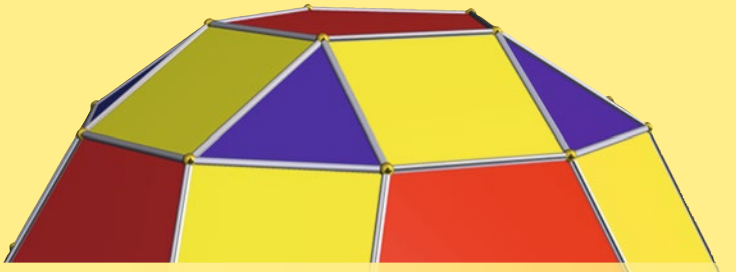
Themistocles M. Rassias
Christodoulos A. Floudas · Sergiy Butenko
Editors



Optimization in Science and Engineering

In Honor of the 60th Birthday of Panos M. Pardalos

Springer

Optimization in Science and Engineering

Themistocles M. Rassias • Christodoulos A. Floudas
Sergiy Butenko
Editors

Optimization in Science and Engineering

In Honor of the 60th Birthday of Panos M. Pardalos

Springer

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

Sergiy Butenko
Industrial and Systems Engineering
Texas A&M University
College Station, TX, USA

Christodoulos A. Floudas
Department of Chemical
and Biological Engineering
Princeton University
Princeton, NJ, USA

ISBN 978-1-4939-0807-3 ISBN 978-1-4939-0808-0 (eBook)
DOI 10.1007/978-1-4939-0808-0
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014936203

Mathematics Subject Classification (2010): 05C85, 47N10, 65K10, 65K15, 90C26, 91A40

Cover illustration: The cover figure is one of the Archomedian solids, rhombicosidodecahedron, which has 60 vertices and 120 edges. Created by Robert Webb's Stella software: http://www.software3d.com/Stella.php

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This book is dedicated to
the 60th birthday of
Panos M. Pardalos

Preface

Panos Pardalos was born to parents Calypso and Miltiades on June 17, 1954, in Mezilo (now Drossato), Greece. Ever since his grandmother Sophia taught him how to count in his early childhood, Panos has been fascinated with mathematics. The remote location of the mountain village and rather unfavorable economic conditions Panos grew up in did not stop him from pursuing knowledge. When he was 15, Panos wrote a letter to the Greek Ministry of Education describing his aspirations and the obstacles he faced in his quest for learning. The government responded by providing a scholarship to support his studies at Athens University.

After obtaining a bachelor's degree in mathematics in 1977, Panos continued his education in the USA. In 1978, he earned a master's degree in mathematics and computer science from Clarkson University (Potsdam, NY) and started Ph.D. studies in computer and information sciences at the University of Minnesota. In 1985, Panos successfully defended his dissertation, which served as the basis for his first book *Constrained Global Optimization: Algorithms and Applications* (Springer-Verlag, 1987) coauthored with his Ph.D. advisor, Judah Ben Rosen. This book became a landmark publication in the emerging field of global optimization and helped Panos to establish himself as one of the leading researchers in the field. By the time of the book's publication he already started his independent academic career as an assistant professor of computer science at Pennsylvania State University.

In 1991, Panos moved to the Department of Industrial and Systems Engineering at the University of Florida (UF), where he currently holds the position of Distinguished Professor and University of Florida Research Foundation Professor. He also serves as the director of Center for Applied Optimization. At UF, Panos is also an affiliated faculty of Computer & Information Science & Engineering Department, Biomedical Engineering Department, McKnight Brain Institute, and the Genetics Institute.

Panos compiled a very impressive record over the years of his (still very active) academic career, which includes nearly 20 coauthored books and over 300 journal articles. He is also an editor of numerous books, including a 7-volume *Encyclopedia of Optimization* co-edited with Christodoulos Floudas and published by Springer. He served as the editor-in-chief and an editorial board member of many highly

respected journals and as the managing editor of several book series. He has organized conferences and gave plenary lectures in world leading institutions. Over 50 of his former Ph.D. students enjoy successful careers in academia and industry, making the impact of his mentoring felt all over the world.

Panos has been honored with a number of awards for his scholastic achievements. His notable recognitions include the Constantin Carathéodory Prize (2013) and EURO Gold Medal (2013); Honorary Doctorates from N.I. Lobachevski State University of Nizhni Novgorod, Russia (2005), V.M. Glushkov Institute of Cybernetics of The National Academy of Sciences of Ukraine (2008), and Wilfrid Laurier University, Canada (2012); Honorary Professorships from the Graduate School of Information Technology & Mathematical Sciences, University of Ballarat, Australia (2010) and from Anhui University of Sciences and Technology, China (2013). He was elected a Foreign Associate Member of Reial Académia de Doctors, Spain (1998), a Foreign Member of Lithuanian Academy of Sciences (1999), Petrovskaya Academy of Sciences and Arts, Russia (2000), and the National Academy of Sciences of Ukraine (2003), as well as an Honorary Member of the Mongolian Academy of Sciences (2005). He is also the recipient of a medal in recognition of broad contributions in science and engineering of the University of Catania, Italy (2013).

Ivan V. Sergienko, Academician of the National Academy of Sciences of Ukraine (NASU), presents the diploma of a foreign member of NASU to Panos M. Pardalos (2003)

As impressive as his academic accomplishments are, it is safe to say that his personal qualities and friendship are the primary reasons Panos is so much loved and respected by his colleagues and students. As he likes to say, “Whatever it is that we do, we are humans first.” His enthusiasm for science is just a reflection of his positive, energetic, and happy personality. He always remembers his roots and knows how to enjoy simple things in life. Many of the readers might have heard the

following story about Panos that is very characteristic of his caring nature. When he was a Ph.D. student at the University of Minnesota, Panos planted a grapefruit seed in a pot, and a tree started growing. When he moved to Penn State a few years later, he brought the plant with him. The next destination for Panos and the tree was Gainesville, Florida, where the climate was finally warm enough for planting a grapefruit tree outside. After some proficient treatment from Panos's father, the tree thrived as did Panos's career at UF, bearing so much highest-quality fruit that it was plenty not only for the Pardalos family but also for Panos's colleagues and students in the department to enjoy.

