is bounded from above in M:

there exists $v \in M$ such that $x_n \leq v$, $\forall n$ (hence, $v \in M_u$);

and this (along with the arbitrariness of our sequence) tells us that $(M_u, \leq)$ is sequentially inductive. Taking (BB-CR) into account, we get a $(\leq, \psi)$-maximal element $v \in M_u$; in contradiction with our working hypothesis.

II) Let $x \in M_u$ be arbitrary fixed; hence, $\psi(x) > \beta(x)$. Further, let $\varepsilon > 0$ be arbitrary fixed. By definition, there exists some point $y \in M(x, \leq)$ with

$\beta(x) \leq \psi(y) < \beta(x) + \varepsilon$ (hence, $0 \leq \psi(y) - \beta(x) < \varepsilon$).

This means that $y \in M_u$; and, moreover (as ψ=decreasing)

$$x \leq y \leq u \leq v \Longrightarrow \psi(y) \geq \psi(u) \geq \psi(v) \geq \beta(v) \geq \beta(u) \geq \beta(y) \geq \beta(x);$$

wherefrom (by our conventions)

$$y \leq u \leq v \Longrightarrow d(u,v) < \varepsilon;$$

which tells us that $(M_u, \leq)$ is weakly regular (modulo d). Combining with the sequential inductivity of $(M_u, \leq)$ we just established, it results that (BB-GH) applies to our data; wherefrom it must be at least one $(\leq, \psi)$-maximal element $v \in M_u$; in contradiction with the same working hypothesis.

Note that, by the same procedure, one gets a slight extension of this result, due to Cârjă et al. [10, Chap. 2, Sect. 2.1]. Further metrical versions of (BB) may be found in Turinici [46].

Monotone EVP

A basic application of these facts is to "monotone" variational principles.

(A) Let X be a nonempty set; and $(\leq)$ be some quasi-order on it. By a *generalized metric* over X we mean, as in Luxemburg [33] and Jung [29], any map $(x,y) \mapsto d(x,y)$ from $X \times X$ to $R_+ \cup \{\infty\} = [0,\infty]$, endowed with all properties of a standard metric (over this extended half-axis):

d is *triangular*: $d(x,z) \leq d(x,y) + d(y,z)$, $\forall x,y,z \in X$
d is reflexive-sufficient: $d(x,y) = 0$ iff $x = y$
d is *symmetric*: $d(x,y) = d(y,x)$, $\forall x,y \in X$.

Suppose that we fixed such an object in the sequel. Call the subset Z of X, $(\leq)$-*closed* (modulo d) when

the limit of each ascending (modulo $(\leq)$) sequence in Z belongs to Z.

Clearly, any closed (modulo d) part of X is $(\leq)$-closed (modulo d) too. The converse is not in general true: just take $X = R$ (endowed with the usual order/metric) and $Z =]0,1]$. Further, call the quasi-order $(\leq)$, *self-closed* (modulo d) provided

$X(x,\leq)$ is $(\leq)$-closed (modulo d), for each $x \in X$;
or, equivalently:
the limit of each ascending sequence is an upper bound of it (modulo $(\leq)$).

For example, this is the case when (cf. Nachbin [36, Appendix])

$(\leq)$ is *semi-closed* (modulo d):
$X(x,\leq)$ is closed (modulo d), for each $x \in X$.

Finally, call the ambient metric d, $(\leq)$-*complete* provided

each ascending (modulo $(\leq)$) d-Cauchy sequence (in X) is d-convergent.

As before, if d is complete, then it is $(\leq)$-complete too. The reciprocal is not in general true; take $X =]0,1]$ endowed with the standard order and metric.

We are now in position to state the announced result. Let the quasi-order $(\leq)$ and the generalized metric $d : X \times X \to R_+ \cup \{\infty\}$ over X be such that

(c01) $(\leq)$ is self-closed (modulo d)
(c02) d is $(\leq)$-complete (over X).

Further, take a function $\varphi : X \to R \cup \{\infty\}$, fulfilling

(c03) φ is inf-proper ($\mathrm{Dom}(\varphi) \neq \emptyset$ and $\inf[\varphi(X)] > -\infty$)
(c04) φ is $(\leq, d)$-lsc: $\liminf_n \varphi(x_n) \geq \varphi(x)$,
whenever (x_n) is ascending (modulo $(\leq)$) and $x_n \xrightarrow{d} x$;
or, equivalently:
$[\varphi \leq t] := \{x \in X; \varphi(x) \leq t\}$ is $(\leq)$-closed (modulo d), for each $t \in R$.

Theorem 7 *Let the prescribed conditions be admitted. Then, for each $u \in \mathrm{Dom}(\varphi)$ there exists $v \in \mathrm{Dom}(\varphi)$, with*

*(**31-a**) $u \leq v,\ d(u,v) \leq \varphi(u) - \varphi(v)$ (hence $\varphi(u) \geq \varphi(v)$)*
*(**31-b**) $x \in X$, $v \leq x$ and $d(v,x) \leq \varphi(v) - \varphi(x)$ imply $v = x$.*

Proof Let $(\preceq)$ stand for the relation (over X)

$$(x, y \in X):\ x \preceq y \text{ iff } x \leq y,\ d(x,y) + \varphi(y) \leq \varphi(x).$$

It is not hard to see that $(\preceq)$ acts as an *order* (antisymmetric quasi-order) on $\mathrm{Dom}(\varphi)$. Further, denote

$X_u := X(u, \preceq)$ (i.e.: $X_u = \{x \in X; u \preceq x\}$).

Clearly, $\emptyset \neq X_u \subseteq \mathrm{Dom}(\varphi)$; moreover (by the admitted conditions)

$$X_u \text{ is}(\preceq)\text{−closed (modulo } d\text{); hence } d \text{ is } (\preceq)\text{−complete on } X_u.$$

We claim that conditions of Brezis-Browder ordering principle (BB) are fulfilled on the ordered structure $(X_u, \preceq)$. In fact, let (x_n) be an ascending (modulo $(\preceq)$) sequence in X_u:

$x_n \leq x_m$ and $d(x_n, x_m) \leq \varphi(x_n) - \varphi(x_m)$, if $n \leq m$.

The sequence $(\varphi(x_n))$ is descending and bounded from below; hence a Cauchy one. This, along with the working relation, shows that (x_n) is an ascending (modulo $(\preceq)$) d-Cauchy sequence; wherefrom $(X_u, \preceq)$ is regular (modulo d). Moreover, the obtained properties give us (by the properties of X_u) some $y \in X_u$ with $x_n \xrightarrow{d} y$. Combining with our working hypothesis, one derives

$$x_n \leq y, d(x_n, y) \leq \varphi(x_n) - \varphi(y),\ \text{(i.e. : } x_n \preceq y\text{)},\ \text{for all } n.$$

In other words, $y \in X_u$ is an upper bound (modulo $(\preceq)$) of (x_n); and this shows that $(X_u, \preceq)$ is sequentially inductive. On the other hand, φ (restricted to X_u) is strictly decreasing:

$x, y \in X_u, x \preceq y, x \neq y$ implies $\varphi(x) > \varphi(y)$;

or, equivalently,

$x, y \in X_u, x \preceq y$, and $\varphi(x) = \varphi(y)$ imply $x = y$.

Hence, (BB) is indeed applicable to our data. As a consequence of this, it then follows that, for the starting $u \in X_u$, there exists some $v \in X_u$ with

(31-c) $u \preceq v$ (whence: $u \leq v, d(u,v) \leq \varphi(u) - \varphi(v)$)
(31-d) v is $(\preceq, \varphi)$-maximal in X_u
(i.e.: $x \in X_u,\ v \leq x,\ d(v,x) \leq \varphi(v) - \varphi(x) \Longrightarrow \varphi(v) = \varphi(x)$).

The former of these is just the first conclusion of the statement. And the latter one gives at once the second conclusion of the same. In fact, assume by contradiction that there would be some $x \in X \setminus \{v\}$, with

$v \le x$ and $d(v,x) \le \varphi(v) - \varphi(x)$; hence, $v \preceq x$.

As $u \preceq v$, we must have $u \preceq x$; i.e.: $x \in X_u$; so, combining with the above

$x \in X_u \setminus \{v\}$, $v \le x$ and $d(v,x) \le \varphi(v) - \varphi(x)$.

Note that, a direct consequence of this relation is

$\varphi(v) > \varphi(x)$ (since $d(v,x) > 0$).

On the other hand, the same premises give

$\varphi(v) = \varphi(x)$ (as v is $(\preceq, d$-maximal in X_u).

The contradiction at which we arrived shows that our working assumption cannot be accepted; and this establishes our claim. The proof is thereby complete.

A basic particular case of this variational result (established—with a different proof—in Turinici [45]) corresponds to the choice

$(\le) = X \times X$ (=the *trivial quasi-order* on X).

The regularity condition (c04) may then be written as

φ is d-lsc: $\liminf_n \varphi(x_n) \ge \varphi(x)$, whenever $x_n \xrightarrow{d} x$;
or, equivalently:
$[\varphi \le t]$ is closed (modulo d), for each $t \in R$;

and Theorem 7 is nothing but Ekeland's variational principle (EVP). On the other hand, the same requirement holds under

$(\le)$ is self-closed (modulo d) and φ is decreasing (modulo $(\le)$);

[The proof is immediate; so, we do not give details]. For this reason, Theorem 7 will be called the *monotone* version of (EVP) (in short: (EVPm)). Note that, by the remarks above, it may be also derived from either of the maximal principles (BB-GH) and (BB-CR); we do not give further details.

(B) From the developments above, we have the implications:

(DC) $\Longrightarrow$ (BB-af) $\Longrightarrow$ (BB-ap) $\Longrightarrow$ (BB-GH) $\Longrightarrow$ (BB)
(DC) $\Longrightarrow$ (BB-CR) $\Longrightarrow$ (BB) $\Longrightarrow$ (EVPm).

So, it is natural asking whether these may be reversed. Clearly, the natural setting for solving this problem is the *strongly reduced* Zermelo-Fraenkel system (ZF-AC). To state a basic result in this direction, some preliminaries are needed.

Let $(X, \le)$ be a partially ordered structure. We say that $(\le)$ has the *inf-lattice* property, provided:

$x \wedge y := \inf(x,y)$ exists, for all $x, y \in X$.

Further, call $z \in X$, $(\leq)$-*maximal* if $X(z, \leq) = \{z\}$; the class of all these points will be denoted as $\max(X, \leq)$. In this case, $(\leq)$ is termed a *Zorn order* when

$\max(X, \leq)$ is nonempty and *cofinal* in X:
for each $u \in X$ there exists a $(\leq)$-maximal $v \in X$ with $u \leq v$.

Further aspects are to be described in a metrical setting. Let $d : X \times X \to R_+$ be a metric over X; and $\varphi : X \to R_+$ be some function. Then, the natural choice for $(\leq)$ above is

$x \leq_{(d,\varphi)} y$ iff $d(x, y) \leq \varphi(x) - \varphi(y)$;

referred to as the Brøndsted order [8] attached to (d, φ). Denote

$X(x, \rho) = \{u \in X; d(x, u) < \rho\}, x \in X, \rho > 0$
(the *open sphere* with center x and radius ρ).

Call the ambient metric space (X, d), *discrete* when

for each $x \in X$ there exists $\rho = \rho(x) > 0$ such that $X(x, \rho) = \{x\}$.

Note that, under such a hypothesis, any function $\psi : X \to R$ is continuous over X. However, the stronger property of this object

$|\psi(x) - \psi(y)| \leq Ld(x, y), x, y \in X$ (for some $L > 0$)

cannot be assured, in general. This will be referred to as: ψ is (L, d)-*Lipschitz*; when $L = 1$ we then say that ψ is *d-nonexpansive*.

Now, let X be a nonempty set, $d : X \times X \to R_+$ be a metric over X and $\varphi : X \to R_+$ be a function. Remember that d is called $(\leq_{(d,\varphi)})$-*complete*, provided

each ascending (modulo $(\leq_{(d,\varphi)})$) d-Cauchy sequence is d-convergent.

An apparently stronger version of this is the following: call the ambient metric d, $(\leq_{(d,\varphi)})$-*strongly-complete* provided

each ascending (modulo $(\leq_{(d,\varphi)})$) sequence is d-convergent.

In fact, these conventions are equivalent to each other, as results from

Proposition 6 *Suppose (under these conventions) that d is $(\leq_{(d,\varphi)})$-complete. Then, d is $(\leq_{(d,\varphi)})$-strongly-complete.*

Proof Let (x_n) be an ascending (modulo $(\leq_{(d,\varphi)})$) sequence in X:

$d(x_n, x_m) \leq \varphi(x_n) - \varphi(x_m)$, if $n \leq m$.

The sequence $(\varphi(x_n))$ is descending and bounded from below; hence a Cauchy one. This, along with the working relation, shows that (x_n) is an ascending (modulo $(\leq_{(d,\varphi)})$) d-Cauchy sequence; and then, by completeness, all is clear.

We may now pass to the basic part of our developments. The statement below is a particular case of our monotone variational principle (EVPm).

Theorem 8 *Let the nonempty set X, the metric d over X and the function $\varphi : X \to R_+$ be such that*

(c05) (X, d) *is discrete and bounded*
(c06) $(\leq_{(d,\varphi)})$ *has the inf-lattice property*
(c07) d *is* $(\leq_{(d,\varphi)})$*-complete*
(or, equivalently: d *is* $(\leq_{(d,\varphi)})$*-strongly-complete)*
(c08) φ *is* d*-nonexpansive and* $\varphi(X)$ *is countable.*

Then, $(\leq_{(d,\varphi)})$ *is a Zorn order.*

Proof (Sketch) As φ is in particular d-continuous, the (partial) order $(\leq_{(d,\varphi)})$ is semi-closed; hence, all the more, self-closed. Moreover, again by the continuous property, φ is necessarily d-lsc; hence, $(\leq_{(d,\varphi)}, d)$-lsc. Summing up, (EVPm) is indeed applicable to our context; and from this, the conclusion follows.

We shall refer to this statement as: the discrete Lipschitz countable version of (EVPm) (in short: (EVPm-dLc)). By the above developments, (EVPm) $\Longrightarrow$ (EVPm-dLc). The remarkable fact to be added is that this last principle yields (DC); so, it completes the circle between all these.

Proposition 7 *We have*

(EVPm-dLc) $\Longrightarrow$ *(DC), in (ZF-AC).*

So, the maximal/variational principles (BB-af), (BB-ap), (BB-GH), (BB-CR), (BB) and (EVPm) are all equivalent with (DC); hence, mutually equivalent.

Proof Let M be a nonempty set; and $\mathscr{R}$ stand for some proper relation over it. Fix in the following $a \in M$, $b \in M(a, \mathscr{R})$. For each $p \geq 2$ in N (= the set of natural numbers), let $N(p, >) := \{0, \ldots, p-1\}$ stand for the *initial segment* determined by p. Denote, for simplicity

X_p = the class of all finite sequences $x : N(p, >) \to M$ with:
$x(0) = a$, $x(1) = b$, and $x(n)\mathscr{R}x(n+1)$ for $0 \leq n \leq p-2$.

In this case, $N(p, >)$ is just $\mathrm{Dom}(x)$ (the *domain* of x); and $p = \mathrm{card}(N(p, >))$ will be referred to as the *order* of x [denoted as $\omega(x)$]. Concerning the effectiveness of this construction, we have (by the Finite Dependent Choice property)

$$X_p \text{ is nonempty, for each } p \geq 2.$$

As a consequence of this, $X = \cup\{X_n; n \geq 2\}$ (is well defined and) nonempty. Let $(\preceq)$ stand for the partial order (on X)

$x \preceq y$ iff $\mathrm{Dom}(x) \subseteq \mathrm{Dom}(y)$ and $x = y|_{\mathrm{Dom}(x)}$;

and $(\prec)$ denote its associated strict order:

$x \prec y$ iff $x \preceq y$ and $x \neq y$.

The following auxiliary fact is to be noted.

Lemma 1 *Under these conventions,*

i) *The mapping $x \mapsto \omega(x)$ is increasing:*

$x \preceq y$ *implies* $\omega(x) \le \omega(y)$.

ii) *The mapping $x \mapsto \omega(x)$ is strictly increasing:*

$x \prec y$ *implies* $\omega(x) < \omega(y)$.

Proof i): By definition, $x \preceq y$ implies $\mathrm{Dom}(x) \subseteq \mathrm{Dom}(y)$; which means $\omega(x) \le \omega(y)$.

ii): Suppose that $x \prec y$ but $\omega(x) = \omega(y)$; i.e.: $\mathrm{Dom}(x) = \mathrm{Dom}(y)$. As $x \preceq y$, we must have

$$x = y|_{\mathrm{Dom}(x)} = y|_{\mathrm{Dom}(y)} = y;$$

in contradiction with $x \neq y$. Hence, $\omega(x) < \omega(y)$; and conclusion follows.

The following auxiliary fact gives a natural way of obtaining the $(a, \mathscr{R})$-iterative sequences in M we are looking for.

Lemma 2 *Suppose that $(X, \preceq)$ admits strictly ascending sequences. Then, necessarily, M admits $(a, \mathscr{R})$-iterative sequences.*

Proof Suppose that there exists a sequence $(z_n; n \ge 0)$ in X, endowed with the strict ascending property:

$i < j$ implies $z_i \prec z_j$; hence, $\omega(z_i) < \omega(z_j)$.

The natural sequence $(p_n := \omega(z_n); n \ge 0)$ is therefore strictly ascending, with

$$p_0 \ge 2;\ \text{hence}, p_n \ge 2 + n,\ n \ge 0.$$

This tells us that the sequence $(c_n = z_n(n); n \ge 0)$ is well defined in M; and, by the choice of $(z_n; n \ge 0)$,

$$c_0 = a, c_n \mathscr{R} c_{n+1},\ \text{for all } n;$$

whence, $(c_n; n \ge 0)$ is $(a, \mathscr{R})$-iterative.

(C) As a consequence of this, it will suffice proving that $(X, \preceq)$ has strictly ascending sequences, to end our argument. To get such a conclusion, we need some conventions and auxiliary facts taken from Turinici [48].

(C1) Let $x, y \in X$ be arbitrary fixed. Denote

$K(x, y) := \{n \in \mathrm{Dom}(x) \cap \mathrm{Dom}(y); x(n) \neq y(n)\}$.

If x and y are *comparable* (i.e.: either $x \preceq y$ or $y \preceq x$; written as: $x <> y$), then $K(x, y) = \emptyset$. Conversely, if $K(x, y) = \emptyset$, then $x \preceq y$ if $\mathrm{Dom}(x) \subseteq \mathrm{Dom}(y)$ and $y \preceq x$ if $\mathrm{Dom}(y) \subseteq \mathrm{Dom}(x)$; hence $x <> y$. Summing up,

$$(x, y \in X): \quad x <> y \text{ if and only if } K(x, y) = \emptyset.$$

The negation of this property means: x and y are *not comparable* (denoted as: $x||y$). By the characterization above, it is equivalent with $K(x, y) \neq \emptyset$. Note that, in such a case, $k(x, y) := \min(K(x, y))$ is well defined as an element of $N(2, \leq)$; and $N(k(x, y), >)$ is the largest initial interval of $\mathrm{Dom}(x) \cap \mathrm{Dom}(y)$ where x and y are identical.

Lemma 3 *The partial order* $(\preceq)$ *has the inf-lattice property. Moreover,*

$$2 \leq \omega(x \wedge y) \leq \min\{\omega(x) - 1, \omega(y) - 1\}, \text{ whenever } x||y.$$

Proof **i)** Let $x, y \in X$ be arbitrary fixed. The case $x <> y$ is clear; so, without loss, one may assume that $x||y$. Note that, by the remark above, $K(x, y) \neq \emptyset$ and $k := k(x, y)$ exists as an element of $N(2, \leq)$. Let the finite sequence $z \in X_k$ be introduced as $z = x|_{N(k,>)} = y|_{N(k,>)}$. For the moment $z \preceq x$ and $z \preceq y$. Suppose that $w \in X_h$ (where $h \geq 2$) fulfills the same properties. Then, the restrictions of x and y to $N(h, >)$ are identical; wherefrom (see above) $h \leq k$ and $w \preceq z$.

ii) As $x||y$, we must have $\omega(x), \omega(y) \geq 3$ and $\omega(x \wedge y) = k(x, y) \geq 2$. This and $k(x, y) \in \mathrm{Dom}(x) \cap \mathrm{Dom}(y)$ give the desired relation.

(C2) Our next objective is to introduce a metrical structure as well as an associated objective function over X, which should have "many" properties required by (EVPm-dLc). Put

$$\varphi(x) = 3^{-\omega(x)}, x \in X;$$

and note that $\varphi(X) = \{3^{-n}; n \geq 2\}$ (hence, φ has countable many strictly positive values). Then, define

$$d(x, y) = |\varphi(x) - \varphi(y)|, \text{ if } x <> y; \ d(x, y) = \varphi(x \wedge y), \text{ when } x||y.$$

Lemma 4 *The mapping* $(x, y) \mapsto d(x, y)$ *is a metric on* X.

Proof Clearly, d is reflexive and symmetric $[d(x, y) = d(y, x), x, y \in X]$. On the other hand, d is sufficient. In fact, assume $d(x, y) = 0$. By a previous evaluation of $\varphi(X)$, it results that x and y are comparable and $\omega(x) = \omega(y)$; wherefrom, $x = y$. Finally, let us verify the triangular property:

$$d(x, z) \leq d(x, y) + d(y, z), \text{ for all } x, y, z \in X.$$

Two alternatives are open before us.

a) The points x and z are comparable ($x <> z$). We start from the obvious relation

$$|\varphi(s) - \varphi(t)| \leq \max\{\varphi(s), \varphi(t)\} \leq \varphi(s \wedge t), \ s, t \in X.$$

Combining with

$$d(x, z) = |\varphi(x) - \varphi(z)| \leq |\varphi(x) - \varphi(y)| + |\varphi(y) - \varphi(z)|$$

yields the desired fact, for all possible cases concerning (x, y) and (y, z).

b) The points x and z are not comparable ($x||z$). Four sub-cases appear:

Sub-case b1): Suppose that $x <> y$, $y <> z$. The alternatives $[x \preceq y, y \preceq z]$ and $[y \preceq x, z \preceq y]$ give $x <> z$; contradiction. So, it remains to discuss the alternatives:

b11) $x \preceq y$, $z \preceq y$. Then, x and z are the restrictions of y to $\mathrm{Dom}(x)$ and $\mathrm{Dom}(z)$, respectively; wherefrom $x <> z$, contradiction.

b12) $y \preceq x$, $y \preceq z$. We start from a direct consequence of our previous evaluation

$$3\max\{\varphi(s), \varphi(t)\} \le \varphi(s \wedge t) \le 3^{-2}, \quad s, t \in X, s||t.$$

The relation to be checked becomes

$$\varphi(x \wedge z) \le 2\varphi(y) - \varphi(x) - \varphi(z).$$

By the imposed conditions, $y \preceq x \wedge z$; wherefrom $\varphi(y) \ge \varphi(x \wedge z)$. A sufficient condition for the desired relation to be true is

$$\varphi(x \wedge z) \le 2\varphi(x \wedge z) - \varphi(x) - \varphi(z);\ \text{i.e.} : \varphi(x) + \varphi(z) \le \varphi(x \wedge z);$$

evident, by the precise consequence.

Sub-case b2): Suppose that $x||y$, $y <> z$. Two logical possibilities occur:

b21) $x||y$, $y \preceq z$. We have to establish that: $\varphi(x \wedge z) \le \varphi(x \wedge y) + \varphi(y) - \varphi(z)$. But, evidently, $x \wedge z \succeq x \wedge y$; wherefrom $\varphi(x \wedge z) \le \varphi(x \wedge y)$; and then, all is clear.

b22) $x||y$, $z \preceq y$ (or, equivalently: $z \preceq y$, $y||x$). The desired relation becomes:

$$\varphi(x \wedge z) \le \varphi(x \wedge y) + \varphi(z) - \varphi(y).$$

For the moment, $x \wedge z \preceq x \wedge y$. If $x \wedge z \prec x \wedge y$, we must get

$$q := \omega(x \wedge z) < \omega(x \wedge y);$$

so (by definition) $x(q) = y(q)$. As $z = y|_{\mathrm{Dom}(z)}$ and $q \in \mathrm{Dom}(x) \cap \mathrm{Dom}(z)$, this yields $y(q) = z(q)$; hence $x(q) = z(q)$, contradiction. Consequently, $x \wedge z = x \wedge y$; and conclusion follows.

Sub-case b3): $x <> y$, $y||z$. As before, two logical possibilities occur:

b31) $y \preceq x$, $y||z$ (or, equivalently: $z||y$, $y \preceq x$). This is just alternative **b21)**, with (z, y, x) in place of (x, y, z).

b32) $x \preceq y$, $y||z$. This is just alternative **b22)**, with (x, y, z) in place of (z, y, x).

Sub-case b4): $x||y$, $y||z$. We have to establish that: $\varphi(x \wedge z) \le \varphi(x \wedge y) + \varphi(y \wedge z)$. As before, the alternative

$$\omega(x \wedge z) \ge \omega(x \wedge y) \text{ or } \omega(x \wedge z) \ge \omega(y \wedge z)$$

gives the desired fact. On the other hand, the alternative

$$q := \omega(x \wedge z) < \min\{\omega(x \wedge y), \omega(y \wedge z)\}$$

yields $x(q) = y(q)$, $y(q) = z(q)$; hence $x(q) = z(q)$, contradiction.

Having discussed all possible cases, the conclusion follows.

(C3) Note that, by a previous remark involving $\varphi(X)$, one has $\mathrm{diam}(X) \leq 3^{-2}$. Further properties of the triplet $(X, d; \varphi)$ are contained in

Lemma 5 *Under the notations above, one has (for each $m \geq 2$)*

$$x, y \in X,\ \omega(x) \leq m,\ d(x,y) < 2 \cdot 3^{-m-1} \Longrightarrow x = y;$$

so that, the metric space (X, d) is discrete.

Proof Assume that $x \neq y$. We show that this cannot be in agreement with our hypothesis. Two cases are open before us:

i) Let x and y be comparable: either $x \prec y$ or $y \prec x$. If $x \prec y$, we have $\omega(x) + 1 \leq \omega(y)$; and then $(1/3)\varphi(x) \geq \varphi(y)$; hence (by definition)

$$d(x,y) = \varphi(x) - \varphi(y) \geq (2/3)\varphi(x) \geq 2 \cdot 3^{-m-1},$$

contradiction. If $y \prec x$, then (by the same way as before)

$$d(x,y) \geq (2/3)\varphi(y) \geq (2/3)\varphi(x) \geq 2 \cdot 3^{-m-1},$$

again a contradiction.

ii) Suppose that x and y are not comparable. Then (by definition)

$$d(x,y) = \varphi(x \wedge y) \geq \varphi(x) \geq 3 \cdot 3^{-m-1},$$

contrary to the hypothesis.

Lemma 6 *Under the same notations above,*

$$|\varphi(x) - \varphi(y)| \leq d(x,y),\ \forall x, y \in X;$$

so, the objective function φ is d-nonexpansive.

Proof If x and y are comparable, then $d(x,y) = |\varphi(x) - \varphi(y)|$; and we are done. If x and y are not comparable then, without loss, one may assume $\omega(x) \leq \omega(y)$; hence $\varphi(x) \geq \varphi(y)$. As $x \wedge y \preceq x$, we have

$$d(x,y) = \varphi(x \wedge y) \geq \varphi(x) \geq \varphi(x) - \varphi(y) = |\varphi(x) - \varphi(y)|;$$

and conclusion follows.

(C4) Given the couple (d, φ) as before, let $(\leq_{(d,\varphi)})$ stand for the Brøndsted order on X; also denoted as $(\leq)$, for simplicity. It is natural to ask which is the relationship between it and the initial order $(\preceq)$ on X.

Lemma 7 *We necessarily have (under these conventions)*

$$x \preceq y \text{ if and only if } x \leq y.$$

That is: these partial orders coincide over X.

Proof Clearly, $x \preceq y$ gives $\omega(x) \leq \omega(y)$; wherefrom $d(x,y) = \varphi(x) - \varphi(y)$; i.e., $x \leq y$. Conversely, assume that $x \leq y$. For the moment, x and y are comparable; since, otherwise, the imposed condition gives

$$\varphi(x \wedge y) \leq \varphi(x) - \varphi(y) \leq \varphi(x) \text{ [hence } \omega(x \wedge y) \geq \omega(x)];$$

in contradiction with a previous relation involving these data. The alternative $y \preceq x$ yields (by the first part) $y \leq x$; wherefrom (as $(\leq)$ is order) $x = y$. Hence, anyway $x \preceq y$; and conclusion follows.

(D) We are now in position to complete the argument. As $(M, \mathscr{R})$ is a proper structure, we necessarily have

$$\max(X, \preceq) = \emptyset; \text{ i.e. : for each } x \in X \text{ there exists } y \in X \text{ with } x \prec y.$$

This, along with (EVPm-dLc), tells us that (cf. a previous characterization)

d is not $(\leq)$-strongly-complete: there is an ascending (modulo $(\leq)$) sequence (x_n) in X, that is not d-convergent.

By the above equivalence property, (x_n) is ascending (modulo $(\preceq)$); hence,

$(\omega(x_n); n \geq 0)$ is ascending: $i \leq j$ implies $\omega(x_i) \leq \omega(x_j)$.

Two alternatives are under discussion.

Case 1. Suppose that there exists some index $k \geq 0$ such that

$x_k = x_n$, (hence, $\omega(x_k) = \omega(x_n)$), for all $n \geq k$.

In this case, by definition, $x_n \xrightarrow{d} x_k$; in contradiction with the initial choice of our sequence.

Case 2. Suppose that

for each rank p, there exists a rank $q > p$, such that $x_p \prec x_q$.

This tells us that

$B(h) := \{n \in N(h, <); x_n \succ x_h\} \neq \emptyset$, for all $h \in N$.

As a consequence of this, the self-map (of N) $(F(n) = \min B(n); n \geq 0)$ is well defined (without any use of DC); in addition, by the very definition above,

$$F(h) > h, x_{F(h)} \succ x_h, \quad \text{for all } h \in N.$$

By the first half of this relation, $(p(n) := F^n(0); n \geq 0)$ is strictly ascending, with $p(0) = 0$; hence, $p(n) \geq n$, for all n. It therefore generates a subsequence $(y_n := x_{p(n)}; n \geq 0)$ of (x_n) with the supplementary property

$$i < j \text{ implies } y_i \prec y_j \text{ (hence, } \omega(y_i) < \omega(y_j)).$$

Note that, as a consequence of this,

$$\omega(y_0) \geq 2; \text{ hence, } \omega(y_n) \geq n + 2, \ n \in N.$$

The sequence $(c_n := y_n(n); n \geq 0)$ is therefore well defined in M; moreover, by this ascending property, $c_0 = a$ and $c_n \mathscr{R} c_{n+1}$, for all n. This gives us the desired conclusion; and completes the argument.

In particular, when the specific assumptions (c06) and (c07) are ignored in Theorem 8, Proposition 7 is comparable with the one in Brunner [9]. For a different perspective about this problem, we refer to Dodu and Morillon [15].

Note finally that—by the developments in Turinici [48] we already quoted—the obtained equivalencies comprise the maximal principles in Bae et al. [2], Brøndsted [7], Dancs et al. [14], Ekeland [16], Szaz [40], and Turinici [43]. Moreover, these include as well some basic statements in Zhu and Li [51], obtainable via separable techniques; we do not give details.

Metrical Efficiency

Let X be a nonempty set. Remember that, by a *generalized metric* over X we mean, as in Luxemburg [33] and Jung [29], any map $(x, y) \mapsto d(x, y)$ from $X \times X$ to $R_+ \cup \{\infty\} = [0, \infty]$, endowed with the properties

d is *triangular*: $d(x, z) \leq d(x, y) + d(y, z)$, $\forall x, y, z \in X$
d is *reflexive-sufficient*: $d(x, y) = 0$ iff $x = y$
d is *symmetric*: $d(x, y) = d(y, x)$, $\forall x, y \in X$;

in this case, (X, d) will be referred to as a *generalized metric space*. Assume that we fixed such a structure in the sequel. The sequential convergence $(\xrightarrow{d})$ attached to d is defined as in the standard metrical case:

$$x_n \xrightarrow{d} x \text{ iff } d(x_n, x) \to 0 \text{ as } n \to \infty;$$

and reads: x is a d-limit of (x_n); when such elements x exist, we say that (x_n) is *d-convergent*. Further, the d-Cauchy property of a sequence is also introduced as in the standard metrical case:

$$d(x_n, x_m) \to 0, \text{ as } n, m \to \infty.$$

Finally, let $(\leq)$ be some quasi-order on X. Each ascending (modulo $(\leq)$) d-convergent sequence is d-Cauchy too; when the reciprocal is also true, we say that d is $(\leq)$-*complete*. Remember that the part M of X is called $(\leq)$-*closed* (modulo d), provided

the limit of each ascending (modulo $(\leq)$) sequence in M belongs to M.

In the same context, we say that $(\leq)$ is *self-closed* (modulo d), provided

$X(u, \leq)$ is $(\leq)$-closed (modulo d), for each $u \in X$.

(A) Having these precise, we may now pass to the effective part of our developments. Let (X, d) be a generalized metric space; and $(\leq)$ some quasi-order on it. Further, let A be some nonempty part of X. We say that $z \in A$ is Pareto $(\leq)$-*efficient* when

$A(z, \leq) = \{z\}$ [i.e.: $x \in A, z \leq x \Longrightarrow z = x$].

The class of all these will be denoted $\text{Eff}(A; \leq)$. To get sufficient conditions for the existence of such points, the monotone Ekeland Variational Principle (EVPm) will be used. This will be done under the basic regularity conditions upon our data

(d01) $(\leq)$ is self-closed and A is $(\leq)$-closed (modulo d)
(d02) d is $(\leq)$-complete (see above).

The specific concept to be considered is being constructed with the aid of certain pairs (u, g); where u is an element of A and $(x, y) \mapsto g(x, y)$ is a function from $\text{gr}(\leq) := \{(x, y) \in X \times X; x \leq y\}$ to $R_+ \cup \{\infty\}$. Precisely, let us say that $u \in A$ is *starting* (modulo $(d; \leq; A, g)$), when

(d03) g is subordinated to $(d; \leq; A, u)$:
$x, y \in A(u, \leq), x \leq y \Longrightarrow d(x, y) \leq g(x, y)$
(d04) g is $(\leq; A, u)$-antitriangular:
$g(x, z) \geq g(x, y) + g(y, z)$, if $x, y, z \in A(u, \leq), x \leq y \leq z$
(d05) $x \mapsto \varphi(x) := g(u, x)$ is bounded above on $A(u, \leq)$.

We are now in position to state our first main result in this exposition.

Theorem 9 *Let the general conditions (d01)+(d02) be admitted. Then, for each starting (modulo $(d; \leq; A, g)$) $u \in A$, there exists $v = v(u) \in A$, with*

*(**41-a**) $u \leq v$, $d(u, v) \leq g(u, v) - g(u, u)$*
*(**41-b**) v is Pareto $(\leq)$-efficient (see above).*

Proof Denote $M = A(u, \leq)(= A \cap X(u, \leq))$. By (d01), M $(\leq)$ is self-closed (modulo d) and d is $(\leq)$-complete over M. Wherefrom (taking (d02) into account)

$$(\leq) \text{ is self} - \text{closed (modulo } d) \text{ and } d \text{ is } (\leq) - \text{complete over} M.$$

Let $x, y \in M$ be arbitrary fixed with $x \leq y$. From (d03), $d(x, y) \leq g(x, y)$. On the other hand, (d04) yields (with (u, x, y) in place of (x, y, z))

$$g(u, y) \geq g(u, x) + g(x, y);$$

so, combining with the above,

$$(x, y \in M) : \ x \leq y \Longrightarrow d(x, y) \leq \varphi(y) - \varphi(x) \ (\text{hence } \varphi(y) \geq \varphi(x)).$$

Finally, (d05) yields

$$0 \leq \varphi(x) = g(u, x) < \infty, \ \forall x \in M.$$

Summing up, (EVPm) applies in (M, d) and $(\leq, -\varphi)$ (under its "purely" monotone version). So, for the point $u \in M$ there must be another one $v \in M$ with

(41-c) $u \leq v, d(u, v) \leq \varphi(v) - \varphi(u)$
(41-d) $x \in M, v \leq x, d(v, x) \leq \varphi(x) - \varphi(v) \Longrightarrow v = x$.

The former of these yields the first half of our conclusions, by the very definition of φ. And the latter one gives at once $v \in \text{Eff}(A; \leq)$; because, in view of $v \in M$ (hence, $u \leq v \in A$) we have (by a previous relation)

$$v \leq x \in A \Longrightarrow v \leq x \in M \Longrightarrow d(v, x) \leq \varphi(x) - \varphi(v);$$

whence (by this choice) $v = x$. The proof is thereby complete.

(B) A linear normed version of this result may be given as follows. Let X be a real linear space. By a *generalized norm* over X we shall mean any map $||.|| : X \to R_+ \cup \{\infty\}$, endowed with

$||.||$ is *subadditive*: $||x + y|| \leq ||x|| + ||y||, \forall x, y \in X$
$||.||$ is *absolutely homogeneous*: $||\tau x|| = |\tau| \cdot ||x||, \forall \tau \in R \setminus \{0\}, \forall x \in X$
$||.||$ is *reflexive sufficient*: $||x|| = 0 \Longleftrightarrow x = 0$.