Panos with his son, Akis, and wife, Rosemary, next to the famous grapefruit tree, February 1, 2014

On behalf of all the authors of the chapters, we are very pleased to dedicate this book to Panos Pardalos on the occasion of his 60th birthday, and wish him many more happy, healthy, and productive years. We would like to thank all the contributors as well as Razia Amzad and Elizabeth Loew of Springer for making this publication possible.

Χρόνια Πολλά Πάνο!

Athens, Greece — *Themistocles M. Rassias*
Princeton, NJ — *Christodoulos A. Floudas*
College Station, TX — *Sergiy Butenko*

Contents

Piecewise Linear Classifiers Based on Nonsmooth Optimization Approaches

Adil M. Bagirov, Refail Kasimbeyli, Gürkan Öztürk, and Julien Ugon

1 Introduction

Nonsmooth optimization provides efficient algorithms for solving many machine learning problems. For example, nonsmooth optimization approaches to the cluster analysis and supervised data classification problems lead to the design of very efficient algorithms for their solution (see, e.g., [1, 3, 4, 10, 13]). Here our aim is to demonstrate how nonsmooth optimization algorithms can be applied to develop efficient piecewise linear classifiers. We use a max–min and a polyhedral conic separabilities as well as an incremental approach to design such classifiers. This chapter contains results which are extensions of those obtained in [14–16].

The problem of separating finite sets has many applications in applied mathematics. One such application is the design of supervised data classification algorithms. If convex hulls of the sets do not intersect, then they are linearly separable and one hyperplane provides complete separation. However, in many real-world applications this is not the case. In most data sets, classes are disjoint, but their convex hulls intersect. In this situation, the decision boundary between the classes is nonlinear. It can be approximated using piecewise linear functions. Over the last three decades different algorithms to construct piecewise linear decision boundaries between finite sets have been designed and applied to solve data classification problems (see, e.g., [2, 10, 11, 18, 19, 24, 27, 31, 39–41]).

Piecewise linear classifiers are very simple to implement and their memory requirements are very low. Therefore they are suitable for small reconnaissance

A.M. Bagirov (✉) • J. Ugon
School of Science, Information Technology and Engineering, Federation University Australia, Ballarat, VIC 3353, Australia
e-mail: a.bagirov@federation.edu.au; j.ugon@ballarat.edu.au

R. Kasimbeyli • G. Öztürk
Department of Industrial Engineering, Anadolu University, Eskisehir 26480, Turkey
e-mail: rkasimbeyli@anadolu.edu.tr; gurkan.o@anadolu.edu.tr

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_1,

robots, intelligent cameras, imbedded and real-time systems, and portable devices [27]. In general, the determination of piecewise linear boundaries is a complex global optimization problem [40]. The objective function in this problem is nonconvex and nonsmooth. It may have many local minimizers, yet only global minimizers provide piecewise linear boundaries with the least number of hyperplanes. Additionally, the number of hyperplanes needed to separate sets is not known a priori. Newton-like methods cannot be applied to solve such problems. As a result piecewise linear classifiers require a long training time, which creates difficulties for their practical application.

In order to reduce the training time most techniques try to avoid solving optimization problems when computing piecewise linear boundaries. Instead they use some form of heuristics to determine the number of hyperplanes. Most of these techniques apply fast clustering algorithms (such as k-means) to find clusters in each class. Then they compute hyperplanes separating pairs of clusters from different classes. The final piecewise linear boundary is obtained as a synthesis of those hyperplanes (see [18, 19, 24, 27, 31, 39–41]). These techniques try to train hyperplanes *locally*. Despite the fact that these algorithms are quite fast they do not always find minimizers, even local ones of the classification error function.

In this chapter, we propose a different approach to design piecewise linear classifiers. This approach is based on the use of (1) hyperboxes, which can be described by the very simple piecewise linear functions, to identify data points which are away from boundaries between pattern classes; (2) polyhedral conic separability to accurately identify data points lying on or near the boundaries between the classes; (3) max–min separability and an incremental approach to find piecewise linear boundaries between pattern classes.

Following these steps first, we approximate classes using hyperboxes and identify data points which are away from the boundaries between classes. Such points can be easily classified using only approximating hyperboxes. In the next iteration we remove all these points from the further consideration and apply the polyhedral conic separability to more accurately identify data points which are on or close to boundaries between classes. In this iteration we also identify regions which can be classified using the polyhedral conic functions (PCFs). Then we remove all data points from these regions and apply max–min separability to the rest of the data set to find piecewise linear boundaries between the sets. Piecewise linear boundaries are built by gradually adding new hyperplanes until separation is obtained with respect to some tolerance. Such an approach allows one to significantly reduce computational effort to train piecewise linear classifiers and considerably improve their classification accuracy. We apply the proposed classifiers to solve supervised data classification problems in 12 publicly available data sets, report the results of numerical experiments, and compare the proposed classifiers with nine other mainstream classifiers.

The rest of this chapter is organized as follows: In Sect. 2 we give an overview of existing piecewise linear classifiers. The definition and some results related to max–min separability are given in Sect. 3. The classification algorithm based on the PCF is described in Sect. 4. Section 5 presents the incremental max–min separability

algorithm. The hybrid polyhedral conic and max–min separability (HPCAMS) algorithm, its implementation, and classification rules are given in Sect. 6. Results of numerical experiments are presented in Sect. 7. Section 8 concludes the chapter.

2 Review of Piecewise Linear Classifiers

Piecewise linear classifiers have been a subject of study for more than three decades. Despite the fact that the computation of piecewise linear boundaries is not an easy task, piecewise linear classifiers are simple to implement, provide a fast (real-time) classification time and have a low memory requirement. Another advantage of piecewise linear classifiers is that they do not depend on parameters. The simplicity of their implementation makes them very suitable for many applications [27].

Existing piecewise linear classifiers can be divided into two classes. The first class contains classifiers in which each segment of the piecewise linear boundary is constructed independently. An optimization problem is formulated for each segment separately. Thus these segments are found as a solution to different optimization problems. We call such an approach a *multiple optimization approach*.

The second class contains classifiers in which the problem of finding a piecewise linear boundary is formulated as an optimization problem. In this case a single optimization problem is solved to find piecewise linear boundaries. We call such an approach *a single optimization approach*.

2.1 Classifiers Based on a Multiple Optimization Approach

To the best of our knowledge, the first approach to construct a piecewise linear classifier was described in [39] (see also [40]). This paper introduces a procedure to locally train piecewise linear decision boundaries. Correctly classified patterns provide adjustments only in those segments of the decision boundary that are affected by those patterns.

The method proposed in [32] is based on the cutting of straight line segments joining pairs of opposed points (i.e., points from distinct classes) in n-dimensional space. The authors describe a procedure to nearly minimize the number of hyperplanes required to cut all of these straight lines. This method does not require parameters to be specified by users, an improvement over methods proposed in [39]. This piecewise linear classifier provides a much faster decision than the k-nearest neighbors classifier for a similar accuracy. In [28], the piecewise linear classifier is compared with a neural network classifier. The latter performs slightly better than the former, but it requires a much longer training time.

In the paper [41] a modification of the method from [32] is proposed. This method constructs the hyperplanes of a piecewise linear classifier so as to keep

a correct recognition rate over a threshold for the training set. The threshold is determined automatically by the Minimum Description Length criterion so as to avoid overfitting of the classifier to the training set.

The paper [37] presents a learning algorithm which constructs a piecewise linear classifier for multi-class data classification problems. In the first step of the algorithm linear regression is used to determine the initial positions of the discriminating hyperplanes for each pair of classes. An error function is minimized by a gradient descent procedure for each hyperplane separately. A clustering procedure decomposing the classes into appropriate subclasses can be applied when the classes are not linearly separable. This classifier was included in the STATLOG project where it achieved good classification results on many data sets [29].

The paper [18] proposes an approach to construct a piecewise linear classifier using neural networks. The training set is split into several linearly separable training subsets, and the separability is preserved in subsequent iterations. In [27] the piecewise linear boundary is represented as a collection of segments of hyperplanes created as perpendicular bisectors of the line segments linking centroids of the classes or parts of classes.

The paper [31] proposes a piecewise linear classifier which starts with a linear classifier. If it fails to separate the classes, then the sample space of one of the classes is divided into two subsample spaces. This sequence of splitting, redesigning, and evaluating continues until the overall performance is no longer improved.

In [19] the authors propose a linear binary decision tree classifier, where the decision at each non-terminal node is made using a genetic algorithm. They apply this piecewise linear classifier to cell classification.

2.2 Classifiers Based on a Single Optimization Approach

There are different approaches to design piecewise linear classifiers based on a single optimization approach. The notion of the bilinear separation was introduced in [17]. In this approach two hyperplanes were used to separate classes. An algorithm for finding those hyperplanes was also developed.

The paper [2] introduces the concept of polyhedral separability which is a generalization of linear separability. In this case one of the sets is approximated by a polyhedral set and the rest of the space is used to approximate the second set. The error function is a sum of nonsmooth convex and nonsmooth nonconvex functions. An algorithm for minimizing the error function is developed where the problem of finding the descent directions is reduced to a linear programming problem.

The concept of max–min separability was introduced in [10]. In this approach two sets are separated using a continuous piecewise linear function. Max–min separability is a generalization of linear, bilinear, and polyhedral separabilities [11]. It is proven that any two finite point sets can be separated by a piecewise linear function. The error function in this case is nonconvex and nonsmooth. An algorithm for

minimizing the error function is developed. Results presented in [11] demonstrate that the algorithm based on max–min separability is effective for solving supervised data classification problems in many large-scale data sets.

Polyhedral conic separability was introduced in [22] where PCFs are used to separate classes. An algorithm for finding such separating PCFs was also designed.

Incremental learning algorithms are becoming increasingly popular in supervised and unsupervised data classification. This type of approach breaks up the data set into observations that can be classified using simple separators, and observations that require more elaborate ones. This allows one to simplify the learning task by eliminating the points that can be more easily classified. Furthermore, at each iteration, information gathered during prior iterations can be exploited. In the case of piecewise linear classifiers, this approach allows us to compute as few hyperplanes as needed to separate the sets, without any prior information. Additionally, this approach allows us to reach a near global solution of the classification error function by using the piecewise linear function obtained at a given iteration as a starting point for the next iteration. Thus it reduces computational effort and avoids possible overfitting. Papers [14–16] present different incremental piecewise linear classifiers. Piecewise linear classifiers introduced in these papers are based in the max–min and polyhedral conic separabilities. In these classifiers simple piecewise linear separators such as hyperboxes are used to find the set of easily classifiable points.

3 Max–Min Separability

The approach we propose in this chapter finds piecewise linear boundaries of classes. These boundaries are determined using max–min separability, a concept which was introduced in [10] (see also [11]). In this section we briefly recall the main definitions from these papers.

3.1 Definition and Properties

Let A and B be given disjoint sets containing m and p n-dimensional vectors, respectively:

$$A = \{a^1, \ldots, a^m\}, a^i \in \mathbb{R}^n, i = 1, \ldots, m,$$
$$B = \{b^1, \ldots, b^p\}, b^j \in \mathbb{R}^n, j = 1, \ldots, p.$$

Consider a collection of hyperplanes $H = \left\{\{x^{ij}, y_{ij}\},\ j \in J_i,\ i \in I\right\}$, where $x^{ij} \in \mathbb{R}^n$, $y_{ij} \in \mathbb{R}^1$, $j \in J_i$, $i \in I$, and $I = \{1, \ldots, l\}$, $l > 0$, $J_i \neq \emptyset\ \forall i \in I$.