In this case, its associated function $d : X \times X \to R_+ \cup \{\infty\}$, introduced as:

$$d(x, y) = ||x - y||, x, y \in X$$

is a generalized metric (on X); compatible with the linear structure of X

$$d(x + a, y + a) = d(x, y), d(\lambda x, \lambda y) = |\lambda| d(x, y), \ \forall x, y, a \in X, \ \forall \lambda \in R \setminus \{0\}.$$

All notions related to d may be now interpreted as $||.||$-notions. For example, the sequential convergence $(\xrightarrow{d})$ attached to d will be also written as $(\xrightarrow{||.||})$; because

$$x_n \xrightarrow{||.||} x \text{ iff } ||x_n - x|| \to 0 \text{ as } n \to \infty;$$

when such elements x exist, we say that (x_n) is $||.||$-*convergent*. Likewise, the d-Cauchy property of the sequence (x_n) will be also referred to as $||.||$-Cauchy; because it is characterized as

$$||x_n - x_m|| \to 0, \text{ as } n, m \to \infty.$$

Finally, let $(\leq)$ be some quasi-order on X. Each ascending (modulo $(\leq)$) $||.||$-convergent sequence is $||.||$-Cauchy; when the reciprocal is also true, we say that $||.||$ is $(\leq)$-*complete*. For each part M of X, its $(\leq)$-closedness (modulo d) will be also referred to as $(\leq)$-closedness (modulo $||.||$). In particular, the self-closedness (modulo d) of $(\leq)$ is to be viewed as a self-closedness (modulo $||.||$); and means:

$X(u, \leq)$ is $(\leq)$-closed (modulo $||.||$), for each $u \in X$.

Now, let $(X, ||.||)$ be taken as above. By an *additive cone* in X we shall understand any part K of X, with

K is *additive* $[K + K \subseteq K]$, and *pointed* $[0 \in K]$.

Let $(\leq)$ stand for the induced quasi-order

$(x, y \in X)\ x \leq y$ iff $y - x \in K$.

All notions related to it may be also viewed as K-notions; the list of these is the one we just described. Let A be some nonempty part of X. The concept of Pareto $(\leq)$-*efficient* point was already introduced; it will be referred to as a Pareto K-*efficient* point. The class of all these will be denoted $\mathrm{Eff}(A; \leq)$; or, equivalently, $\mathrm{Eff}(A; K)$. To get sufficient conditions for the existence of such points in this linear setting, some general and specific hypotheses must be accepted. The general ones may be written as (see above)

(d06) K and A are K-closed (modulo $||.||$)
(d07) $||.||$ is K-complete (see above).

The specific ones are being expressed in terms of a certain function $\psi : K \to R_+ \cup \{\infty\}$. Precisely, assume that

(d08) ψ is subordinated to $(||.||; K)$: $x \in K \Longrightarrow ||x|| \leq \psi(x)$
(d09) ψ is super-additive on K: $\psi(x + y) \geq \psi(x) + \psi(y), \forall x, y \in K$.

Further, call $u \in A$, (normed) *admissible* (modulo ψ) if

(n-ad) $x \mapsto \varphi(x) := \psi(x - u)$ is bounded above on $A \cap (u + K)$.

Note that, as a consequence of this,

$\psi(0) = 0$; hence, $0 \in \mathrm{Dom}(\psi)$.

In fact, by the (normed) admissible condition,

$\varphi(x) < \infty$, for some $x \in A \cap (u + K)$;
so that, $\psi(z) < \infty$, where $z := x - u \in K$.

On the other hand, by the super-additive property,

$\psi(z) = \psi(z + 0) \geq \psi(z) + \psi(0)$ (whence, $\psi(0) \leq 0$);

and, from this, we are done.

The following "normed" efficiency result is valid.

Theorem 10 *Let the general conditions (d06)+(d07) be admitted; as well as the specific ones (d08)+(d09). Then, for each admissible (modulo ψ) point $u \in A$, there exists $v = v(u) \in A$, such that*

*(**42-a**)* $u \leq v$, $||u - v|| \leq \psi(v - u)$
*(**42-b**)* v *is Pareto K-efficient (see above).*

Proof Let the function $g : \mathrm{gr}(\leq) \to R_+ \cup \{\infty\}$ be defined as:

$g(x, y) = \psi(y - x)$, $(x, y) \in \mathrm{gr}(\leq)$.

We claim that the conditions of Theorem 9 are fulfilled by $(d, \leq)$ and (u, g) over A. This is evident—via (d06)+(d07)—for the couple (d01)+(d02); so, we only have to show that—under (d08)+(d09)—the admissible (modulo ψ) point $u \in A$ is starting (modulo $(d; \leq; A, g)$); i.e.: the requirements (d03)-(d05) are fulfilled here. To do this, note that (d03) follows from (d08); and (d05) is a direct consequence of admissible property (if we remember the definition of g). Finally, let $x, y, z \in A(u, \leq)$ be such that $x \leq y \leq z$. By definition, $y - z \in K$, $z - y \in K$ (hence $z - x \in K$); and this, along with (d09) yields

$$g(x, z) = \psi(z - x) \geq \psi(y - x) + \psi(z - y) = g(x, y) + g(y, z);$$

hence (d04) follows as well. The proof is thereby complete.

This result includes the one in Isac [27, Theorem 3], obtained (with our notations) under the stronger form of (normed) admissible condition

(n-ad-s) $x \mapsto \psi(x - u)$ is bounded above continuous on $A \cap (u + K)$.

Hence, the continuity condition may be removed; this is also true for the condition

K is strongly pointed: $K \cap (-K) = \{0\}$;

we do not give details.

Finally, note that the multiplicative properties of our linear space are not used; so, this result is extendable to the case of X being a topological Abelian group. Further aspects may be found in Isac and Tammer [28].

Fang Spaces

Let X be some nonempty set. By a *generalized pseudometric* over X we shall mean any map $d : X \times X \to R_+ \cup \{\infty\}$, fulfilling

d is *reflexive*: $d(x, x) = 0, \forall x \in X$;

if in addition

d is *symmetric*: $d(x, y) = d(y, x), \forall x, y \in X$

then d is referred to as a *generalized s-pseudometric*.

(A) Let $(\Lambda, \leq)$ be some directed quasi-ordered structure without maximal element. We say that the family $\mathscr{D} = (d_\lambda; \lambda \in \Lambda)$ of generalized s-pseudometrics over X is Λ-*admissible*, when

(e01) $(d_\lambda(x,y) \leq d_\mu(x,y), \forall x,y \in X)$, if $\lambda \leq \mu$ [$\mathscr{D}$ is Λ-monotone]
(e02) $\forall \lambda \in \Lambda, \exists \mu \in \Lambda(\lambda, \leq)$, with
$d_\lambda(x,z) \leq d_\mu(x,y) + d_\mu(y,z), \forall x,y,z \in X$ [$\mathscr{D}$ is Λ-triangular].

Assume that we fixed such a family; with, in addition

(e03) $(d_\lambda(x,y) = 0, \forall \lambda \in \Lambda)$ imply $x = y$ [$\mathscr{D}$ is sufficient].

Its associated family of relations $\mathscr{V} = \{V(\lambda, r); \lambda \in \Lambda, r > 0\}$, where

$$V(\lambda, r) = \{(x,y) \in X \times X; d_\lambda(x,y) < r\}, \lambda \in \Lambda, r > 0$$

is a fundamental system of entourages for a uniform structure $\mathscr{U} = \mathscr{U}(\mathscr{D})$ over X (cf. Bourbaki [5, Chap. 2, Sect. 1]). This structure, introduced in 1996 by Fang [17], became a very useful instrument in the probabilistic and fuzzy metric spaces theory. As a rule, the "uniform" terminology refers to it. However (as results directly by definition), all $\mathscr{U}$-notions are in fact $\mathscr{V}$-notions; so, we shall work with $\mathscr{V}$ in place of $\mathscr{U}$. Moreover, all these $\mathscr{V}$-notions may be translated in terms of $\mathscr{D}$; so, the initial uniform structure $(X, \mathscr{V})$ may be written as $(X, \mathscr{D})$; and referred to as a *Fang uniform space*. To motivate our assertion, remember that the associated (sequential) convergence structure $(\mathscr{V})$ on X may be described as

$$x_n \xrightarrow{(\mathscr{V})} x \text{ iff } \forall V \in \mathscr{V}, \exists n(V) : \ n \geq n(V) \Longrightarrow (x_n, x) \in V.$$

For simplicity, it will be referred to as: x is the $\mathscr{V}$-*limit* of (x_n); if such elements x exist, we shall say that (x_n) is $\mathscr{V}$-*convergent*. Likewise, the $\mathscr{V}$-*Cauchy* property for a sequence (x_n) in X may be introduced as

$$\forall V \in \mathscr{V}, \exists n(V) : n(V) \leq p \leq q \Longrightarrow (x_p, x_q) \in V.$$

Now, both these notions may be translated in terms of $\mathscr{D}$. In fact, the convergence relation $x_n \xrightarrow{(\mathscr{V})} x$ means:

$$x_n \xrightarrow{d_\lambda} x, \text{ for each } \lambda \in \Lambda; \text{ written as: } x_n \xrightarrow{\mathscr{D}} x.$$

This, by convention, reads: x is a $\mathscr{D}$-*limit* of (x_n); when such elements x exist, we say that (x_n) is $\mathscr{D}$-*convergent*. On the other hand, the $\mathscr{V}$-Cauchy property of a sequence (x_n) in X amounts to:

(x_n) is d_λ-Cauchy, for each $\lambda \in \Lambda$; referred to as: (x_n) is $\mathscr{D}$-*Cauchy*.

Putting these together, proves our assertion.

Finally, let $(\leq)$ be some quasi-order on X. Each ascending (modulo $(\leq)$) $\mathscr{D}$-convergent sequence is $\mathscr{D}$-Cauchy; when the reciprocal holds too, we say that $\mathscr{D}$ is *sequentially* $(\leq)$-*complete*. Given the (nonempty) part M of X, call it *sequentially* $(\leq)$-*closed* (modulo $\mathscr{D}$), provided:

the $\mathscr{D}$-limit of each ascending (modulo $(\leq)$) sequence in M belongs to M.

In particular, $(\leq)$ will be referred to as *sequentially self-closed* (modulo $\mathscr{D}$), when

$X(u, \leq)$ is sequentially $(\leq)$-closed (modulo $\mathscr{D}$), for each $u \in X$.

Technically speaking, the uniformity $\mathscr{U} = \mathscr{U}(\mathscr{D})$ is not in general metrizable; i.e., it is not generated by a standard metric. But, from a generalized perspective, this is possible. Precisely, define the mapping $\Delta : X \times X \to R_+ \cup \{\infty\}$ as

$$\Delta(x, y) = \sup\{d_\lambda(x, y); \lambda \in \Lambda\}, x, y \in X.$$

By the imposed properties of $\mathscr{D}$ (and Λ), it follows that

Δ is *triangular*: $\Delta(x, z) \leq \Delta(x, y) + \Delta(y, z), \forall x, y, z \in X$
Δ is *reflexive-sufficient:* $\Delta(x, y) = 0$ iff $x = y$
Δ is *symmetric*: $\Delta(x, y) = \Delta(y, x), \forall x, y \in X$.

In other words, Δ is a *generalized metric* over X, as in Luxemburg [33] and Jung [29]. Its associated uniform structure ($\mathscr{U} = \mathscr{U}(\Delta)$) is the one for which the family of relations $\mathscr{V} = \{V(\varepsilon); \varepsilon > 0\}$, where

$$V(\varepsilon) = \{(x, y \in X \times X; \Delta(x, y) < \varepsilon\}, \varepsilon > 0$$

is a fundamental system of entourages. This, in turn, gives us the associated (sequential) convergence ($\xrightarrow{\Delta}$) and Cauchy structure. The natural question to be posed is that of clarifying the relationships between these and the ones attached to the family $\mathscr{D} = (d_\lambda; \lambda \in \Lambda)$.

Proposition 8 *The generic local inclusions hold:*

*(**51-1**) $(\forall (x_n), \forall x)$ $[x_n \xrightarrow{\Delta} x] \Longrightarrow [x_n \xrightarrow{\mathscr{D}} x]$*
*(**51-2**) $(\forall$ sequence$)$ Δ-Cauchy $\Longrightarrow$ $\mathscr{D}$-Cauchy.*

As a consequence of this, we have the generic implication (for $M \in (2)^X$)

*(**51-3**) sequentially $(\leq)$-closed (modulo $\mathscr{D}$) $\Longrightarrow$ $(\leq)$-closed (modulo Δ).*

The proof is immediate, by the involved conventions; we do not give details. Note that the reciprocal of these is not in general true; because the uniform structure attached to $\mathscr{D}$ is strictly finer (in general) than the one induced by the generalized metric Δ.

Proposition 9 *Under these notations,*

*(**52-1**) $(\forall (x_n), \forall x)$ $[(x_n)$ is Δ-Cauchy$]$ and $[x_n \xrightarrow{\mathscr{D}} x]$ imply $[x_n \xrightarrow{\Delta} x]$*
*(**52-2**) $\mathscr{D}$ is sequentially $(\leq)$-complete $\Longrightarrow$ Δ is $(\leq)$-complete.*

Proof The second part in the statement follows at once from the first part of the same; so, it will suffice proving that the first part of our statement holds. Let (x_n) be some Δ-Cauchy sequence in X, so as (for some $x \in X$)

$$x_n \xrightarrow{\mathscr{D}} x \text{ (hence } d_\lambda(x_n, x) \to 0, \text{ for each } \lambda \in \Lambda).$$

By definition, for each $\beta > 0$ there exists some rank $n(\beta)$ in such a way that

$(d_\lambda(x_i, x_j) \leq \beta, \forall \lambda \in \Lambda)$, whenever $n(\beta) \leq i \leq j$.

Let the rank $i \geq n(\beta)$ be arbitrary fixed; and, for each $\lambda \in \Lambda$, let $\mu \in \Lambda(\lambda, \leq)$ be the index assured by the Λ-triangular property of $\mathscr{D}$. From the above relation we then have, for all such (λ, μ),

$$d_\lambda(x_i, x) \leq d_\mu(x_i, x_j) + d_\mu(x_j, x) \leq \beta + d_\mu(x_j, x), \ \forall j \geq i.$$

Passing to limit upon j gives (for all such i)

$$d_\lambda(x_i, x) \leq \beta, \ \forall \lambda \in \Lambda \ \ (\text{hence } \Delta(x_i, x) \leq \beta).$$

This, by the arbitrariness of β, gives $x_n \xrightarrow{\Delta} x$; as claimed.

(B) A linear (locally convex) version of this result may be given as follows. Let X be a real linear space. By a *(generalized) seminorm* on X we mean any function $|.| : X \times X \to R_+ \cup \{\infty\}$, endowed with the properties

$|.|$ is *subadditive*: $|x + y| \leq |x| + |y|, \forall x, y \in X$
$|.|$ is *absolutely homogeneous*: $|\tau x| = |\tau| \cdot |x|, \forall \tau \in R \setminus \{0\}, \forall x \in X$
$|.|$ is *reflexive*: $|0| = 0$.

In this case, its associated function $d : X \times X \to R_+ \cup \{\infty\}$ introduced as:

$d(x, y) = |x - y|, x, y \in X$

is a generalized semimetric over X; which, in addition, is compatible with the linear structure of X. All notions related to d may be now interpreted as $|.|$-notions. For example, the sequential convergence $(\xrightarrow{d})$ will be also written as $(\xrightarrow{|.|})$; because

$x_n \xrightarrow{|.|} x$ iff $|x_n - x| \to 0$ as $n \to \infty$;

when such elements x exist, we say that (x_n) is $|.|$-*convergent*. Likewise, the d-Cauchy property of a sequence will be also referred to as $|.|$-Cauchy, in view of

(x_n) is d-Cauchy iff $|x_n - x_m| \to 0$ as $n, m \to \infty$.

Finally, let $(\leq)$ be some quasi-order on X. Clearly, each ascending (modulo $(\leq)$) $|.|$-convergent sequence is $|.|$-Cauchy; when the reciprocal is also true, we say that $|.|$ is $(\leq)$-*complete*. For each part M of X, its $(\leq)$-closedness (modulo d) will be also referred to as $(\leq)$-closedness (modulo $|.|$). In particular, the self-closedness (modulo d) of $(\leq)$ is to be viewed as a self-closedness (modulo $|.|$) of $(\leq)$.

Now, let $(\Lambda, \leq)$ be a directed set without maximal element; and $\mathscr{P} = \{|.|_\lambda; \lambda \in \Lambda\}$ be a family of generalized seminorms over X, with

(e04) $\lambda \leq \mu \Longrightarrow |.|_\lambda \leq |.|_\mu$ [$\mathscr{P}$ is Λ-monotone]
(e05) $(|x|_\lambda = 0, \forall \lambda \in \Lambda) \Longrightarrow x = 0$. [$\mathscr{P}$ is separated].

[Note that there is a natural possibility of getting such families, by starting from each (infinite) directed separated family of (generalized) seminorms. In fact, let I be an infinite (index) set; and $\mathscr{Q} = \{||.||_i; i \in I\}$ be a family of generalized seminorms over X; supposed to be *separated* (see above) and *directed*:

(e06) $\forall i, j \in I, \exists k \in I$, such that $\max\{||.||_i, ||.||_j\} \leq ||.||_k$.

Denote by Λ the class of all (nonempty) finite parts of I; and let the quasi-order $(\leq)$ over it be the usual inclusion; clearly, $(\Lambda, \leq)$ is directed and has no maximal elements. For each $\lambda \in \Lambda$, define the generalized seminorm: $|.|_\lambda = \max\{||.||_i; i \in \lambda\}$. The family $\mathscr{P} = \{|.|_\lambda; \lambda \in \Lambda\}$ of all such objects fulfills (e04)+(e05), as it can be directly seen. In addition, (e06) tells us that it is equivalent with the family $\mathscr{Q}$; see, for instance, Precupanu [38, Chap. 3, Sect. 3.1]; and this proves the claim]. Now, for each $\lambda \in \Lambda$, let d_λ stand for the associated generalized semimetric (see above). As a consequence, the family $\mathscr{D} = \{d_\lambda; \lambda \in \Lambda\}$ is Λ-admissible and separated; so, it induces a uniform structure $\mathscr{U} = \mathscr{U}(\mathscr{D})$ over X by means of the fundamental system of entourages $\mathscr{V} = \{V(\lambda, \tau); \lambda \in \Lambda, \tau > 0\}$, we just described. In addition, it is separated by (e05); see Bourbaki [5, Chap. 2, Sect. 1.2] for details. On the other hand, the family of subsets $\mathscr{Z} = \{Z(\lambda, r); \lambda \in \Lambda, r > 0\}$, where

$$Z(\lambda, r) = \{x \in X; |x|_\lambda < r\}, \lambda \in \Lambda, r > 0,$$

is a fundamental system of convex zero neighborhoods for a linear topology $\mathscr{T} = \mathscr{T}(\mathscr{P})$ over X; referred to as the locally convex topology induced by $\mathscr{P}$; note that, by (e05) again, $\mathscr{T}$ is separated. Concerning the relationships between these, the uniformity $\mathscr{U}$ over X is just the one induced by $\mathscr{T}$ via the canonical map $Z(\lambda, r) \mapsto V(\lambda, r)$; conversely, the locally convex topology $\mathscr{T}$ over X is just the one induced by $\mathscr{U}$, by means of the canonical map $V(\lambda, r) \mapsto Z(\lambda, r)$; we do not give details. As a consequence of this, any concept related to $\mathscr{D}$ may be interpreted as a $\mathscr{P}$-concept. For example, the associated sequential convergence structure $(\xrightarrow{\mathscr{D}})$ will be denoted as $(\xrightarrow{\mathscr{P}})$; and means

$$x_n \xrightarrow{\mathscr{P}} x \text{ iff } x_n \xrightarrow{|.|_\lambda} x, \text{ for each } \lambda \in \Lambda;$$

and reads: x is a *$\mathscr{P}$-limit* of (x_n); if such elements x exist, we say that (x_n) is *$\mathscr{P}$-convergent*. Also, the $\mathscr{D}$-Cauchy property of a sequence (x_n) be denoted as: (x_n) is *$\mathscr{P}$-Cauchy*; and means

(x_n) is $|.|_\lambda$-Cauchy, for each $\lambda \in \Lambda$.