This collection of hyperplanes defines the following max–min function on $\mathbb{R}^n$:

$$\varphi(z) = \max_{i \in I} \min_{j \in J_i} \left\{\langle x^{ij}, z\rangle - y_{ij}\right\},\ z \in \mathbb{R}^n. \tag{1}$$

Here $\langle \cdot, \cdot \rangle$ is an inner product in $\mathbb{R}^n$.

Definition 1. The sets A and B are max–min separable if there exist a finite number of hyperplanes $\{x^{ij}, y_{ij}\}$ with $x^{ij} \in \mathbb{R}^n,\ y_{ij} \in \mathbb{R}^1,\ j \in J_i,\ i \in I$ such that

1. for all $i \in I$ and $a \in A$
$$\min_{j \in J_i} \left\{ \langle x^{ij}, a \rangle - y_{ij} \right\} < 0;$$
2. for any $b \in B$ there exists at least one $i \in I$ such that
$$\min_{j \in J_i} \left\{ \langle x^{ij}, b \rangle - y_{ij} \right\} > 0.$$

Remark 1. It follows from Definition 1 that if the sets A and B are max–min separable then $\varphi(a) < 0$ for any $a \in A$ and $\varphi(b) > 0$ for any $b \in B$, where the function φ is defined by (1). Thus the sets A and B can be separated by a function represented as a max–min of linear functions. Therefore this kind of separability is called max–min separability.

Remark 2. The notions of max–min and piecewise linear separabilities are equivalent. The sets A and B are max–min separable if and only if they are disjoint: $A \bigcap B = \emptyset$ [10].

3.2 Error Function

Given any set of hyperplanes $\{x^{ij}, y_{ij}\},\ j \in J_i,\ i \in I$ with $x^{ij} \in \mathbb{R}^n,\ y_{ij} \in \mathbb{R}^1$ an averaged error function is defined as (see [10, 11])

$$f(X,Y) = f_1(X,Y) + f_2(X,Y) \tag{2}$$

where

$$f_1(X,Y) = (1/m) \sum_{k=1}^{m} \max \left[0, \max_{i \in I} \min_{j \in J_i} \{ \langle x^{ij}, a^k \rangle - y_{ij} + 1 \} \right],$$
$$f_2(X,Y) = (1/p) \sum_{t=1}^{p} \max \left[0, \min_{i \in I} \max_{j \in J_i} \{ -\langle x^{ij}, b^t \rangle + y_{ij} + 1 \} \right],$$

and $X = (x^{11}, \ldots, x^{lq_l}) \in \mathbb{R}^{nL},\ Y = (y_{11}, \ldots, y_{lq_l}) \in \mathbb{R}^L,\ L = \sum_{i \in I} q_i,\ q_i = |J_i|,\ i \in I = \{1, \ldots, l\}$. $|J_i|$ denotes the cardinality of the set J_i. It is clear that $f(X,Y) \geq 0$ for all $X \in \mathbb{R}^{nL}$ and $Y \in \mathbb{R}^L$.

Remark 3. The error function (2) is nonconvex and if the sets A and B are max–min separable with the given number of hyperplanes, then the global minimum of this function $f(X^*, Y_*) = 0$ and the global minimizer is not always unique. Moreover, $X = 0 \in \mathbb{R}^{nL}$ cannot be an optimal solution [10].

The problem of max–min separability is reduced to the following mathematical programming problem:

$$\text{minimize } f(X,Y) \text{ subject to } (X,Y) \in \mathbb{R}^{(n+1)L} \tag{3}$$

where the objective function f is described by Eq. (2).

In the paper [11], an algorithm for solving problem (3) is presented. This algorithm exploits special structures of the error function such as piecewise partial separability (for the definition of piecewise partial separability, see [12]). In this algorithm it is assumed that the number of hyperplanes is known a priori. However this information is not always available. The classification accuracy is highly dependent on this number. A large number of hyperplanes may lead to overfitting of the training set. It is therefore imperative to calculate as few hyperplanes as needed to separate classes with respect to a given tolerance. An incremental approach can be applied to solve this problem.

The complexity of the error function (2) computation depends on the number of data points. For data sets containing tens of thousands of points the error function becomes expensive to compute, and the algorithms proposed in [10, 11] become very time consuming. In most large data sets not all data points contribute to the piecewise linear functions separating classes. Such points are away from the boundaries between classes. Identification of such data points is decisive to reduce (and sometimes significantly) computational effort to evaluate the error function. An incremental approach allows one to reduce the number of points at each iteration by eliminating points easily classified using simpler piecewise linear separators calculated at previous iterations. Also, this scheme allows us to reduce the risk of overfitting by only considering the data points that are relevant.

4 Classification Algorithms Based on PCFs

In this section we describe classification algorithms based on the separation via PCFs.

It is impossible to overestimate the importance of theorems on the existence of a separating hyperplane for two disjoint convex sets. A large number of methods for solving single-objective optimization problems are based on these theorems. In convex vector optimization, it is a common practice to characterize efficient points of sets as support points by positive or strictly positive linear support functionals. Many classification algorithms are based on the linear separation theorems. Unfortunately convex hulls of many data sets encountered in classification problems are not disjoint and therefore the linear separation theorems of convex analysis used in data classification problems leads to difficulties. The simple reason is that, for disjoint nonconvex sets a separating hyperplane may not exist. Therefore, nonconvex analysis requires special separation theorems.

The main reason of difficulties arising when passing from the convex analysis to the nonconvex one is that the nonconvex cases may arise in many different forms and each case may require a special approach. Some problems of nonconvex optimization, in more generalized form, have been studied in the framework of the

abstract convexity (see [30, 33–36, 38]). Abstract convexity suggests variety of approaches which can be used to analyze different nonconvex problems. It generalizes the existing supporting philosophy for convex sets and suggests different ways to support nonconvex sets by using a suitable class of real functions alternatively to the class of linear functions used in convex analysis. These investigations demonstrate the importance of finding a specific class of functions defining special nonlinear supporting surfaces which are suitable to analyze the given nonconvex problem.

In [20] Gasimov suggested a special type of PCFs and with their help obtained characterization theorems for Benson properly efficient points in vector optimization without any convexity and boundedness conditions. By using the same class of PCFs, recently Kasimbeyli [25] proved a nonlinear separation theorem for nonconvex cones.

In the following subsection we present the class of PCFs. The subsequent subsection demonstrates how the separation technique based on PCFs is used in nonconvex nonsmooth optimization and then we give the corresponding classification algorithms.

4.1 Polyhedral Conic Functions

The class of PCFs that we consider in this subsection consists of functions $g_{(w,\xi,\gamma,a)}$: $\mathbb{R}^n \to \mathbb{R}$ defined as:

$$g_{(w,\xi,\gamma,a)}(x) = \langle w, x-a\rangle + \xi \|x-a\|_1 - \gamma, \tag{4}$$

where $w, a \in \mathbb{R}^n$, $\xi, \gamma \in \mathbb{R}$, and $\|x\|_1 = |x_1| + \cdots + |x_n|$ is an l_1-norm of the vector $x \in \mathbb{R}^n$.

Lemma 1. *A graph of the function $g_{(w,\xi,\gamma,a)}$ defined in Eq. (4) is a polyhedral cone with vertex at $(a, -\gamma) \in \mathbb{R}^n \times \mathbb{R}$.*

Proof. To prove the lemma we show that:

1. A graph of the function is a cone with vertex at $(a, -\gamma) \in \mathbb{R}^n \times \mathbb{R}$, and
2. Each sublevel set of this function is a convex polyhedron.

To prove the first part, consider a set $\text{graph}(g_{(w,\xi,\gamma,a)}) - (a, -\gamma)$:

$$\text{graph}(g_{(w,\xi,\gamma,a)}) - (a, -\gamma) = \{(x-a, \alpha+\gamma) : \langle w, x-a\rangle + \xi \|x-a\|_1 - \gamma = \alpha\}.$$

By letting $x - a = y$, $\alpha + \gamma = \beta$ this set can be written also as

$$\text{graph}(g_{(w,\xi,\gamma,a)}) - (a, -\gamma) = \{(y, \beta) : \langle w, y\rangle + \xi \|y\|_1 = \beta\}. \tag{5}$$

It is obvious that this set is a cone with vertex at the origin. Indeed if $(y, \beta) \in \text{graph}(g_{(w,\xi,\gamma,a)}) - (a, -\gamma)$ then

$$\langle w, y\rangle + \xi \|y\|_1 = \beta.$$

Hence, for any $\lambda > 0$ we have:

$$\lambda\langle w, y\rangle + \lambda\xi \|y\|_1 = \lambda\beta,$$

or

$$\langle w, \lambda y\rangle + \xi \|\lambda y\|_1 = \lambda\beta,$$

which implies that $(\lambda y, \lambda\beta)$ also belongs to $\text{graph}(g_{(w,\xi,\gamma,a)}) - (a, -\gamma)$ and therefore this set is a cone with vertex at the origin.

Now we show the second part of the proof. Let α be a real number. Then the sublevel set of the function $g_{(w,\xi,\gamma,a)}$ given by (4) is:

$$S_\alpha = \left\{x \in \mathbb{R}^n : g_{(w,\xi,\gamma,a)}(x) = \langle w, x-a\rangle + \xi \|x-a\|_1 - \gamma \le \alpha\right\}.$$

By using a definition of l_1-norm, this set can equivalently be written as

$$S_\alpha = \left\{x \in \mathbb{R}^n : \langle \tilde{w}, x-a\rangle - \gamma \le \alpha\right\},$$

where

$$\langle \tilde{w}, x-a\rangle = \sum_{i=1}^{n} \left(w_i + \xi sgn(x_i - a_i)\right)(x_i - a_i).$$

This means that the sublevel set S_α is an intersection of utmost 2^n half spaces and therefore is a convex polyhedron. The proof is completed. □

Definition 2. A function $g : \mathbb{R}^n \to \mathbb{R}$ is called *polyhedral conic* if its graph is a cone and all its sublevel sets

$$S_\alpha = \left\{x \in \mathbb{R}^n : g(x) \le \alpha\right\},$$

for $\alpha \in \mathbb{R}$, are polyhedrons.

It follows from Lemma 1 that each function of the form (4) is a PCF.

4.2 Conical Supporting Surfaces in Nonsmooth Optimization

By using the PCFs, Azimov and Gasimov introduced the notion of the weak subdifferential, which is a generalization of the classic subdifferential [6, 7]. With the help of this notion, a collection of zero duality gap conditions for a wide class of nonconvex and nonsmooth optimization problems was derived. In this subsection we give some important properties of the weak subdifferentials and study some relationships between the weak subdifferentials and the directional derivatives in the nonconvex case. We recall the concept of the supporting cones and the weak subdifferentials (see [6, 7, 21, 23, 26]).

Let $(X, \|\cdot\|_X)$ be a real normed space, and let X^* be the topological dual of X. Let $(x^*, c) \in X^* \times \mathbb{R}_+$, where $\mathbb{R}_+$ is the set of nonnegative real numbers. We define the conic surface $C(\bar{x}; x^*, c) \subset X$ with vertex at $\bar{x} \in X$ as follows:

$$C(\bar{x};x^*,c)=\{x\in X:\langle x^*,x-\bar{x}\rangle-c\,\|x-\bar{x}\|=0\}. \tag{6}$$

Then the corresponding upper- and lower-conic halfspaces are, respectively, defined as

$$C^+(\bar{x};x^*,c)=\{x\in X:\langle x^*,x-\bar{x}\rangle-c\,\|x-\bar{x}\|\le 0\} \tag{7}$$

and

$$C^-(\bar{x};x^*,c)=\{x\in X:\langle x^*,x-\bar{x}\rangle-c\,\|x-\bar{x}\|\ge 0\}. \tag{8}$$

Note that if $c=0$, the conic surface $C(\bar{x};x^*,c)$ becomes a hyperplane. Hence the supporting cone defined below is a simple generalization of the supporting hyperplane.