Finally, let $(\leq)$ be some quasi-order on X. Each ascending (modulo $(\leq)$) $\mathscr{P}$-convergent sequence is $\mathscr{P}$-Cauchy; when the reciprocal holds too, we say that $\mathscr{P}$ is *sequentially* $(\leq)$-*complete*. For each part M of X, its sequential $(\leq)$-closedness (modulo $\mathscr{D}$) will be also referred to as a sequential $(\leq)$-closedness (modulo $\mathscr{P}$); and means:

the $\mathscr{P}$-limit of each ascending (modulo $(\leq)$) sequence in M belongs to M.

In particular, the sequential self-closedness (modulo $\mathscr{D}$) of $(\leq)$ is to be viewed as a sequential self-closedness (modulo $\mathscr{P}$) of $(\leq)$; and may be characterized as:

$X(u, \leq)$ is sequentially $(\leq)$-closed (modulo $\mathscr{P}$), for each $u \in X$.

(C) Technically speaking, the locally convex topology $\mathscr{T} = \mathscr{T}(\mathscr{P})$ is not in general normable; i.e., it is not generated by a standard norm. But, from a generalized perspective, this is possible. Precisely, denote

$(||.|| : X \to R_+ \cup \{\infty\})$: $||x|| = \sup\{|x|_\lambda; \lambda \in \lambda\}, x \in X.$

This map is a generalized seminorm on X; and, by (e05), it is also sufficient; hence, $||.||$ is a generalized norm on X. Its associated map

$(\Delta : X \times X \to R_+ \cup \{\infty\})$: $\Delta(x, y) = ||x - y||, x, y \in X$

is therefore a generalized metric on X; which is also compatible with the linear structure of X. As before, all notions related to Δ may be also viewed as $||.||$-notions; we do not give details.

Summing up, we have two "parallel" structures over X induced by $\mathscr{P} = \{|.|_\lambda; \lambda \in \Lambda\}$ and $||.||$, respectively. So, it would be useful to establish a lot of relationships between the associated concepts we just introduced.

Proposition 10 *The generic local inclusions hold:*

*(**53-1**)* $(\forall (x_n), \forall x)$ $[x_n \xrightarrow{||.||} x] \Longrightarrow [x_n \xrightarrow{\mathscr{P}} x]$
*(**53-2**) ($\forall$ sequence) $||.||$-Cauchy $\Longrightarrow$ $\mathscr{P}$-Cauchy.*

As a consequence of this, we have the generic implication (for $M \in (2)^X$)

*(**53-3**) sequentially $(\leq)$-closed (modulo $\mathscr{P}$) $\Longrightarrow$ $(\leq)$-closed (modulo $||.||$).*

The reciprocal of these is not in general true; because the locally convex structure attached to $\mathscr{P}$ is strictly finer (in general) than the one induced by the generalized norm $||.||$.

Proposition 11 *Under these notations,*

*(**54-1**) $(\forall (x_n), \forall x)$ [(x_n) is $||.||$-Cauchy] and $[x_n \xrightarrow{\mathscr{P}} x]$ imply $[x_n \xrightarrow{||.||} x]$*
*(**54-2**) $\mathscr{P}$ is sequentially $(\leq)$-complete $\Longrightarrow$ $||.||$ is $(\leq)$-complete.*

The proof of these statements is the linear version of the (uniform) one above; so, we do not give details.

Uniform Pareto Efficiency

In the following, a uniform variant is given for the metrical type results involving Pareto efficient points.

(A) Let X be a nonempty set. Take some directed quasi-ordered structure $(\Lambda, \leq)$ without maximal element; as well as a family $\mathscr{D} = (d_\lambda; \lambda \in \Lambda)$ of s-pseudometrics over X; supposed to be Λ-admissible and sufficient (see above). Further, let $(\leq)$ be a quasi-order on X; and A be some nonempty part of X. The notion of Pareto $(\leq)$-*efficiency* is the described one: we say that $z \in A$ is Pareto $(\leq)$-*efficient*, when

$A(z, \leq) = \{z\}$ [i.e.: $x \in A, z \leq x \Longrightarrow z = x$].

To establish an existence result (involving such points) in our extended setting, we start from the basic conditions

(f01) $(\leq)$ is sequentially self-closed and A is sequentially $(\leq)$-closed (modulo $\mathscr{D}$)
(f02) $\mathscr{D}$ is sequentially $(\leq)$-complete.

The specific ones are being expressed in terms of a certain pair (u, g); where u is an element of A and $g : \mathrm{gr}(\leq) \to R_+ \cup \{\infty\}$ is a function. Precisely, let us say that $u \in A$ is *starting* (modulo $(\mathscr{D}; \leq; A, g)$), when

(f03) g is subordinated to $(\mathscr{D}; \leq; A, u)$:
$x, y \in A(u, \leq), x \leq y \Longrightarrow d_\lambda(x, y) \leq g(x, y)$, for all $\lambda \in \Lambda$
(f04) g is $(\leq; A, u)$-antitriangular:
$x, y, z \in A(u, \leq), x \leq y \leq z \Longrightarrow g(x, z) \geq g(x, y) + g(y, z)$
(f05) $x \mapsto \varphi(x) := g(u, x)$ is bounded above on $A(u, \leq)$.

We are now in position to state an appropriate answer to our question.

Theorem 11 *Let the general conditions (f01)+(f02) be admitted. Then, for each starting (modulo $(\mathscr{D}; \leq; A, g)$) point $u \in A$, there exists $v = v(u) \in A$, in such a way that*

*(**61-a**)* $u \leq v,\ d_\lambda(u, v) \leq g(u, v) - g(u, u), \forall \lambda \in \Lambda$
*(**61-b**) v is Pareto $(\leq)$-efficient (see above).*

Proof Let Δ stand for the generalized metric on X attached to $\mathscr{D}$. From our general results involving such structures, it follows that (by means of imposed conditions)

(f06) $(\leq)$ is self-closed and A is $(\leq)$-closed (modulo Δ)
(f07) Δ is $(\leq)$-complete (see above)
(f08) g is subordinated to $(\Delta; \leq; A, u)$:
$x, y \in A(u, \leq), x \leq y \Longrightarrow \Delta(x, y) \leq g(x, y)$.

This, along with (f04)+(f05), tells us that Theorem 9 is applicable to these data. It gives us, for the starting point $u \in A$, some other point $v = v(u) \in A$, with the properties described in the quoted statement. But, from this, our desired conclusions are clear. The proof is complete.

From the argument above, it follows that Theorem 11 is deductible from Theorem 9. The reciprocal is also true; just take Λ (hence $\mathscr{D}$ as well) as a singleton. Hence, these two results are equivalent to each other. Some related aspects may be found in Hamel [25].

(B) In the following, a linear version of the result above is established. Let X be a (real) linear space, $(\Lambda, \leq)$ be a directed structure without maximal element, and $\mathscr{P} = \{|.|_\lambda; \lambda \in \Lambda\}$ be a Λ-monotone separated family of generalized seminorms (see above). Further, let K be an additive pointed cone in X; and let $(\leq)$ stand for its induced quasi-order. Finally, take some part A of X. The notion of Pareto $(\leq)$-*efficient* (or, equivalently, K-*efficient*) point is the already precise one. To get sufficient conditions for the existence of such points, some general and specific hypotheses must be accepted. The general ones may be written as

(f09) K and A are sequentially K-closed (modulo $\mathscr{P}$)
(f10) $\mathscr{P}$ is sequentially K-complete.

The specific ones are being expressed in terms of a certain function $\psi : K \to R_+ \cup \{\infty\}$. Precisely, assume that

(f11) ψ is subordinated to $(\mathscr{P}; K)$: $x \in K \Longrightarrow |x|_\lambda \leq \psi(x), \forall \lambda \in \Lambda$
(f12) ψ is super-additive on K: $\psi(x + y) \geq \psi(x) + \psi(y), \forall x, y \in K$.

Finally, call $u \in A$, *admissible* (modulo ψ) if

(lc-ad) $x \mapsto \varphi(x) := \psi(x - u)$ is bounded above on $A \cap (u + K)$.

The following "locally convex" efficiency result is valid.

Theorem 12 *Let the general conditions (f09)+(f10) be admitted; as well as the specific ones (f11)+(f12). Then, for each admissible (modulo ψ) $u \in A$, there exists $v = v(u) \in A$, in such a way that*

*(**62-a**) $u \leq v,\ |u - v|_\lambda \leq \psi(v - u), \forall \lambda \in \Lambda$*
*(**62-b**) v is Pareto K-efficient (see above).*

Proof For each $\lambda \in \Lambda$, let d_λ stand for the generalized semimetric attached to $|.|_\lambda$; and $\mathscr{D} = \{d_\lambda; \lambda \in \Lambda\}$ stand for the family of all these. Further, let the function $g : \mathrm{gr}(\leq) \to R \cup \{\infty\}$ be defined as:

$g(x, y) = \psi(y - x),\ (x, y) \in \mathrm{gr}(\leq)$.

It is not hard to see that conditions of the preceding statement hold for these data; and, from this, we are done.

(C) Note that an alternate proof of this result is available, by reducing it to the normed Pareto efficient point statement we already established. In fact, let $||.||$ stand for the generalized norm on X attached to $\mathscr{P}$ (see above). By our general developments, all conditions in the quoted statement are fulfilled; and then, by its conclusions, one gets all desired facts.

In particular, the regularity condition (lc-ad) holds under

(lc-ad-c) $x \mapsto \varphi(x) := \psi(x - u)$ is bounded above and
sequentially continuous on $A \cap (u + K)$.

Note that in such a case, our locally convex result includes a related statement in Isac [27, Theorem 6]. Hence, the continuity condition may be removed; this is also true for the condition

K is strongly pointed: $K \cap (-K) = \{0\}$;

we do not give details.

Further, when Λ reduces to a single point, the obtained result is nothing else than Theorem 10. The reciprocal is also true (by the proposed proof). Hence Theorem 12 is equivalent with its normed variant.

Finally, note that the multiplicative properties of our linear space are not used; so, this result is extendable to the case of X being a topological Abelian group. Further aspects will be discussed elsewhere.

Semigroup Anti-Measures

In what follows, the concept of *semigroup anti-measure* is introduced; and some basic properties of it are discussed.

(A) Let $(X, \leq)$ be some quasi-ordered structure. By a *semigroup* over X we shall mean any map $(t, x) \mapsto S(t)x$, from $R_+ \times X$ to X, with

(g01) $S(0)x = x$, $\forall t \geq 0$, $\forall x \in X$
(g02) $S(t+s)x = S(t)(S(s)x)$, $\forall t, s \geq 0$, $\forall x \in X$.

Assume that we fixed such an object; which, in addition, is *monotone*:

(g03) $t \leq s, x \leq y \Longrightarrow S(t)x \leq S(s)y$.

Denote in the following, for $(x, u) \in \mathrm{gr}(\leq)$,

$$\Theta(\leq, S; x, u) = \{t \in R_+; S(t)x \leq u\},\ \ \theta(\leq, S; x, u) = \sup \Theta(\leq, S; x, u).$$

This yields a couple of functions $\Theta(.,.) := \Theta(\leq, S; ., .)$ and $\theta(.,.) := \theta(\leq, S; ., .)$ from $\mathrm{gr}(\leq)$ to $(2)^{R_+}$ (=the class of all nonempty subsets of R_+) and $R_+ \cup \{\infty\}$, respectively; the latter of these will be referred to as the *anti-measure* attached to $(\leq, S)$. Technically speaking, this construction may be related to the one in Turinici [47]; but we must say that our setting is rather different from the quoted one. A motivation for our terminology will be offered later. For the moment, we shall be interested in giving some basic properties of these maps, to be used further.

(i) We start by noting that, for each $(x, u) \in \mathrm{gr}(\leq)$,

$$\Theta(x, u) \text{ is hereditary} : s \in \Theta(x, u) \Longrightarrow [0, s] \subseteq \Theta(x, u);$$

so, it is a nonempty (R_+)-initial interval $[0, \lambda[$ (where $0 < \lambda \leq \infty$) or $[0, \lambda]$ (where $0 \leq \lambda < \infty$).

(ii) Further, it would be useful to precise the behavior of $(x,u) \mapsto \theta(x,u)$ with respect to its variables. The following answer to this is valid:

$$x_1 \le x_2 \le u_2 \le u_1 \Longrightarrow \theta(x_1,u_1) \ge \theta(x_2,u_2).$$

In other words: the map $(x,u) \mapsto \theta(x,u)$ is decreasing in the first variable and increasing in the second one. [Since the proof is immediate, we do not give details].

(iii) In a close connection with the preceding fact, we have the following property:

Proposition 12 *Let $x_1, x_2, x_3 \in X$ be such that $x_1 \le x_2 \le x_3$. Then*

$$\theta(x_1,x_3) \ge \theta(x_1,x_2) + \theta(x_2,x_3) \quad (\mathit{super-additivity}).$$

Proof Without loss, one may assume that $\theta(x_1,x_2) > 0$, $\theta(x_2,x_3) > 0$. Let $t < \theta(x_1,x_2)$, $s < \theta(x_2,x_3)$ be arbitrary fixed. By definition, $S(t)x_1 \le x_2$, $S(s)x_2 \le x_3$; wherefrom (by the semigroup properties)

$$S(t+s)x_1 = S(s)(S(t)x_1) \le S(s)x_2 \le x_3; \text{ so, } t+s \in \Theta(x_1,x_3).$$

This, and the definition of the map θ, ends the argument.

This relation motivates our terminology; because for a standard (*subadditive*) measure of these order intervals (taken as in Halmos [24, Chap. 2, Sect. 9]) the inequality in our statement above is written with the dual sign $(\le)$.

(iv) Finally, a basic question about the map θ is that of describing its finite/infinite values. Call the pair (x,u) in gr$(\le)$, *admissible* (modulo θ) when $\theta(x,u) < \infty$. Note that, by the interval monotonicity of θ, one has the hereditary type property

if (x,u) is admissible (modulo θ) and $x \le x' \le u' \le u$
then (x',u') is admissible too (modulo θ).

This tells us that the class of all such couples (x,u) is large enough; we do not give further details.

(B) In the following, a linear construction of this type is considered. Let X be a (real) *vector space*. Take a *convex cone* H of X; i.e.

$$\alpha H + \beta H \subseteq H, \text{ for each } \alpha, \beta \in R_+;$$

and let $(\le)$ stand for its induced quasi-order. Further, choose some point $k^0 \in H \setminus (-H)$; and put (for $x \in H$)

$$\Gamma(H;k^0;x) = \{s \in R_+; k^0 s \le x\}, \quad \gamma(H;k^0;x) = \sup \Gamma(H;k^0;x).$$

We therefore defined a couple of functions $\Gamma(.) := \Gamma(H;k^0;.)$ and $\gamma(.) := \gamma(H;k^0;.)$ from H to $(2)^{R+}$ (=the class of all nonempty subsets in R_+) and $R_+ \cup \{\infty\}$, respectively; the latter of these will be referred to as the *gauge* function attached to $(H;k^0)$. Such objects were introduced (in the normed context) by Gerth

(Tammer) and Weidner [18]. Some refinements of these (to the locally convex setting) were provided by Goepfert et al. [20]. The present developments may be viewed as a non-topological extension of the quoted ones.

(j) We start by noting that, for each $x \in H$,

$$\Gamma(x) \text{ is hereditary } (s \in \Gamma(x) \Longrightarrow [0, s] \subseteq \Gamma(x));$$

so, it is a nonempty R_+-initial interval $[0, \alpha[$ (where $0 < \alpha \leq \infty$) or $[0, \alpha]$ (where $0 \leq \alpha < \infty$).

(jj) An interesting question is that of $\infty \in \gamma(H)$ being or not valid. According to Cristescu [13, Chap. 5, Sect. 1], let us say that H is *Archimedean*, provided

$[h \in X, v \in H$ and $\Gamma(H; h; v) = R_+]$ imply $h \in -H$.

Proposition 13 *Assume that H is Archimedean. Then, for each* $x \in H$,

*(***72-1***)* $\gamma(x) \neq \infty$*; hence,* $0 \leq \gamma(x) < \infty$
*(***72-2***)* $\gamma(x) \in \Gamma(x)$*; so that,* $\Gamma(x) = [0, \gamma(x)]$.

Proof The first part is clear, by the choice of k^0 (and Archimedean hypothesis). For the second one, let $x \in H$ be arbitrary fixed. If $\gamma(x) = 0$, then $\Gamma(x) = \{0\} = [0, \gamma(x)]$; and we are done. If $0 < \gamma(x) < \infty$, we have (by definition of supremum)

$k^0(\gamma(x) - t) \leq x$, if $0 < t \leq \gamma(x)$; so, $(k^0\gamma(x) - x)s \leq k^0$, $\forall s \in R_+$;
wherefrom (cf. the above conventions): $\Gamma(H; k^0\gamma(x) - x; k^0) = R_+$.

This and the Archimedean property of H give the desired fact.

(jjj) Returning to the general case, note that Γ and γ are *positively homogeneous*:

$$\Gamma(tx) = t\Gamma(x),\ \gamma(tx) = t\gamma(x),\ \forall t \in R_+,\ \forall x \in H.$$

On the other hand, the choice of k^0 yields

$$\gamma(0) = 0,\ \gamma(k^0) = 1;\ \text{ hence } \gamma(k^0 t) = t,\ \forall t \in R_+.$$

Further information may be obtained from the increasing property of γ

$$x_1, x_2 \in H, x_1 \leq x_2 \text{ implies } \gamma(x_1) \leq \gamma(x_2).$$

In fact, as a consequence of this,

$$\gamma(y) \leq t,\ \text{ whenever } y \in H, y \leq k^0 t;$$

so, $\mathrm{Dom}(\gamma) := \{x \in H; \gamma(x) < \infty\}$ is “large” enough.