Definition 3. $C(\bar{x};x^*,c)$ is called the supporting cone to the set $S\subset X$ if $S\subset C^+(\bar{x};x^*,c)$ (or $S\subset C^-(\bar{x};x^*,c)$) and $\mathrm{cl}(S)\cap C(\bar{x};x^*,c)\neq\emptyset$.

It is clear that the lower-conic halfspace $C^-(\bar{x};x^*,c)$ is a convex cone with vertex at $\bar{x}$.

Definition 4. Let $F:X\to\mathbb{R}$ be a single-valued function, and let $\bar{x}\in X$ be the given point where $F(\bar{x})$ is finite. A pair $(x^*,c)\in X^*\times\mathbb{R}_+$ is called the weak subgradient of F at $\bar{x}$ if

$$F(x)-F(\bar{x})\ge\langle x^*,x-\bar{x}\rangle-c\|x-\bar{x}\| \text{ for all } x\in X. \tag{9}$$

The set

$$\partial^w F(\bar{x})=\{(x^*,c)\in X^*\times\mathbb{R}_+:F(x)-F(\bar{x})\ge\langle x^*,x-\bar{x}\rangle-c\|x-\bar{x}\| \text{ for all } x\in X\}$$

of all weak subgradients of F at $\bar{x}$ is called the weak subdifferential of F at $\bar{x}$. If $\partial^w F(\bar{x})\neq\emptyset$, then F is called weakly subdifferentiable at $\bar{x}$. If (9) is satisfied only for $x\in S$, where $S\subset X$, then we say that F is weakly subdifferentiable at $\bar{x}$ on S. The weak subdifferential of F at $\bar{x}$ on S will be denoted by $\partial^w_S F(\bar{x})$.

Remark 4. It is obvious that, when F is subdifferentiable at $\bar{x}$ (in the classical sense), then F is also weakly subdifferentiable at $\bar{x}$; that is, if $x^*\in\partial F(\bar{x})$, then by definition $(x^*,c)\in\partial^w F(\bar{x})$ for every $c\ge 0$. It follows from Definition 4 that the pair $(x^*,c)\in X^*\times\mathbb{R}_+$ is a weak subgradient of F at $\bar{x}\in X$ if there is a continuous (superlinear) concave function

$$g(x)=\langle x^*,x-\bar{x}\rangle+F(\bar{x})-c\,\|x-\bar{x}\| \tag{10}$$

such that $g(x)\le F(x)$ for all $x\in X$ and $g(\bar{x})=F(\bar{x})$. The set $\mathrm{hypo}(g)=\{(x,\alpha)\in X\times\mathbb{R}:g(x)\ge\alpha\}$ is a closed convex cone in $X\times\mathbb{R}$ with vertex at $(\bar{x},F(\bar{x}))$. Indeed,

$$\begin{aligned}&\mathrm{hypo}(g)-(\bar{x},F(\bar{x}))\\&=\{(x-\bar{x},\alpha-F(\bar{x}))\in X\times\mathbb{R}:\langle x^*,x-\bar{x}\rangle-c\,\|x-\bar{x}\|\ge\alpha-F(\bar{x})\}\\&=\{(u,\beta)\in X\times\mathbb{R}:\langle x^*,u\rangle-c\,\|u\|\ge\beta\}.\end{aligned}$$

Thus, it follows from (9) and (10) that

$$\mathrm{graph}(g)=\{(x,\alpha)\in X\times\mathbb{R}:g(x)=\alpha\}$$

is a conic surface which is a supporting cone to

$$\operatorname{epi}(F) = \{(x,\alpha) \in X \times \mathbb{R} : F(x) \leq \alpha\}$$

at the point $(\bar{x}, F(\bar{x}))$ in the sense that

$$\operatorname{epi}(F) \subset \operatorname{epi}(g), \text{ and } \operatorname{cl}(\operatorname{epi}(F)) \cap \operatorname{graph}(g) \neq \emptyset.$$

For presentation of the main theorem of this section, we use the following standard assumption.

Assumption 1 *Let*

- *function $F : \mathbb{R}^n \to \mathbb{R}$ be directionally differentiable at $\bar{x} \in \mathbb{R}^n$,*
- *the directional derivative $F'(\bar{x})$ of F at $\bar{x}$ be bounded from below on some neighborhood of $0_{\mathbb{R}^n}$, and*
- *the following apply:*

$$F(x) - F(\bar{x}) \geq F'(\bar{x})(x - \bar{x}) \quad \text{for all } x \in \mathbb{R}^n. \tag{11}$$

Theorem 1. *Let Assumption 1 be satisfied for function $F : \mathbb{R}^n \to \mathbb{R}$. Then F is weakly subdifferentiable at $\bar{x} \in \mathbb{R}^n$ and*

$$F'(\bar{x})(h) = \sup\{\langle x^*, h\rangle - c\|h\| : (x^*, c) \in \partial^w F(\bar{x})\} \quad \text{for all } h \in \mathbb{R}^n, \tag{12}$$

where $F'(\bar{x})(h)$ denotes the directional derivative of F at $\bar{x}$ in the direction h.

The following theorem gives necessary and sufficient optimality conditions in the nonconvex case. First we give a definition of the starshaped set.

Definition 5. A nonempty subset S of a real linear space is called starshaped with respect to some $\bar{x} \in S$ if for all $x \in S$,

$$\lambda x + (1-\lambda)\bar{x} \in S \quad \forall \lambda \in [0,1]. \tag{13}$$

Theorem 2. *Let S be a nonempty subset of $\mathbb{R}^n$ starshaped with respect to $\bar{x} \in S$, and let $F : \mathbb{R}^n \to \mathbb{R}$ be a given function. Suppose that F has a directional derivative at $\bar{x}$ in every direction $x - \bar{x}$ with arbitrary $x \in S$ and that*

$$F(x) - F(\bar{x}) \geq F'(\bar{x})(x - \bar{x}) \quad \text{for all } x \in S.$$

(a) If $\bar{x} \in S$ is a minimal point of F on S, then

$$\sup\{\langle x^*, x - \bar{x}\rangle - c\|x - \bar{x}\| : (x^*, c) \in \partial_S^w F(\bar{x})\} \geq 0 \text{ for all } x \in S. \tag{14}$$

(b) If for some $\bar{x} \in S$ the inequality (14) is satisfied, then $\bar{x}$ is a minimal point of F on S.