(jv) Some other useful properties of such objects are formulated in

Proposition 14 *The gauge function γ is super-additive:*

*(**73-1**) $\gamma(x_1 + x_2) \geq \gamma(x_1) + \gamma(x_2)$, for each $x_1, x_2 \in H$.*

As a consequence, γ is subtractive:

*(**73-2**) $\gamma(x_1 - x_2) \leq \gamma(x_1) - \gamma(x_2)$,*
whenever $x_1, x_2 \in H$ fulfill $x_1 \geq x_2$ and $\gamma(x_1) - \gamma(x_2)$ exists.

Proof i): The case of $0 \in \{\gamma(x_1), \gamma(x_2)\}$ is clear, via γ=increasing; so, it remains to discuss the case of $\gamma(x_1) > 0$, $\gamma(x_2) > 0$. By definition (and hereditary property),

$x_1 \geq k^0 t_1, x_2 \geq k^0 t_2$, whenever $0 \leq t_1 < \gamma(x_1), 0 \leq t_2 < \gamma(x_2)$;

and this yields (for all such (t_1, t_2))

$x_1 + x_2 \geq k^0[t_1 + t_2]$ (i.e.: $\gamma(x_1 + x_2) \geq t_1 + t_2$).

This, and the arbitrariness of the precise couple, ends the argument.

ii): By the previous relation, we have

$$\gamma(x_1) = \gamma(x_1 - x_2 + x_2) \geq \gamma(x_1 - x_2) + \gamma(x_2).$$

If $\gamma(x_2) = \infty$, then (by the relation we just obtained) $\gamma(x_1) = \infty$; in contradiction with $\gamma(x_1) - \gamma(x_2)$ being well defined. Hence, necessarily, $0 \leq \gamma(x_2) < \infty$; and then, conclusion follows in either case involving $\gamma(x_1)$.

In particular, we have the sup-translation property

$$\gamma(y + k^0 t) \geq \gamma(y) + t, \text{ for each } y \in H, \ t \in R_+;$$

further aspects may be found in Turinici [46].

(C) The following remark about these developments is in effect for us. Let X be a (real) linear space; and H, some *convex cone* of it (see above). Denote by $(\preceq)$ the induced quasi-order; and pick some $k^0 \in H \setminus (-H)$. The mapping (from $R_+ \times X$ to X)

$S(t)x = x + tk^0, \ t \in R_+, x \in X$

is a monotone semigroup over X as it can be directly seen; we shall term it, the *semigroup attached to* (H, k^0). Let $(x, u) \mapsto \theta(x, u)$ stand for the anti-measure attached to $(\preceq, S)$; in fact, a natural position is to consider that it is attached to (H, k^0). The study of this object is entirely reduced to the one generated by our gauge function γ (attached to (H, k^0)); because

$$\theta(x, u) = \gamma(u - x), \text{ for all } (x, u) \in \text{gr}(\preceq).$$

As a consequence of this, all properties of the anti-measure θ are obtainable from the properties of the gauge function γ. Clearly, this may be also used in the converse way; because, from $\gamma(x) = \theta(0, x)$, $x \in H$, all properties of γ are obtainable from the ones in θ. We however preferred to deduce them in a direct way, for technical reasons. Further aspects may be found in Tataru [42].

(D) A non-trivial extension of our construction is to be performed under the lines below. By a *monotone almost semigroup* on X we shall mean any map $(t, x) \mapsto T(t)x$ from $R_+ \times X$ to X fulfilling

(mas-1) $T(0)x = x, \forall t \geq 0, \forall x \in X$
(mas-2) $T(t + s)x \leq T(t)(T(s)x)), \forall t, s \geq 0, \forall x \in X$
(mas-3) $t \leq s, x \leq y \Longrightarrow T(t)x \leq T(s)y.$

It is not hard to see that the construction of our couple (Θ, θ) is possible in this relaxed setting; and the properties above remain true. Note that, such objects are, in a certain sense, "complementary" to the ones introduced by S. Turinici and M. Turinici [49]. A concrete example of monotone almost semigroup is that given as

$$T(t)x = S(\varphi(t))x, \ t \in R_+, x \in X;$$

where $(t, x) \mapsto S(t)x$ is a monotone semigroup and $\varphi : R_+ \to R_+$ is a function with

$$\varphi(0) = 0, \varphi(t + s) \leq \varphi(t) + \varphi(s), \ \forall t, s \geq 0.$$

We shall discuss these facts elsewhere.

Semigroup Pareto Efficiency

In the following, a semigroup version of the Pareto efficient point results over metric/normed structures is given, via introduced concepts.

(A) Let $(X, \leq)$ be a quasi-ordered structure; and $d : X \times X \to R_+ \cup \{\infty\}$ be a generalized metric over X. Given the nonempty part A of X, remember that $z \in A$ is Pareto efficient when

$$A(z, \leq) = \{z\}; \text{i.e.: } z \leq w \in A \text{ implies } z = w;$$

the class of all these will be denoted $\mathrm{Eff}(A; \leq)$. As precise, sufficient conditions for existence of such points are obtainable via monotone Ekeland Variational Principle (EVPm). It is our aim in the following to show that such a procedure is valid as well in a semigroup context. To do this, let the basic regularity conditions upon our data be the standard ones

(h01) $(\leq)$ is self-closed and A is $(\leq)$-closed (modulo d)
(h02) d is $(\leq)$-complete (see above).

For the remaining (specific) ones, we need some conventions. Take a semigroup $(t, x) \mapsto S(t)x$ over X; which in addition is monotone with respect to $(\leq)$; and let θ stand for its attached anti-measure. The relation $(\preceq)$ over X introduced as

$$x_1 \preceq x_2 \text{ iff } x_1 \leq x_2, S(d(x_1, x_2))x_1 \leq x_2$$

is reflexive and transitive; hence a quasi-order. In addition,

($\preceq$) it is coarser than ($\leq$): $x_1 \preceq x_2 \Longrightarrow x_1 \leq x_2$.

Call the point $u \in A$, $(\leq, S)$-*starting* when

(h03) $x_1, x_2 \in A(u, \leq), x_1 \leq x_2 \Longrightarrow S(d(x_1, x_2))x_1 \leq x_2$
(h04) $x \mapsto \varphi(x) := \theta(u, x)$ is bounded above on $A(u, \leq)$.

We are now in position to state our announced Pareto efficient point statement.

Theorem 13 *Let the general conditions above (h01)+(h02) be accepted. Then, for each $(\leq, S)$-starting point $u \in A$, there exists some other point $v \in A$, with*

*(**81-a**) $u \leq v$, $d(u, v) \leq \theta(u, v) - \theta(u, u)$*
*(**81-2**) v is Pareto $(\leq)$-efficient (see above).*

Proof Denote for simplicity $M = A(u, \leq)$. We claim that (EVPm) holds for (M, d) and $(\leq, -\varphi)$; and this will complete the argument. Let x_1, x_2 be points in M with $x_1 \leq x_2$. By the starting condition and anti-triangular property of θ, we have

$$d(x_1, x_2) \leq \theta(x_1, x_2) \leq \varphi(x_2) - \varphi(x_1) \text{ (hence } \varphi(x_1) \leq \varphi(x_2));$$

wherefrom, φ is $(\leq)$-increasing along M. Finally, (h04) yields

$0 \leq \varphi(x) = \theta(u, x) < \infty$, $\forall x \in M$.

Summing up, (EVPm) is indeed applicable here; and, from its conclusions, we are done.

(B) Let $(X, ||.||)$ be a (real) Banach space. Take some convex cone H of X; with, in addition,

H is closed (hence, Archimedean).

As usual, denote by $(\leq)$ the associated quasi-order. Fix $k^0 \in H \setminus (-H)$ and put

$H(k^0) = \{x \in H; k^0||x|| \leq x\}$;
or, equivalently: $H(k^0) = \{x \in H; ||x|| \leq \gamma(x)\}$;

where γ is the gauge function attached to (H, k^0). This is a closed convex cone of H which "approximates" H when $||k^0||$ approaches zero; precisely

$$H(\varepsilon k^0) \text{ ``tends'' to } H \text{ as } \varepsilon \to 0.$$

So, the efficient points relative to the convex cone $K := H(k^0)$ may be viewed as a good approximation for the efficient points relative to H, when $||k^0||$ is small enough. As we shall see, the problem of determining such points may be solved by the Pareto efficient point results over normed structures.

Let $(\preceq)$ stand for the quasi-order attached to K. Further, take some nonempty part A of X. The general conditions above are assumed to hold, in our normed setting; that is (via K=closed)

(h05) A is K-closed (modulo $||.||$)
(h06) $||.||$ is K-complete (see above).

And the specific ones are to be written in terms of the gauge function $\psi := \gamma$. In this direction, note that

(subord) γ is subordinated to $(||.||; K)$ (by definition)
(s-ad) γ is super-additive on K (cf. the properties of gauge function).

So, the only condition to be added writes

(h07) $u \in A$ is *admissible* (modulo γ):
$x \mapsto \varphi(x) := \gamma(x - u)$ is bounded above on $A \cap (u + K)$.

The following Pareto efficient point result is then available.

Theorem 14 *Let the precise conditions be in use. Then, for each admissible (modulo γ) $u \in A$, there exists a Pareto K-efficient point $v = v(u) \in A$ with $u \preceq v$.*

Proof By the remarks above, Theorem 10 is applicable to these data. In this case, given $u \in A$ as before, there must be some $v \in \mathrm{Eff}(A, K)$, with

$$u \preceq v, \ ||u - v|| \leq \gamma(v - u).$$

This gives us all desired facts.

Some remarks are in order. A first proof of Theorem 14 was obtained by Goepfert and Tammer [19], through an iterative procedure. Further, Isac [27] provided another proof of the same, by means of a general constructive test. Note that, in both cases, it was assumed that $k^0 \in \mathrm{int}(H)$; the present approach shows that this may be replaced by the weaker condition $k^0 \in H \setminus (-H)$. Finally, note that an algebraic counterpart of Theorem 14 is obtainable under the methods in Turinici [46]; we do not give details. Further aspects may be found in Németh [37].

References

1. M. Altman, A generalization of the Brezis-Browder principle on ordered sets. Nonlinear Anal. **6**, 157–165 (1982)
2. J.S. Bae, E.W. Cho, S.H. Yeom, A generalization of the Caristi-Kirk fixed point theorem and its applications to mapping theorems. J. Korean Math. Soc. **31**, 29–48 (1994)
3. P. Bernays, A system of axiomatic set theory: Part III. Infinity and enumerability analysis. J. Symb. Log. **7**, 65–89 (1942)
4. C.E. Blair, The Baire category theorem implies the principle of dependent choice. Bull. Acad. Pol. Sci. (Sér. Math. Astronom. Phys.) **10**, 933–934 (1977)
5. N. Bourbaki, *General Topology (Chapters 1–4)* (Springer, Berlin, 1989)
6. H. Brezis, F.E. Browder, A general principle on ordered sets in nonlinear functional analysis. Adv. Math. **21**, 355–364 (1976)
7. A. Brøndsted, On a lemma of Bishop and Phelps. Pac. J. Math. **55**, 335–341 (1974)
8. A. Brøndsted, Fixed points and partial orders. Proc. Am. Math. Soc. **60**, 365–366 (1976)

9. N. Brunner, Topologische Maximalprinzipien. Z. Math. Logik Grundl. Math. **33**, 135–139 (1987)
10. O. Cârjă, M. Necula, I.I. Vrabie, *Viability, Invariance and Applications.* North Holland Mathematics Studies, vol. 207 (Elsevier B. V., Amsterdam, 2007)
11. P.J. Cohen, *Set Theory and the Continuum Hypothesis* (Benjamin, New York, 1966)
12. V. Conserva, S. Rizzo, Maximal elements in a class of order complete metric subspaces. Math. Jpn. **37**, 515–518 (1992)
13. R. Cristescu, *Topological Vector Spaces* (Noordhoff International Publishers, Leyden, 1977)
14. S. Dancs, M. Hegedus, P. Medvegyev, A general ordering and fixed-point principle in complete metric space. Acta Sci. Math. (Szeged) **46**, 381–388 (1983)
15. J. Dodu, M. Morillon, The Hahn-Banach property and the Axiom of Choice. Math. Log. Q. **45**, 299–314 (1999)
16. I. Ekeland, Nonconvex minimization problems. Bull. Am. Math. Soc. (New Series) **1**, 443–474 (1979)
17. J.X. Fang, The variational principle and fixed point theorems in certain topological spaces. J. Math. Anal. Appl. **202**, 398–412 (1996)
18. C. Gerth (Tammer), P. Weidner, Nonconvex separation theorems and some applications in vector optimization. J. Optim. Theory Appl. **67**, 297–320 (1990)
19. A. Goepfert, C. Tammer, A new maximal point theorem. Z. Anal. Anwend. (J. Analysis Appl.) **14**, 379–390 (1995)
20. A. Goepfert, C. Tammer, C. Zălinescu, On the vectorial Ekeland's variational principle and minimal points in product spaces. Nonlinear Anal. **39**, 909–922 (2000)
21. A. Goepfert, H. Riahi, C. Tammer, C. Zălinescu, *Variational Methods in Partially Ordered Spaces.* Canadian Mathematical Society Books in Mathematics, vol. 17 (Springer, New York, 2003)
22. R. Goldblatt, On the role of the Baire Category theorem and Dependent Choice in the foundation of logic. J. Symb. Log. **50**, 412–422 (1985)
23. A. Granas, C.D. Horvath, On the order-theoretic Cantor theorem. Taiwan. J. Math. **4**, 203–213 (2000)
24. P.R. Halmos, *Measure Theory* (Springer, New York, 1974)
25. A. Hamel, Equivalents to Ekeland's variational principle in uniform spaces. Nonlinear Anal. **62**, 913–924 (2005)
26. D.H. Hyers, G. Isac, T.M. Rassias, *Topics in Nonlinear Analysis and Applications* (World Scientific Publishing, Singapore, 1997)
27. G. Isac, On Pareto efficiency. A general constructive existence principle, in *Combinatorial and Global Optimization*, ed. by P.M. Pardalos et al. (World Scientific Publishing, Singapore, 2002), pp. 133–144
28. G. Isac, C. Tammer, Nuclear and full nuclear cones in product spaces: Pareto efficiency and an Ekeland type variational principle. Positivity **14**, 1–28 (2004)
29. C.F.K. Jung, On generalized complete metric spaces. Bull. Am. Math. Soc. **75**, 113–116 (1969)
30. O. Kada, T. Suzuki, W. Takahashi, Nonconvex minimization theorems and fixed point theorems in complete metric spaces. Math. Jpn. **44**, 381–391 (1996)
31. B.G. Kang, S. Park, On generalized ordering principles in nonlinear analysis. Nonlinear Anal. **14**, 159–165 (1990)
32. S. Kasahara, On some generalizations of the Banach contraction theorem. Publ. Res. Inst. Math. Sci. Kyoto Univ. **12**, 427–437 (1976)
33. W.A.J. Luxemburg, On the convergence of successive approximations in the theory of ordinary differential equations (II). Ind. Math. **20**, 540–546 (1958)
34. G.H. Moore, *Zermelo's Axiom of Choice: Its Origin, Development and Influence* (Springer, New York, 1982)
35. Y. Moskhovakis, *Notes on Set Theory* (Springer, New York, 2006)
36. L. Nachbin, *Topology and Order* (D. van Nostrand, Princeton, 1965)
37. A.B. Németh, Between Pareto efficiency *and Pareto ε-efficiency*. Optimization **20**, 615–637 (1989)

38. T. Precupanu, *Linear Topological Spaces and Fundamentals of Convex Analysis (Romanian)* (Editura Academiei Române, Bucureşti, 1992)
39. E. Schechter, *Handbook of Analysis and its Foundation* (Academic Press, New York, 1997)
40. A. Szaz, An improved Altman type generalization of the Brezis Browder ordering principle. Math. Commun. **12**, 155–161 (2007)
41. A. Tarski, Axiomatic and algebraic aspects of two theorems on sums of cardinals. Fund. Math. **35**, 79–104 (1948)
42. D. Tataru, Viscosity solutions of Hamilton-Jacobi equations with unbounded nonlinear terms. J. Math. Anal. Appl. **163**, 345–392 (1992)
43. M. Turinici, A generalization of Brezis-Browder's ordering principle. An. Şt. Univ. "A. I. Cuza" Iaşi (S I-a: Mat) **28**, 11–16 (1982)
44. M. Turinici, Metric variants of the Brezis-Browder ordering principle. Demonstratio Math. **22**, 213–228 (1989)
45. M. Turinici, A monotone version of the variational Ekeland's principle. An. Şt. Univ. "A. I. Cuza" Iaşi (S. I-a: Mat) **36**, 329–352 (1990)
46. M. Turinici, Minimal points in product spaces. An. Şt. Univ. "Ovidius" Constanţa (Ser. Math.) **10**, 109–122 (2002)
47. M. Turinici, Projective maximal principles in general vector spaces. Libertas Math. **29**, 25–36 (2008)
48. M. Turinici, Brezis-Browder principle and Dependent Choice. An Şt. Univ. "Al. I. Cuza" Iaşi (Mat.) **57**, 263–277 (2011)
49. S. Turinici, M. Turinici, Projective metrics on abstract ordered sets. Mathematica (Cluj) **34**(57), 81–88 (1992)
50. E.S. Wolk, On the principle of dependent choices and some forms of Zorn's lemma. Can. Math. Bull. **26**, 365–367 (1983)
51. J. Zhu, S.J. Li, Generalization of ordering principles and applications. J. Optim. Theory Appl. **132**, 493–507 (2007)

New Two-Slope Parameterized Achievement Scalarizing Functions for Nonlinear Multiobjective Optimization

Outi Wilppu, Marko M. Mäkelä, and Yury Nikulin

Abstract Most of the methods for multiobjective optimization utilize some scalarization technique where several goals of the original multiobjective problem are converted into a single-objective problem. One common scalarization technique is to use the achievement scalarizing functions. In this paper, we introduce a new family of two-slope parameterized achievement scalarizing functions for multiobjective optimization. This family generalizes both parametrized ASF and two-slope ASF. With these two-slope parameterized ASF, we can guarantee (weak) Pareto optimality of the solutions produced, and every (weakly) Pareto optimal solution can be obtained. The parameterization of this kind gives a systematic way to produce different solutions from the same preference information. With two weighting vectors depending on the achievability of the reference point, there is no need for any assumptions about the reference point. In addition to theory, we give graphical illustrations of two-slope parameterized ASF and analyze sparsity of the solutions produced in convex and nonconvex testproblems.

Keywords Achievement scalarizing functions • Multiobjective optimization • Parameterization • Pareto optimal solutions

Mathematics Subject Classification 90C29, 65K05, 49M37

Introduction

In many applications, the aim is to optimize several objectives and to find a solution which is as good as possible for every objective at the same time. Usually, these objectives are conflicting, and due to that it is not possible to find a solution being optimal for every objective simultaneously. That is why compromises between these

O. Wilppu (✉) • M.M. Mäkelä • Y. Nikulin
Department of Mathematics and Statistics, University of Turku, FI-20014 Turku, Finland
e-mail: omwilp@utu.fi

N.J. Daras, T.M. Rassias (eds.), *Operations Research, Engineering, and Cyber Security*, Springer Optimization and Its Applications 113,
DOI 10.1007/978-3-319-51500-7_18

conflicting objectives are needed. The compromise is optimal if none objective can be improved without impairing at least one of the other objectives. The problem of this kind is called a multiobjective optimization problem, and its optimal solution is called Pareto optimal.

Usually, there are several mathematically equally good Pareto optimal solutions, and someone needs to choose the best solution for a particular problem. This person is called a decision maker. (S)he has an insight into the problem, and it is also possible to obtain some additional information of the problem from the him/her.

As it was said, the problem setting of the multiobjective optimization problem differs a lot from the single-objective optimization. Solving only one objective of the multiobjective optimization problem with a single-objective method can lead to an arbitrary bad solution with respect to other objectives for the original multiobjective problem. Thus, different methods are needed in order to solve multiobjective problems. Several methods are described in [2, 10, 17] and the references therein. Most of the methods for the multiobjective optimization utilize a scalarization. In scalarization, several goals of the original multiobjective problem are converted into a single-objective, and then some suitable single-objective method is applied. Several scalarization techniques are introduced and compared in [11].

One of the most common scalarization techniques is the use of achievement scalarizing functions [4, 5, 10, 11, 15, 16, 18, 19]. In this approach, a reference point is asked from the decision maker. After that, an achievement scalarizing function is optimized in order to find a solution being the closest to the reference point. In [5], an approach is proposed to generate the set of equivalent reference points producing the same solutions, which can be used to assist the decision maker to select the reference point.

Chebyshev type achievement scalarizing functions [19] are one of the most popular achievement scalarizing functions. In Fig. 1, one example is given to illustrate the Chebyshev type achievement scalarizing function in the objective space with two different reference points. If the reference point is unachievable, the distance from the reference point to the feasible objective region is minimized. In Fig. 1a, the right-angled contours are increasing from the unachievable reference point towards the feasible objective region. The optimal solution is the first point from the feasible objective region touching the contour. On the other hand, if the reference point is achievable, the maximum value of the negative difference between the reference point and the nondominated set (i.e., the set of Pareto optimal solutions in objective space) is minimized. In Fig. 1b, the optimal solution for scalarized problem is the nondominated point touching the contour last.

The wide usage of Chebyshev type achievement scalarizing function is due to its good mathematical properties. With this L_∞ metric, any weakly Pareto optimal solution can be obtained. In addition, other type of metrics can be used: for example, linear L_1 metric. But unlike Chebyshev metric, with L_1 metric not every weakly Pareto optimal solution is necessarily obtained in the nonconvex case since there might exist nonsupported solutions. To overcome this drawback, in [16] L_1 based metric is presented to ensure that every weakly Pareto optimal solution can be obtained.

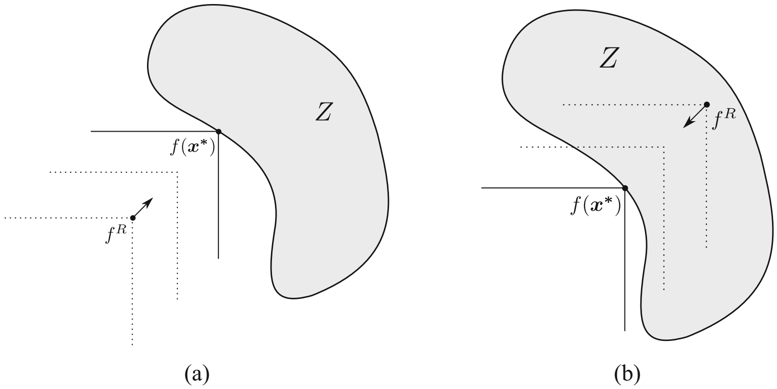


Fig. 1 Graphical illustration of Chebyshev type achievement scalarizing function. (**a**) Unachievable reference point. (**b**) Achievable reference point

In this paper, we propose a new family of two-slope parameterized achievement scalarizing functions (TSPASF). These functions are based on the parameterized achievement scalarizing functions (PASF) introduced in [15]. By using parameterization, it is possible to utilize metrics varying in different combinations from Chebyshev metric to the linear metric. We generalize the PASF by utilizing the idea of two different weighting vectors depending on the achievability of the reference point described in [4]. Thus, the new TSPASF is a generalization of both PASF and two-slope ASF. In the case of the unachievable reference point, we get back to the PASF. In the case of the linear metric, TSPASF is reduced to two-slope ASF. The advantage of this new TSPASF is that any Pareto optimal solution can be found by moving the reference point or changing the weighting vectors. Another advantage compared with PASF is that we need neither assume anything about the reference point nor test reference point achievability. This occurs since the formulation of the problem guarantees that the right weighting vector is used in every case.

This paper is organized as follows: In section “Preliminaries”, we recall some basic results of multiobjective optimization and describe the ideas of ASF and PASF. Section “Two-Slope Parameterized Achievement Scalarizing Functions” is dedicated to a new TSPASF, and a special case of the multiobjective problem with three objectives is analyzed and illustrated in section “Case of Three Objectives”. In section “Computational Experiments”, we give some numerical examples in convex and nonconvex case. Section “Conclusion” contains some final remarks.

Preliminaries

We consider a multiobjective optimization problem where all the objectives are minimized simultaneously. This problem is of the form

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \\ \text{s. t.} \quad & \boldsymbol{x} \in X, \end{aligned} \tag{1}$$

where the partial objective functions are defined $f_i : X \to \mathbb{R}$, $i \in \mathbb{N}_m = \{1, \ldots, m\}$ and they are assumed to be lower semicontinuous. A set $X \subset \mathbb{R}^n$ is a nonempty compact set of feasible solutions. The image of this set X is called a feasible objective region $Z = f(X)$. The objective functions are also assumed to be conflicting. Thus, it is impossible to have a solution being minimal for every objective function.

We recall some basic results from multiobjective optimization. For more details we refer to [2, 10]. In the following we use notation $\boldsymbol{x} < \boldsymbol{y}$ if $x_i < y_i$ for all $i \in \mathbb{N}_n$ and notation $\boldsymbol{x} \leq \boldsymbol{y}$ if $x_i \leq y_i$ for all $i \in \mathbb{N}_n$.

A solution of problem (1) is called Pareto optimal if none objective can be improved without deteriorating some other objective at the same time. We can formally define that a solution $\boldsymbol{x}^* \in X$ of the problem (1) is *Pareto optimal* if there exists no point $\boldsymbol{x} \in X$ such that $f_i(\boldsymbol{x}) \leq f_i(\boldsymbol{x}^*)$ for all $i \in \mathbb{N}_m$ and $f_j(\boldsymbol{x}) < f_j(\boldsymbol{x}^*)$ for at least one index $j \in \mathbb{N}_m$. Under the assumptions of problem (1), Pareto optimal solutions exist [17]. Usually, there exist many mathematically equally good Pareto optimal solutions and a set of these Pareto optimal solutions is called the *Pareto set*.

We can also define a generalized concept where a solution $\boldsymbol{x}^* \in X$ is called *weakly Pareto optimal* if there exists no another point $\boldsymbol{x} \in X$ such that $f_i(\boldsymbol{x}) < f_i(\boldsymbol{x}^*)$ for all $i \in \mathbb{N}_m$. Note that the set of Pareto optimal solutions is a subset of the set of weakly Pareto optimal solutions.

To get some information about Pareto optimal solutions, an ideal and a nadir vector, f^I and f^N, can be calculated giving a lower and an upper bound for the range of Pareto optimal solutions, respectively. The components of the *ideal vector* are obtained by minimizing every objective separately. Thus, the i:th component of the ideal vector can be defined by solving the problem $\min_{\boldsymbol{x} \in X} f_i(\boldsymbol{x})$. The ideal vector tells how good solutions can be found, but normally the ideal vector is not feasible. If the ideal vector is feasible, then it will clearly be also an optimal solution of problem (1).

The *nadir vector* relates the upper bound for Pareto optimal solutions representing the worst solution. The components of the nadir vector can be calculated by maximizing objectives over the set of Pareto optimal solutions. Due to this optimization over the Pareto set, it is usually difficult to obtain the nadir vector, but it can be approximated [1, 2, 10].

A *utopian vector* gives a strictly better solution than any of the Pareto optimal solutions and even better than the ideal vector. The components of the utopian vector are of the form $f_i^U = f_i^I - \varepsilon_i$ where $\varepsilon_i > 0$ is a sufficient small constant.

A point consisting of desirable values for objective functions is called a *reference point* $f^R = (f_i^R, \ldots, f_m^R)$. These desirable values have been provided by the decision maker who tells what (s)he wishes to achieve. The reference point is said to be *achievable* if $f^R \in Z + \mathbb{R}^m_+$ where $\mathbb{R}^m_+ = \{\boldsymbol{y} \in \mathbb{R}^m \,|\, y_i \geq 0 \text{ for } i \in \mathbb{N}_m\}$. Otherwise the reference point is said to be *unachievable*.

In this paper, we are focusing on achievement scalarizing functions (ASF) [18, 19] in order to scalarize multiobjective problem (1). This scalarized problem is of the form

$$\min_{\boldsymbol{x}\in X} s_R(f(\boldsymbol{x}), \boldsymbol{\lambda}). \tag{2}$$

One example of achievement scalarizing functions is Chebyshev type

$$s_R(f(\boldsymbol{x}), \boldsymbol{\lambda}) = \max_{i\in\mathbb{N}_m} \left\{\lambda_i(f_i(\boldsymbol{x}) - f_i^R)\right\}, \tag{3}$$

where the vector f^R is a reference point and the value $\lambda_i > 0$ is a weighting coefficient for the objective function f_i specifying the direction of the projection from the reference point to the Pareto frontier.

If the reference point is unachievable, then the ASF is minimizing the distance from the reference point to the feasible set. On the other hand, if the reference point is achievable, we are minimizing the maximum value of the negative difference between the reference point and the nondominated set. By moving the reference point or manipulating $\boldsymbol{\lambda}$, any (weakly) Pareto optimal solution can be obtained [10].

In order to guarantee that problem (2) generates Pareto optimal solutions, the following properties of the ASF can be described.

Definition 1 ([21]) An achievement scalarizing function $s_R : \mathbb{R}^m \times \mathbb{R}^m_+ \to \mathbb{R}$ is said to be

1. *increasing* if for any $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^m, \boldsymbol{y}_1 \leq \boldsymbol{y}_2$, then $s_R(\boldsymbol{y}_1, \boldsymbol{\lambda}) \leq s_R(\boldsymbol{y}_2, \boldsymbol{\lambda})$.
2. *strictly increasing* if for any $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^m, \boldsymbol{y}_1 < \boldsymbol{y}_2$, then $s_R(\boldsymbol{y}_1, \boldsymbol{\lambda}) < s_R(\boldsymbol{y}_2, \boldsymbol{\lambda})$.
3. *strongly increasing* if for any $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^m, \boldsymbol{y}_1 \leq \boldsymbol{y}_2$ and $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$, then $s_R(\boldsymbol{y}_1, \boldsymbol{\lambda}) < s_R(\boldsymbol{y}_2, \boldsymbol{\lambda})$.

Note that any strongly increasing ASF is also strictly increasing and any strictly increasing ASF is also increasing. For example, a function of Chebyshev type (3) is strictly increasing but not strongly increasing.

The following two theorems specify necessary and sufficient conditions to (weak) Pareto optimality:

Theorem 1 ([20, 21]) *The following two statements are true:*

1. *Let s_R be strongly (strictly) increasing. If $\boldsymbol{x}^* \in X$ is an optimal solution of problem* (2)*, then $\boldsymbol{x}^*$ is (weakly) Pareto optimal for problem* (1)*.*
2. *If s_R is increasing and the solution $\boldsymbol{x}^* \in X$ of problem* (2) *is unique, then $\boldsymbol{x}^*$ is Pareto optimal for problem* (1)*.*

Theorem 2 ([10]) *If s_R is strictly increasing and $\boldsymbol{x}^* \in X$ is weakly Pareto optimal solution for problem* (1)*, then it is a solution of problem* (2) *with $f^R = f(\boldsymbol{x}^*)$ and the optimal value of s_R is zero for any weight vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$, $\lambda_i > 0$ for all $i \in \mathbb{N}_m$.*

A starting point for this paper is a parameterized achievement scalarizing function developed in [15] extending the ideas of an additive achievement scalarizing function introduced in [16]. Let I^q be a subset of $\mathbb{N}_m$ of cardinality q. The parameterized achievement scalarizing function (PASF) [15] is of the form

$$\tilde{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}) = \max_{I^q \subseteq \mathbb{N}_m : |I^q| = q} \left\{ \sum_{i \in I^q} \max[\lambda_i (f_i(\boldsymbol{x}) - f_i^R), 0] \right\},$$

where $q \in \mathbb{N}_m$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$, $\lambda_i > 0$, $i \in \mathbb{N}_m$. The problem to be solved is then

$$\min_{\boldsymbol{x} \in X} \tilde{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}). \tag{4}$$

Due to the formation of PASF, the value of $\tilde{s}_R^q$ is always nonnegative. With different values of the parameter q, different metrics varying in different combinations between L_1 to L_∞ metrics are obtained. Extreme cases are L_1 metric with $q = m$, where m is the number of objectives, and L_∞ metric (3) with $q = 1$.

The following two properties were proven for $\tilde{s}_R^q$ in [15].

Theorem 3 ([15]) *Given problem (4), let f^R be a reference point such that there exists no feasible solution whose image strictly dominates f^R and $\lambda_i > 0$ for all $i \in \mathbb{N}_m$. Then any optimal solution of problem (4) is a weakly Pareto optimal solution for problem (1).*

Theorem 4 ([15]) *Given problem (4), let f^R be any reference point and $\lambda_i > 0$ for all $i \in \mathbb{N}_m$. Then, among the optimal solutions of problem (4) there exists at least one Pareto optimal solution for problem (1).*

Theorem 4 implies that if $\boldsymbol{x}^*$ is a unique solution of problem (4), then it is a Pareto optimal solution for problem (1).

With the PASF, several Pareto optimal solutions can be found by moving the reference point or by manipulating the weighting coefficients meanwhile the reference point stays fixed. A limitation of the PASF is that the reference point should not be strictly dominated by some feasible point. With the two-slope parameterized ASF described in the next section, this limitation can be forgot.

Two-Slope Parameterized Achievement Scalarizing Functions

Next we introduce a new extended parameterized ASF taking achievability of the reference point into account as in [4]. A two-slope parameterized ASF (TSPASF) is defined as follows:

$$\hat{s}^q_R(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) = \max_{I^q \subseteq \mathbb{N}_m:\, |I^q|=q} \left\{ \sum_{i \in I^q} \Big[\max\left\{\lambda^U_i (f_i(\boldsymbol{x}) - f^R_i), 0\right\} + \min\left\{\lambda^A_i (f_i(\boldsymbol{x}) - f^R_i), 0\right\} \Big] \right\}, \tag{5}$$

where $q \in \mathbb{N}_m$, $\boldsymbol{\lambda}^U = (\lambda^U_1, \dots, \lambda^U_m)$ and $\boldsymbol{\lambda}^A = (\lambda^A_1, \dots, \lambda^A_m)$, $\lambda^U_i, \lambda^A_i > 0$, $i \in \mathbb{N}_m$. Either of these two different weighting vectors $\boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A$ is used depending on whether the reference point is unachievable or achievable, respectively. In the following, we denote

$$J^q = \{I^q \subseteq \mathbb{N}_m \mid |I^q| = q\}.$$

The problem to be solved is then

$$\min_{\boldsymbol{x} \in X} \hat{s}^q_R(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A). \tag{6}$$

Notice that if $q = 1$, then $\hat{s}^q_R$ has the same formation than two-slope ASF proposed in [4].

For the TSPASF, it is possible to prove similar result as Theorem 3. Note that the assumption of nonexisting feasible solution which image strictly dominates f^R is not needed.

Theorem 5 *Given problem* (6)*, let* $\lambda^U_i, \lambda^A_i > 0$ *for all* $i \in \mathbb{N}_m$*. Then any optimal solution of problem* (6) *is a weakly Pareto optimal solution for problem* (1)*.*

Proof Let $\boldsymbol{x}^*$ be an optimal solution of problem (6). Assume that $\boldsymbol{x}^*$ is not weakly Pareto optimal. Then there exists a feasible solution $\boldsymbol{x}' \in X$ such that $f_i(\boldsymbol{x}') < f_i(\boldsymbol{x}^*)$ for all $i \in \mathbb{N}_m$.

For any $\boldsymbol{x} \in X$, denote $I_{\boldsymbol{x}} = \left\{i \in \mathbb{N}_m \mid f^R_i \leq f_i(\boldsymbol{x})\right\}$ and $J_{\boldsymbol{x}} = \{i \in \mathbb{N}_m \mid f^R_i > f_i(\boldsymbol{x})\}$. Since $I_{\boldsymbol{x}'} \subseteq I_{\boldsymbol{x}^*}$ and $J_{\boldsymbol{x}'} \supseteq J_{\boldsymbol{x}^*}$, we obtain

$$\begin{aligned}
&\hat{s}^q_R(f(\boldsymbol{x}'), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) \\
&= \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{\lambda^U_i (f_i(\boldsymbol{x}') - f^R_i), 0\right\} + \min\left\{\lambda^A_i (f_i(\boldsymbol{x}') - f^R_i), 0\right\}\right] \right\} \\
&= \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{\boldsymbol{x}'}} \lambda^U_i (f_i(\boldsymbol{x}') - f^R_i) + \sum_{i \in I^q \bigcap J_{\boldsymbol{x}'}} \lambda^A_i (f_i(\boldsymbol{x}') - f^R_i) \right\} \\
&< \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{\boldsymbol{x}'}} \lambda^U_i (f_i(\boldsymbol{x}^*) - f^R_i) + \sum_{i \in I^q \bigcap J_{\boldsymbol{x}'}} \lambda^A_i (f_i(\boldsymbol{x}^*) - f^R_i) \right\}
\end{aligned}$$

$$\begin{aligned}
&\leq \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{x^*}} \lambda_i^U (f_i(\boldsymbol{x}^*) - f_i^R) + \sum_{i \in I^q \bigcap J_{x^*}} \lambda_i^A (f_i(\boldsymbol{x}^*) - f_i^R) \right\} \\
&= \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}^*) - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}^*) - f_i^R), 0 \right\} \right] \right\} \\
&= \hat{s}_R^q (f(\boldsymbol{x}^*), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A).
\end{aligned}$$

Inequality $\hat{s}_R^q(f(\boldsymbol{x}'), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) < \hat{s}_R^q(f(\boldsymbol{x}^*), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$ contradicts the assumption of $\boldsymbol{x}^*$ being an optimal solution of problem (6). This implies that $\boldsymbol{x}^*$ is weakly Pareto optimal. □

In addition, the following result similar to Theorem 4 can be proven for the TSPASF.