4.3 PCF Algorithm

We state now our algorithm for solving the separation problem. Since the algorithm is based on PCFs we will call it PCF Algorithm [22].

We consider the problem of separation of two nonempty finite point sets A and B in $\mathbb{R}^n$. An iterative algorithm generating a nonlinear separating function by using PCFs and therefore called a PCF algorithm is developed. This algorithm is based on solutions of linear programming subproblems. A solution of these subproblems at each iteration results in the PCF which separates a certain part of the set A from the whole set B. By excluding these points from A, algorithm passes to the next iteration and so on. The resulting separation function is defined as a point-wise minimum of all the functions generated. We show that the algorithm terminates in a finite number of iterations and the maximum number of iterations required for separating two arbitrary finite point sets does not exceed the number of elements in one of these sets. An illustrative example has been constructed and application on classification problems has been implemented.

Let A and B be two given sets in $\mathbb{R}^n$:

$$A = \{a^i \in \mathbb{R}^n : i \in I\}, B = \{b^j \in \mathbb{R}^n : j \in J\}$$

where $I = \{1, \ldots, m\}$, $J = \{1, \ldots, p\}$. The algorithm presented below generates at each iteration a function of the form (4) by calculating the parameters w, ξ and γ as a solution of a certain linear programming (LP) subproblem. These parameters are used to define a PCF whose sublevel set divides the whole space into two parts such that all the points of B remain "outside," and as many points of A as possible remain "inside" of this sublevel set. By excluding these latter points from A, algorithm passes to the next iteration and generates new separating function for this modified set. The process continues until an empty set is obtained. The resulting separating function is defined as a point-wise minimum of all functions so generated. We will prove that the algorithm terminates in finite steps.

Algorithm 1. PCFs classification algorithm

Initialization Step: Set $I_1 := I, A_1 := A$, and $l := 1$.

Step 1: Let a^l be an arbitrary point of A_l. Solve subproblem P_l:

$$(P_l) \qquad \min\left(\frac{\langle y, e_m\rangle}{m}\right) \tag{15}$$

subject to

$$\langle w, a^i - a^l\rangle + \xi \left\|a^i - a^l\right\|_1 - \gamma + 1 \le y_i, \qquad \forall i \in I_l, \tag{16}$$

$$-\langle w, b^j - a^l\rangle - \xi \left\|b^j - a^l\right\|_1 + \gamma + 1 \le 0, \qquad \forall j \in J, \tag{17}$$

$$y = (y_1, \ldots, y_m) \in \mathbb{R}^m_+, w \in \mathbb{R}^n, \xi \in \mathbb{R}, \gamma \ge 1. \tag{18}$$

Let $w^l, \xi^l, \gamma^l, y^l$ be a solution of (P_l). Set

$$g_l(x) := g_{(w^l,\xi^l,\gamma^l,a^l)}(x) \tag{19}$$

and go to Step 2.

Step 2: Set $I_{l+1} := \{i \in I_l : g_l(a^i) + 1 > 0\}, A_{l+1} := \{a^i \in A_l : i \in I_{l+1}\}, l := l+1$. If $A_l \neq \emptyset$ then go to Step 1.

Step 3: Define the function $g(x)$ (separating the sets A and B) as

$$g(x) = \min_l g_l(x) \tag{20}$$

and stop.

At each iteration l, the algorithm arbitrarily chooses some element a^l from the set A_l and calculates parameters (w^l, ξ^l, γ^l) by solving a linear subproblem (P_l). All these parameters are then used in (19) for defining the function g_l. It follows from Lemma 1 that the graph of the function g_l consisting of points $(x,z) \in \mathbb{R}^n \times \mathbb{R}$ with $z = g_l(x)$ is a cone with vertex at $(a^l, -\gamma^l)$. A constraint $\gamma \geq 1$ stated in constraint set (18) ensures that the vertex of this cone has to be placed "under" the hyperplane $z = 0$, that is in the half-space $\mathbb{R}^n \times (0, -\infty)$. The constraint set (16) ensures that the point a^l and all the points of the set A^l which are "close" to a^l have to be in the polyhedron corresponding to the sublevel set $\{x : g_l(x) \leq -1\}$. The closeness of these points of A_l to a^l is defined by the optimal value of the objective function (15) in (P_l). That is the sublevel set $\{x : g_l(x) \leq -1\}$ will include as much elements of A_l (besides of a^l) as the value of this objective function is close to zero. Thus the objective function (15) and the constraint sets (16) and (18) ensure that all the elements of A_l will be enclosed to the sublevel set $\{x : g_l(x) \leq -1\}$ of g_l if this minimum is zero. On the other hand the constraint set (17) ensures that all the elements of the set B have to be remained outside of the sublevel set $\{x : g_l(x) < 1\}$ at each iteration. Note that such a "separability" at each iteration becomes possible due to the characteristics of PCFs described in Lemma 1. We will call the method of separation described in the algorithm the PCF separation.

The following theorem proves that the presented algorithm terminates in a finite number of iterations and the resulting function g defined by (20) separates arbitrary disjoint sets A and B consisting of finite number of elements in R^n.

Theorem 3. *PCF Algorithm terminates in a finite number of iterations and the function $g : \mathbb{R}^n \to \mathbb{R}$ defined by (20) strictly separates the sets A and B in the sense that*

$$g(a) < 0, \ \forall a \in A, \tag{21}$$

$$g(b) > 0, \ \forall b \in B. \tag{22}$$

Proof. First show that the problem (P_l) has a solution$(w_l, \xi_l, \gamma_l) \in \mathbb{R}^n \times \mathbb{R}_+ \times [1, \infty)$ such that the corresponding function g_l separates at least one element, (say a_l) of A_l and the whole set B. By taking $w_l = 0, \gamma_l = 1$ we obtain a function $g_l(x) = \xi \|x - a^l\|_1 - 1$, for which we have $g_l(a^l) = -1 < 0$ and $g_l(b^j) = \xi \|b^j - a^l\|_1 - 1, \forall j \in J$.