Theorem 6 *Given problem* (6)*, let* $\lambda_i^U, \lambda_i^A > 0$ *for all* $i \in \mathbb{N}_m$*. Then, among the optimal solutions of problem* (6) *there exists at least one Pareto optimal solution for problem* (1)*.*

Proof Let $\boldsymbol{x}^*$ be an optimal solution of problem (6) but not a Pareto optimal solution of problem (1). Then, according to the definition of Pareto optimality there exists $\boldsymbol{x}' \in X$ such that $f_i(\boldsymbol{x}') \leq f_i(\boldsymbol{x}^*)$ for all $i \in \mathbb{N}_m$ and $f_j(\boldsymbol{x}') < f_j(\boldsymbol{x}^*)$ for at least one index $j \in \mathbb{N}_m$. Now

$$\begin{aligned}
&\hat{s}_R^q (f(\boldsymbol{x}'), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) \\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}') - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}') - f_i^R), 0 \right\} \right] \right\} \\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{x'}} \lambda_i^U (f_i(\boldsymbol{x}') - f_i^R) + \sum_{i \in I^q \bigcap J_{x'}} \lambda_i^A (f_i(\boldsymbol{x}') - f_i^R) \right\} \\
&\quad \leq \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{x^*}} \lambda_i^U (f_i(\boldsymbol{x}^*) - f_i^R) + \sum_{i \in I^q \bigcap J_{x^*}} \lambda_i^A (f_i(\boldsymbol{x}^*) - f_i^R) \right\} \qquad (7) \\
&\quad = \hat{s}_R^q (f(\boldsymbol{x}^*), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A).
\end{aligned}$$

This completes the proof since if the inequality (7) is strict this contradicts the assumption that $\boldsymbol{x}^*$ is an optimal solution for problem (6). If equality (7) holds, then $\boldsymbol{x}'$ is an optimal solution for problem (6) and Pareto optimal for problem (1). □

Theorem 6 implies the following corollary.

Corollary 1 *If an optimal solution of problem* (6) *is unique, then it is a Pareto optimal solution for problem* (1) *for any* $\lambda_i^U, \lambda_i^A > 0$, $i \in \mathbb{N}_m$.

Corollary 1 can be proven also in a different way. According to Theorem 1, if $\hat{s}_R^q$ is increasing and the solution $\boldsymbol{x}^* \in X$ of problem (6) is unique, then $\boldsymbol{x}^*$ is Pareto optimal.

Now $\hat{s}_R^q$ is increasing according to Definition 1, since take $\boldsymbol{x}_1 \in X$ and $\boldsymbol{x}_2 \in X$ with $f_i(\boldsymbol{x}_1) < f_i(\boldsymbol{x}_2)$ for all $i \in \mathbb{N}_m$ and $\lambda_i^U, \lambda_i^A > 0$ for all $i \in \mathbb{N}_m$. Then

$$\begin{aligned}
&\hat{s}_R^q(f(\boldsymbol{x}_1), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)\\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}_1) - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}_1) - f_i^R), 0 \right\} \right] \right\}\\
&\quad \le \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}_2) - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}_2) - f_i^R), 0 \right\} \right] \right\}\\
&\quad = \hat{s}_R^q(f(\boldsymbol{x}_2), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A).
\end{aligned}$$

The following result guarantees that with TSPASF it is possible to obtain every weakly Pareto optimal solution.

Theorem 7 *If* $\boldsymbol{x}^*$ *is weakly Pareto optimal for problem* (1)*, then it is a solution of problem* (6) *with* $f^R = f(\boldsymbol{x}^*)$*, and optimal value is zero for all* $\lambda_i^U, \lambda_i^A > 0$, $i \in \mathbb{N}_m$.

Proof Theorem 2 implies this theorem if $\hat{s}_R^q$ is strictly increasing. Now we prove that $\hat{s}_R^q$ indeed is strictly increasing.

Take $\boldsymbol{x}_1 \in X$ and $\boldsymbol{x}_2 \in X$ with $f_i(\boldsymbol{x}_1) < f_i(\boldsymbol{x}_2)$ for all $i \in \mathbb{N}_m$. Since $\lambda_i^U, \lambda_i^A > 0$ for all $i \in \mathbb{N}_m$, $I_{\boldsymbol{x}_1} \subseteq I_{\boldsymbol{x}_2}$, and $J_{\boldsymbol{x}_1} \supseteq J_{\boldsymbol{x}_2}$, we obtain

$$\begin{aligned}
&\hat{s}_R^q(f(\boldsymbol{x}_1), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)\\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}_1) - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}_1) - f_i^R), 0 \right\} \right] \right\}\\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{\boldsymbol{x}_1}} \lambda_i^U (f_i(\boldsymbol{x}_1) - f_i^R) + \sum_{i \in I^q \bigcap J_{\boldsymbol{x}_1}} \lambda_i^A (f_i(\boldsymbol{x}_1) - f_i^R) \right\}\\
&\quad < \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{\boldsymbol{x}_2}} \lambda_i^U (f_i(\boldsymbol{x}_2) - f_i^R) + \sum_{i \in I^q \bigcap J_{\boldsymbol{x}_2}} \lambda_i^A (f_i(\boldsymbol{x}_2) - f_i^R) \right\}\\
&\quad = \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{ \lambda_i^U (f_i(\boldsymbol{x}_2) - f_i^R), 0 \right\} + \min\left\{ \lambda_i^A (f_i(\boldsymbol{x}_2) - f_i^R), 0 \right\} \right] \right\}\\
&\quad = \hat{s}_R^q(f(\boldsymbol{x}_2), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A).
\end{aligned}$$

□

It is also possible to prove that the convexity of the original objective functions of problem (1) is also preserving in $\hat{s}_R^q$. However, this proof necessitates the following lemma.

Lemma 1 *Let functions f_i be the objective functions of problem* (1) *and sets $I_{\boldsymbol{x}}$ and $J_{\boldsymbol{x}}$ are defined by $I_{\boldsymbol{x}} = \{i \in \mathbb{N}_m \mid f_i^R \leq f_i(\boldsymbol{x})\}$ and $J_{\boldsymbol{x}} = \{i \in \mathbb{N}_m \mid f_i^R > f_i(\boldsymbol{x})\}$. If all the functions f_i are convex and $\boldsymbol{x} \in X$, where X is a convex set, then*

1. $I_{\boldsymbol{x}} \subset (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$
2. $J_{\boldsymbol{x}} \supset (J_{\boldsymbol{x}_1} \bigcap J_{\boldsymbol{x}_2})$,

where $\boldsymbol{x} = \theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2 \in X$, $\theta \in [0, 1]$.

Proof Due to the convexity of the function f_i, the following holds

$$f_i(\boldsymbol{x}) \leq \theta f_i(\boldsymbol{x}_1) + (1-\theta) f_i(\boldsymbol{x}_2) \quad \text{for all } i. \tag{8}$$

In order to proof the first case, consider an index i such that $i \in I_{\boldsymbol{x}}$ and $i \notin (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$. Since $i \in I_{\boldsymbol{x}}$, then $f_i^R \leq f_i(\boldsymbol{x})$. Moreover, since $i \notin (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$, then $f_i^R > f_i(\boldsymbol{x}_1)$ and $f_i^R > f_i(\boldsymbol{x}_2)$. The last implies the following:

$$f_i^R = \theta f_i^R + (1-\theta) f_i^R > \theta f_i(\boldsymbol{x}_1) + (1-\theta) f_i(\boldsymbol{x}_2). \tag{9}$$

Since f_i is convex, inequality (8) holds. From this, inequality (9) and the assumption that $i \in I_{\boldsymbol{x}}$, it follows

$$f_i^R \leq f_i(\boldsymbol{x}) \leq \theta f_i(\boldsymbol{x}_1) + (1-\theta) f_i(\boldsymbol{x}_2) < f_i^R,$$

which contradicts the assumption that $i \notin (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$ and thus $i \in (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$. Same holds for every index of $i \in I_{\boldsymbol{x}}$ and thus $I_{\boldsymbol{x}} \subset (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$.

In the second case, assume that an index $i \in (J_{\boldsymbol{x}_1} \bigcap J_{\boldsymbol{x}_2})$ and thus $f_i^R > f_i(\boldsymbol{x}_1)$ and $f_i^R > f_i(\boldsymbol{x}_2)$. This property and the convexity of f_i imply

$$f_i^R = \theta f_i^R + (1-\theta) f_i^R > \theta f_i(\boldsymbol{x}_1) + (1-\theta) f_i(\boldsymbol{x}_2) \geq f_i(\boldsymbol{x}).$$

Now $f_i^R > f_i(\boldsymbol{x})$ and thus $i \in J_x$. Same holds for every index of $i \in (J_{\boldsymbol{x}_1} \bigcap J_{\boldsymbol{x}_2})$ and thus $J_{\boldsymbol{x}} \supset (J_{\boldsymbol{x}_1} \bigcap J_{\boldsymbol{x}_2})$. □

Theorem 8 *Let X be a convex set, and functions f_i be the objective functions of problem* (1)*. If all the functions f_i are convex, then $\hat{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$ is also convex when $\boldsymbol{x} \in X$.*

Proof According to Lemma 1 we have $I_{\boldsymbol{x}} \subset (I_{\boldsymbol{x}_1} \bigcup I_{\boldsymbol{x}_2})$ and $J_{\boldsymbol{x}} \supset (J_{\boldsymbol{x}_1} \bigcap J_{\boldsymbol{x}_2})$. Thus,

$$\begin{aligned}
&\hat{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)\\
&= \max_{I^q \in J^q} \left\{ \sum_{i \in I^q} \left[\max\left\{\lambda_i^U (f_i(\boldsymbol{x}) - f_i^R), 0\right\} + \min\left\{\lambda_i^A (f_i(\boldsymbol{x}) - f_i^R), 0\right\} \right] \right\}\\
&= \max_{I^q \in J^q} \left\{ \sum_{i \in I^q \bigcap I_{\boldsymbol{x}}} \lambda_i^U (f_i(\boldsymbol{x}) - f_i^R) + \sum_{i \in I^q \bigcap J_{\boldsymbol{x}}} \lambda_i^A (f_i(\boldsymbol{x}) - f_i^R) \right\}.
\end{aligned}$$

Now $\hat{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$ is a maximum of the convex functions and hence $\hat{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$ is also convex function. Thus the convexity is preserved. The proof with more details can be found in [22]. □

Note that in order to guarantee Pareto optimality of the solution produced, we can make the ASF strongly increasing by adding an augmented term [2, 10] to (5), and the following augmented form

$$\hat{s}_R^q + \rho \sum_{i \in \mathbb{N}_m} \lambda_i (f_i(\boldsymbol{x}) - f_i^R), \quad \rho > 0$$

is used in practice.

The advantages of TSPASF are that we always find at least a weakly Pareto optimal solution, and the different weakly Pareto optimal solutions may be obtained by changing the reference point, weighting vectors or the value of the parameter q. Additionally, compared with PASF there are no restrictions for the location of the reference point. Thus, there is no need for any tests of achievability of the reference point since the formulation (5) uses always the right weighting coefficient.

The parameterization used in TSPASF and PASF gives a systematic way to produce possible different (weakly) Pareto optimal solutions from the same preference information with different metrics. The systematic way of this kind may be useful in some interactive methods [10], for example synchronous NIMBUS [12], using several ASF basing on the same preference information. In order to find different (weakly) Pareto optimal solutions, problem (6) can be solved with all values or just some values of the parameter q.

In problem (6), there exists a min-max term. Thus, the problem is nonsmooth even if the objective functions of problem (1) are differentiable. Nonsmooth problems can be solved efficiently with bundle methods [6]. Problem (6) can also be turned into differentiable MINLP form as follows:

$$\begin{aligned}
&\min \ \alpha && (10)\\
&\text{s. t. } \alpha \geq \sum_{i \in I_s^q} \left[\lambda_i^U (1 - z_i^s)(f_i(\boldsymbol{x}) - f_i^R) + z_i^s \lambda_i^A (f_i(\boldsymbol{x}) - f_i^R)\right], \ s = 1, \dots, \binom{m}{q}\\
&\quad f_i^R - f_i(\boldsymbol{x}) \leq z_i^s M, \ i \in I^q, \ s = 1, \dots, \binom{m}{q}\\
&\quad f_i^R - f_i(\boldsymbol{x}) \geq (z_i^s - 1)M, \ i \in I^q, \ s = 1, \dots, \binom{m}{q}\\
&\quad \boldsymbol{x} \in X, \ z_i^s \in \{0, 1\}, \ i \in \mathbb{N}_m, \ s = 1, \dots, \binom{m}{q},
\end{aligned}$$

where s enumerates a q-element subsets I_s^q of an m-element set $\mathbb{N}_m$, z_i^s is a binary variable and M is a sufficiently large number to ensure that $z_i^s = 1$ if and only if $f_i^R - f_i(\boldsymbol{x}) > 0$ and $z_i^s = 0$ if and only if $f_i^R - f_i(\boldsymbol{x}) \leq 0$. Due to the binary variable z_i^s, some mixed-integer programming solver, for example generalized αECP algorithm [3], is needed.

According to Theorem 8, in the case where all the objectives f_i are convex, also $\hat{s}_R^q$ is convex, and thus the global optimum can be found. In general, if the objectives f_i are nonconvex, then problem (6) can be solved with a bundle method, and problem (10) can be solved with αECP algorithm, but only the local optimum can be guaranteed. If the objectives are assumed to be f°-pseudoconvex, then also global optimum can be guaranteed with bundle [8] and αECP [3] method.

Case of Three Objectives

In [4], there is given some graphical illustration of two-slope ASF with two objectives. This illustration is valid also for TSPASF in the case when $q = 1$. In order to illustrate the functioning of TSPASF and to compare it to parameterized ASF, we consider the simplest non-trivial case, in other words the case where $m = 3$. This represents $\hat{s}_R^q$ in the case of three objectives. Now (5) has the form

$$\begin{aligned}
\hat{s}_R^q(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) = \max_{I^q \subseteq \{1,2,3\}: |I^q| = q} \Bigg\{ \sum_{i \in I^q} \Big[&\max\left\{\lambda_i^U (f_i(\boldsymbol{x}) - f_i^R), 0\right\}\\
&+ \min\left\{\lambda_i^A (f_i(\boldsymbol{x}) - f_i^R), 0\right\} \Big] \Bigg\},
\end{aligned}$$

where $q = 1, 2, 3$, $\boldsymbol{\lambda}^U = (\lambda_1^U, \lambda_2^U, \lambda_3^U)$ and $\boldsymbol{\lambda}^A = (\lambda_1^A, \lambda_2^A, \lambda_3^A)$, $\lambda_i^U, \lambda_i^A > 0$, $i \in \mathbb{N}_3$. Now for $q = 1$:

$$\begin{aligned}
&\hat{s}_R^1(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) = \\
&\quad \max\Big\{ \max\left\{\lambda_1^U(f_1(\boldsymbol{x}) - f_1^R), 0\right\} + \min\left\{\lambda_1^A(f_1(\boldsymbol{x}) - f_1^R), 0\right\}; \\
&\quad \max\left\{\lambda_2^U(f_2(\boldsymbol{x}) - f_2^R), 0\right\} + \min\left\{\lambda_2^A(f_2(\boldsymbol{x}) - f_2^R), 0\right\}; \\
&\quad \max\left\{\lambda_3^U(f_3(\boldsymbol{x}) - f_3^R), 0\right\} + \min\left\{\lambda_3^A(f_3(\boldsymbol{x}) - f_3^R), 0\right\} \Big\}
\end{aligned}$$

for $q = 2$:

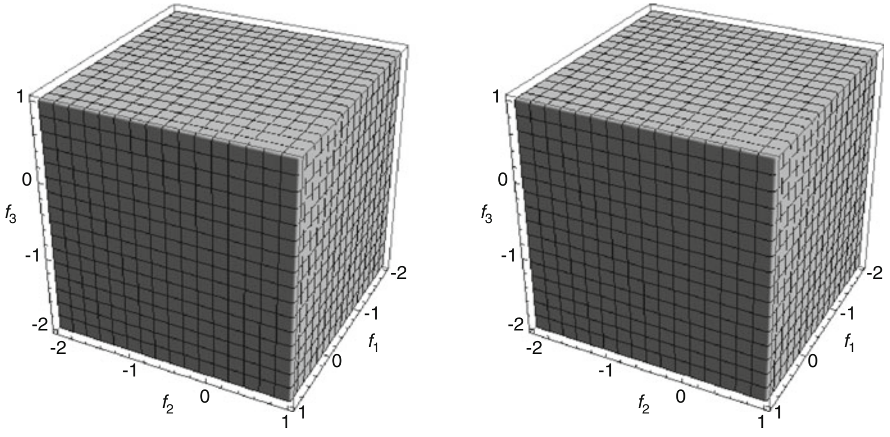
$$\begin{aligned}
&\hat{s}_R^2(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) = \\
&\quad \max\Big\{ \max\left\{\lambda_1^U(f_1(\boldsymbol{x}) - f_1^R), 0\right\} + \min\left\{\lambda_1^A(f_1(\boldsymbol{x}) - f_1^R), 0\right\} \\
&\quad + \max\left\{\lambda_2^U(f_2(\boldsymbol{x}) - f_2^R), 0\right\} + \min\left\{\lambda_2^A(f_2(\boldsymbol{x}) - f_2^R), 0\right\}; \\
&\quad \max\left\{\lambda_1^U(f_1(\boldsymbol{x}) - f_1^R), 0\right\} + \min\left\{\lambda_1^A(f_1(\boldsymbol{x}) - f_1^R), 0\right\} \\
&\quad + \max\left\{\lambda_3^U(f_3(\boldsymbol{x}) - f_3^R), 0\right\} + \min\left\{\lambda_3^A(f_3(\boldsymbol{x}) - f_3^R), 0\right\}; \\
&\quad \max\left\{\lambda_2^U(f_2(\boldsymbol{x}) - f_2^R), 0\right\} + \min\left\{\lambda_2^A(f_2(\boldsymbol{x}) - f_2^R), 0\right\} \\
&\quad + \max\left\{\lambda_3^U(f_3(\boldsymbol{x}) - f_3^R), 0\right\} + \min\left\{\lambda_3^A(f_3(\boldsymbol{x}) - f_3^R), 0\right\} \Big\}
\end{aligned}$$

for $q = 3$:

$$\begin{aligned}
&\hat{s}_R^3(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A) = \\
&\quad \max\left\{\lambda_1^U(f_1(\boldsymbol{x}) - f_1^R), 0\right\} + \min\left\{\lambda_1^A(f_1(\boldsymbol{x}) - f_1^R), 0\right\} \\
&\quad + \max\left\{\lambda_2^U(f_2(\boldsymbol{x}) - f_2^R), 0\right\} + \min\left\{\lambda_2^A(f_2(\boldsymbol{x}) - f_2^R), 0\right\} \\
&\quad + \max\left\{\lambda_3^U(f_3(\boldsymbol{x}) - f_3^R), 0\right\} + \min\left\{\lambda_3^A(f_3(\boldsymbol{x}) - f_3^R), 0\right\}.
\end{aligned}$$

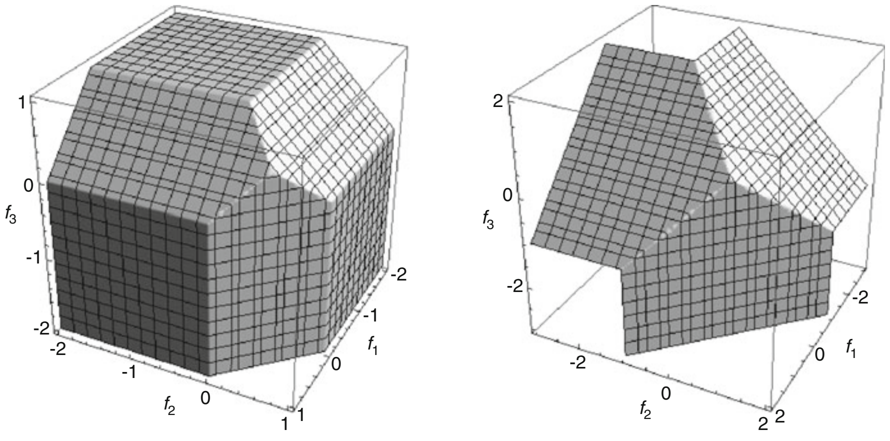
Next we give some graphical illustrations of 1-level sets (i.e., the set of points for which the distance from the reference point equals 1 with respect to the corresponding ASF) in 3-dimensional space for both PASF $\tilde{s}_R^q$ and TSPASF $\hat{s}_R^q$ to see the difference between them. The algebraic form of $\tilde{s}_R^3$ is given in [15]. The view is restricted within a rectangular $\{f = (f_1, f_2, f_3)^T : -2 \le f_i \le 1, i \in N_3\}$, and the reference point is assumed to be $f^R = (0, 0, 0)^T$. All the objective functions are assumed to be identity mappings $f_i(\boldsymbol{x}) = \boldsymbol{x}$, and all the weighting coefficients are equal to one, $\lambda_1^U, \lambda_2^U, \lambda_3^U, \lambda_1^A, \lambda_2^A, \lambda_3^A = 1$. Figures 2a, 3a and 4a show 1-level set of $\tilde{s}_R^1$, $\tilde{s}_R^2$, and $\tilde{s}_R^3$, respectively and Figs. 2b, 3b and 4b show 1-level set of $\hat{s}_R^1$, $\hat{s}_R^2$, and $\hat{s}_R^3$, respectively.

Notice that the choice of the parameter q affects the shape of D-levels. These D-levels may vary from sharp to flat. Those cases where faces are parallel to the



(a) (b)

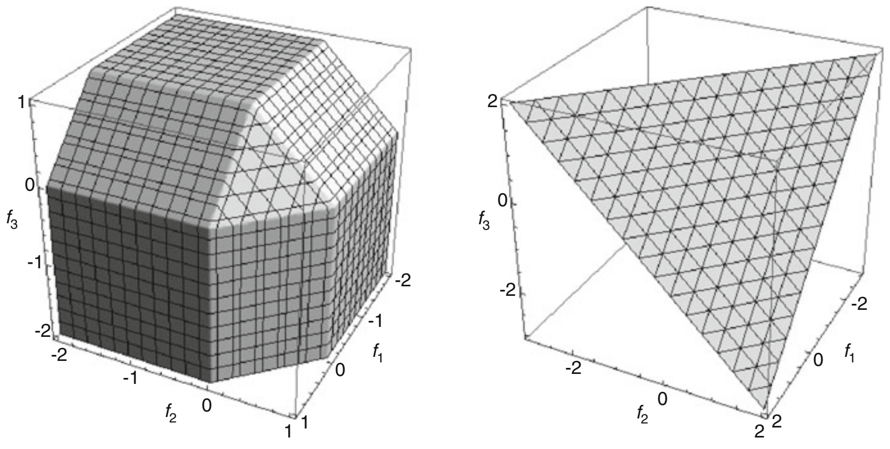
Fig. 2 1-level sets for PASF and TSPASF with $q = 1$. (**a**) 1-level set for $\tilde{s}^1_R(f(\boldsymbol{x}), \boldsymbol{\lambda})$. (**b**) 1-level set for $\hat{s}^1_R(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$



(a) (b)

Fig. 3 1-level sets for PASF and TSPASF with $q = 2$. (**a**) 1-level set for $\tilde{s}^2_R(f(\boldsymbol{x}), \boldsymbol{\lambda})$. (**b**) 1-level set for $\hat{s}^2_R(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$

faces f_1f_2, f_1f_3, or f_2f_3 correspond to the situation then one of the maxima equals to one and other two are less than one or zero. If sum of two maxima equals to one and the third is less than one or zero, it corresponds the case where faces are sloped and parallel to the coordinate rays. If all the three maxima are positive and sum of them equals to one, we have either a flat face (see Fig. 4b) or a triangle pyramid with a top vertex $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in Fig. 3b. These top vertices correspond to those cases when all three maxima are participating.



(a) (b)

Fig. 4 1-level sets for PASF and TSPASF with $q = 3$. (**a**) 1-level set for $\tilde{s}^3_R(f(\boldsymbol{x}), \boldsymbol{\lambda})$. (**b**) 1-level set for $\hat{s}^3_R(f(\boldsymbol{x}), \boldsymbol{\lambda}^U, \boldsymbol{\lambda}^A)$

Note that if the resulting optimal value of D-level set is positive in the case of TSPASF, then it corresponds to the case of unachievable reference point. A negative value signals about reference point achievability. In case of PASF, negative value of D-level set is not possible.

Computational Experiments

In this section, we consider some computational experiments for TSPASF. Compared with PASF in [15] and two-slope ASF in [4], TSPASF generalizes both of them. Due to the theoretical properties of these functions, when the reference point is unachievable, the both TSPASF and PASF give the same solution. In the case of achievable reference point, PASF cannot be used. On the other hand, regardless of the reference point, the same solution is obtained with two-slope ASF and TSPASF when $q = 1$.

In order to explore the behavior of TSPASF, several test problems described in [13] are used. The computational calculations are carried out by applying multiobjective proximal bundle method [7, 9]. This method is designed for nonconvex and constrained problems with possibly several objective functions.

The computational tests are divided into two groups based on their convexity. In both cases twenty reference points between the ideal and nadir point are randomly generated and the used weighting vectors $\boldsymbol{\lambda}^U$, $\boldsymbol{\lambda}^A$ are of the form

$$\boldsymbol{\lambda}^U = \frac{1}{f^N - f^R}, \quad \boldsymbol{\lambda}^A = \frac{1}{f^R - f^I}$$

as suggested in [4].

In these computational tests, we have concentrated on studying three aspects. The first one is to guarantee that by changing the value of the parameter q we can indeed obtain different solutions, not only in theory but also in practice. Another interesting issue is a sparsity of the solutions produced or how much solutions differ. The last aspect, which has been singled out and studied, is the computational time. Since the computational time may vary a lot between convex and nonconvex problems, these both cases are considered. In addition, the Pareto frontier is usually much nicer in convex than in nonconvex case.

According to our results, most of the times the solutions obtained by varying the value of the parameter q differ and the difference is significant both in the convex and the nonconvex test problems. In the convex case, the solution times produced with the different values of the parameter q are the same order whereas in the nonconvex test problem $q = 1$ turns out to be clearly the most time-consuming value of the parameter q. Thus, with TSPASF by varying metrics between Chebyshev and linear metric, the different solutions with good sparsity are obtained without growing computational efforts compared with the two-slope ASF [4] equaling the case $q = 1$ in TSPASF.

In the following, the convex and the nonconvex test problems are analyzed closer.

Convex Case

We consider the following convex Chankong-Haimes test problem [13]

$$\begin{aligned} \min\ & f(\boldsymbol{x}) = ((x_1-1)^2+(x_2-1)^2,\ (x_1-2)^2+(x_2-3)^2, \\ & (x_1-4)^2+(x_2-2)^2) \\ \text{s. t.}\ & x_1+2x_2 \leq 10 \\ & 0 \leq x_1 \leq 10 \\ & 0 \leq x_2 \leq 4, \end{aligned}$$

with three objectives and two variables.

In the following, we refer to the cases where for one reference point a solution is calculated with every value of the parameter $q \in \mathbb{N}_3$. Thus, twenty cases are considered due to the number of the generated reference points. Among these randomly generated reference points, there are 40% of achievable reference points and 60% of unachievable reference points.

In Table 1, the difference and the sparsity of the solutions are analyzed. Here solutions with linear metric ($q = 3$) and Chebyshev metric ($q = 1$) are compared with the case where $q = 2$ and the used metric is something between Chebyshev and linear metric.

Table 1 Differences of solutions

Convex case	$q = 1$	$q = 3$
Cases where $\boldsymbol{x}^*$ with $q = 2$ equals to $\boldsymbol{x}^*$ with	15%	0%
Average relative distance between $f(\boldsymbol{x}^*)$ with $q = 2$ and	0.52065	0.31203
Nonconvex case	$q = 1$	$q = 3$
Cases where $\boldsymbol{x}^*$ with $q = 2$ equals to $\boldsymbol{x}^*$ with	0%	10%
Average relative distance between $f(\boldsymbol{x}^*)$ with $q = 2$ and	201 203.24	0.15565

In Table 1, the first row presents the percentage value of cases where both values $q = 1$ and $q = 3$ give the same solution as the solution obtained with value $q = 2$. As Table 1 indicates, with values $q = 3$ and $q = 2$ we never obtain the same solution, and with values $q = 1$ and $q = 2$ only 15% of cases the solutions were the same. Thus, by varying the value of the parameter q, mostly different solutions are obtained.

In the second row, the sparsity of the solutions is considered by calculating the (relative) distances between solutions. Table 1 reports the averages of these distance calculations. In the comparison of the distances between solutions obtained with $q = 1$ and $q = 2$, the average distance in the objective space is 0.52065. By excluding the cases where the same solution was obtained, every distance belongs to the interval from 0.069274 to 1.59813 in the objective space. The average distance with values $q = 2$ and $q = 3$ is 0.31203 in the objective space, and each of these solutions belongs to the interval from 0.11194 to 0.70154 in the objective space. Based on these calculations, it can be said that the differences between the solutions are significant.

Since the solutions obtained by varying the value of the parameter q are actually various solutions, it is also interesting to know what the price of the different solutions is in terms of the number of the iterations. To explore this aspect, in Table 2 we describe the average number of the iterations and function calls needed when the calculations are carried out with multiobjective proximal bundle method. In these calculations, both the average number of the iterations and the function calls are approximately on the same order regardless of the value of the parameter q.

Nonconvex Case

We scrutinize the following nonconvex water resources planning test problem [13]

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) = (e^{0.001x_1} x_1^{0.02} x_2^2,\ 0.5x_2^2,\ -e^{0.005x_1} x_1^{0.001} x_2^2) \\ \text{s. t.} \quad & 0.01 \le x_1 \le 1.3 \\ & 0.01 \le x_2 \le 10, \end{aligned}$$

Table 2 Computational times

Convex case	$q = 1$	$q = 2$	$q = 3$
Average number of iterations	12.20	16.95	6.00
Average number of function calls	15.10	22.80	8.10
Nonconvex case	$q = 1$	$q = 2$	$q = 3$
Average number of iterations	30.9	7.45	5.85
Average number of function calls	53.3	15.45	13.6

with three objectives and two variables as were also in the example problem in the convex case. Again, in this nonconvex case 20 different reference points are generated randomly. There are now 35% of achievable and 65% of unachievable reference points.

In Table 1, the sparsity of the solutions is analyzed also in the nonconvex case. As Table 1 shows, with values $q = 1$ and $q = 2$ the same solution was never obtained and with values $q = 2$ and $q = 3$ only 10% of cases the solutions produced are the same. Thus, the different solutions are obtained by varying the parameter q also in the nonconvex case. By comparing this with the convex case, we see that the total number of the same solutions is now smaller. In addition, the same solutions are produced only with values $q = 2$ and $q = 3$ whereas in the convex case the same solutions are produced with values $q = 1$ and $q = 2$.

When we consider the (relative) distances between the solutions produced described in Table 1, we see that in the nonconvex case the distances are significant as was also in the convex example. The average distance between the solutions obtained with $q = 1$ and $q = 2$ is 201,203.24 in the objective space and every distance belongs to the interval from 0.0075685 to 762,114. The large end point of the interval is the cause of the relative distance since the solution obtained in the objective space in case $q = 2$ is so close to the zero vector. The distances between the solutions obtained with $q = 2$ and $q = 3$ is 0.15565 in the objective space. By excluding the cases where the same solution was obtained, every distance belongs to the interval from 0.00067393 to 0.99999 in the objective space.

As said above, in Table 2, the computational times are described, and in the convex case the value of the parameter q does not affect computational time significantly. In the nonconvex case, the most time-consuming value of the parameter q is 1 representing Chebyshev type function. In this case, the number of the iterations and the function calls are both significantly larger than the corresponding values for other two values of the parameter q.

Conclusion

In this paper, we have presented a new family of achievement scalarizing functions based on a parameterization utilizing two different weighting vectors depending on whether the reference point is achievable or not. We have proven that we always find at least weakly Pareto optimal solution, and, if the solution is unique, it is Pareto optimal. We have also proven that every weakly Pareto optimal solution can be produced. Furthermore, we can find the different solutions by changing the value of the parameter q, the reference point or the weighting vectors. Additionally, to use TSPASF, there is no need for any assumptions about the reference point nor test the reference point achievability.

We have also illustrated the shapes of the different D-levels, and the computational tests have been performed for both convex and nonconvex problems. These results have shown that the sparsity of the solutions produced is good. The computational time does not grow with the different values of the parameter q in the convex case, but in the nonconvex case, Chebyshev type function turns out the most time-consuming value of the parameter q.

The presented TSPASF gives a systematic way to produce different (weakly) Pareto optimal solutions from the same preference information with different metrics. The property of this kind could be used, for example, in some interactive methods. The different solutions can be calculated with all values of the parameter q or just some of them. Thus, it is interesting to know more about how the value q affects the shape of D-level.

In future, the one possible continuation of this work is to apply TSPASF in synchronous decision making approach. For example, the water treatment facilities design problem with four objectives [14] can be considered. Another possible future research is the use of TSPASF in data envelopment analysis.

Acknowledgements The research has been financially supported by the Finnish Academy of Science and Letters (the Vilho, Yrjö and Kalle Väisälä Foundation), Emil Aaltonen Foundation and University of Turku Graduate School UTUGS Matti programme.

References

1. K. Deb, K. Miettinen, Nadir point estimation using evolutionary approaches: better accuracy and computational speed through focused search, in *Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems*, ed. by M. Ehrgott, B. Naujoks, T.J. Stewart, J. Wallenius (Springer, Berlin, 2010), pp. 339–354
2. M. Ehrgott, *Multicriteria Optimization*, 2nd edn. (Springer, Berlin, 2005)
3. V.P. Eronen, M.M. Mäkelä, T. Westerlund, Extended cutting plane method for a class of nonsmooth nonconvex MINLP problems. Optim.: J. Math. Program. Oper. Res. **64**(3), 1–21 (2013)
4. M. Luque, K. Miettinen, A.B. Ruiz, F. Ruiz, A two-slope achievement scalarizing function for interactive multiobjective optimization. Comput. Oper. Res. **39**, 1673–1681 (2012)
5. M. Luque, L.A. Lopez-Agudo, O.D. Marcenaro-Gutierrez, Equivalent reference points in multiobjective programming. Expert Syst. Appl. **42**, 2205–2212 (2015)
6. M.M. Mäkelä, Survey of bundle methods for nonsmooth optimization. Optim. Methods Softw. **17**(1), 1–29 (2002)
7. M.M. Mäkelä, Multiobjective proximal bundle method for nonconvex nonsmooth optimization: fortran subroutine MPBNGC 2.0. Tech. Rep. B 13/2003, Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing, University of Jyväskylä, Jyväskylä (2003)

8. M.M. Mäkelä, V.P. Eronen, N. Karmitsa, On nonsmooth multiobjective optimality conditions with generalized convexities, in *Optimization in Science and Engineering*, ed. by T.M. Rassias, C.A. Floudas, S. Butenko (Springer, New York, 2014), pp. 333–357
9. M.M. Mäkelä, N. Karmitsa, O. Wilppu, Proximal bundle method for nonsmooth and nonconvex multiobjective optimization, in *Mathematical Modeling and Optimization of Complex Structures*, ed. by T. Tuovinen, S. Repin, P. Neittaanmäki, Computational Methods in Applied Sciences, vol 40 (Springer International Publishing, New York, 2016), pp. 191–204
10. K. Miettinen, *Nonlinear Multiobjective Optimization* (Kluwer Academic Publishers, Boston, 1999)
11. K. Miettinen, M.M. Mäkelä, On scalarizing functions in multiobjective optimization. OR Spectr. **24**, 193–213 (2002)
12. K. Miettinen, M.M. Mäkelä, Synchronous approach in interactive multiobjective optimization. Eur. J. Oper. Res. **170**, 909–922 (2006)
13. K. Miettinen, M.M. Mäkelä, K. Kaario, Experiments with classification-based scalarizing functions in interactive multiobjective optimization. Eur. J. Oper. Res. **175**, 931–947 (2006)
14. K. Miettinen, D. Podkopaev, F. Ruiz, M. Luque, A new preference handling technique for interactive multiobjective optimization without trading-off. J. Glob. Optim. **63**(4), 633–652 (2015)
15. Y. Nikulin, K. Miettinen, M.M. Mäkelä, A new achievement scalarizing function based on parameterization in multiobjective optimization. OR Spectr. **34**, 69–87 (2012)
16. F. Ruiz, M. Luque, F. Miguel, M. del Mar Muñoz, An additive achievement scalarizing function for multiobjective programming problems. Eur. J. Oper. Res. **188**(3), 683–694 (2008)
17. Y. Sawaragi, H. Nakayama, T. Tanino, *Theory of Multiobjective Optimization* (Academic, Orlando, 1985)
18. A.P. Wierzbicki, Basic properties of scalarizing functionals for multiobjective optimization. Optimization **8**, 55–60 (1977)
19. A.P. Wierzbicki, The use of reference objectives in multiobjective optimization, in *Multiple Criteria Decision Making Theory and Applications. MCDM Theory and Applications Proceedings*, ed. by G. Fandel, T. Gal. Lecture Notes in Economics and Mathematical Systems, vol. 177 (Springer, Berlin, 1980), pp. 468–486
20. A.P. Wierzbicki, A methodological approach to comparing parametric characterizations of efficient solutions, in *Large-scale Modelling and Interactive Decision Analysis*, ed. by G. Fandel et al., Lecture Notes in Economics and Mathematical Systems, vol. 273 (Springer, Berlin, 1986), pp. 27–45
21. A.P. Wierzbicki, On the completeness and constructiveness of parametric characterizations to vector optimization problems. OR Spectr. **8**, 73–87 (1986)
22. O. Wilppu, M.M. Mäkelä, Y. Nikulin, Two-slope parameterized achievement scalarizing functions for multiobjective optimization. Tech. Rep. 1114, TUCS Technical Reports, Turku Centre for Computer Science, Turku (2014)