Since $b^j \in B$ we have $\|b^j - a^l\|_1 > 0, \forall j \in J$. Therefore when ξ is sufficiently large, the term $\xi\|b^j - a^l\|_1$ can be made large enough. Then for ξ sufficiently large, we have $g_l(b^j) > 0, \forall j \in J$ which means that the function $g_l(x) = \xi_l\|x - a^l\|_1 - 1$ separates a^l and the set B in the sense that $g_l(a_l) < 0$ and $g_l(b) > 0, \quad \forall b \in B$.

Let $\tilde{A}_l$ be the subset of A_l consisting of elements which are separated from B by the function g_l formed using the solution of the problem (P_l) at the lth iteration, and let A_{l+1} be the subset of A_l consisting of the elements which could not be separated from B by g_l. If this set is not empty the algorithm will be continued. Since the set A has a finite number of elements, the process will be terminated after the finite number of iterations. Thus we will have a partition $\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_L$ of the set A and functions $g_1, g_2, \ldots, g_L$ with properties:

$$A = \bigcup \tilde{A}_l,$$
$$g_l(a) < 0, \forall a \in \tilde{A}_l,$$
$$g_l(b) > 0, \forall b \in B, l = 1, \ldots, L.$$

Then the function $g(x) = \min\limits_l \{g_l(x)\}$ will obviously have the properties (21) and (22). Indeed, since for every $a \in A$ there exists $l \in \{1, 2, \ldots, L\}$ such that $a \in \tilde{A}_l$, we have $g_l(a) < 0$ and therefore $g(a) = \min\limits_l \{g_l(a)\} < 0$. On the other hand, since $g_l(b) > 0$ for all $l \in \{1, 2, \ldots, L\}$, $b \in B$, we have $g(b) > 0$. □

Corollary 1. *Let A and B be two arbitrary sets consisting of finite number of points in $\mathbb{R}^n$. Then*

1. *there exists a partition of $A : A = \bigcup \tilde{A}_l$ such that, $co\tilde{A}_l \cap B = \emptyset$ and functions $g_l(x) = \langle w_l, x\rangle + \xi_l\|x\|_1 - \gamma_l$, for $l = 1, 2, \ldots$, with $g_l(a) < 0, \forall a \in co\tilde{A}_l, g_l(b) > 0$, $\forall b \in B$, and*

2. *the function $g(x) = \min\limits_l g_l(x)$ separates A and B in the sense of (21) and (22). Here co stands for convex hull of a set.*

Proof. 1. The existence of a partition $A : A = \bigcup \tilde{A}_l$ and functions $g_l(x)$ with a property

$$g_l(a) < 0, \forall a \in \tilde{A}_l, \quad g_l(b) > 0, \forall b \in B,$$

follows from the proof of Theorem 3. Let $C_l = \{x \in \mathbb{R}^n : g_l(x) \leq 0\}$. Then $\tilde{A}_l \subset C, B \subset \{x \in \mathbb{R}^n : g_l(x) > 0\}$, and $C \cap B = \emptyset$ by construction. By Lemma 1, C_l is a convex polyhedron. Since it contains $\tilde{A}_l$, it contains also co$\tilde{A}_l$—the smallest convex set containing $\tilde{A}_l$. Thus, $(\text{co}\tilde{A}_l) \cap B = \emptyset$.

2. Is obvious. □

Example 1. Consider two finite point sets A and B in $\mathbb{R}^2$ shown in Fig. 1. Note that the set A is taken to consist of two isolated parts. The coordinates x_1 and x_2 of the points described in this figure are given in Table 1.

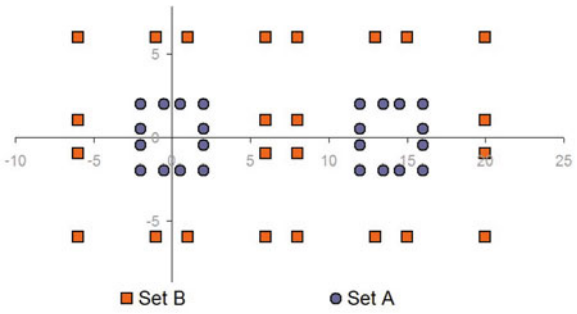


Fig. 1 Two finite point sets A and B in $\mathbb{R}^2$

Table 1 Coordinates of data points

A	1	2	3	4	5	6	7	8	9	10	11	12
x_1	−2	−2	−2	−2	−0.5	−0.5	0.5	0.5	2	2	2	2
x_2	0.5	−0.5	2	−2	2	−2	2	−2	0.5	−2	−0.5	2

	13	14	15	16	17	18	19	20	21	22	23	24
	12	12	12	12	13.5	13.5	14.5	14.5	16	16	16	16
	2	−2	0.5	−0.5	2	−2	2	−2	−0.5	0.5	2	−2

B	1	2	3	4	5	6	7	8	9	10	11	12
x_1	8	6	20	6	15	20	1	−1	−6	20	−6	−6
x_2	−6	−1	6	−6	6	1	6	6	1	−6	−1	−6

	13	14	15	16	17	18	19	20	21	22	23	24
	8	13	20	6	15	13	8	−1	−6	6	8	1
	−1	6	−1	1	−6	−6	1	−6	6	6	6	−6

This example has been solved by PCF algorithm. The geometrical interpretations for separation function obtained by PCF algorithm are presented in Fig. 2.

PCF Algorithm is applied for constructing a separation function. GAMS/CPLEX solver is used for solving the LP subproblems. Algorithm has been terminated in seven iterations. The subsets $\tilde{A}_l, l = 1,\ldots,7$ partitioning the set A and the corresponding PCFs g_l separating these sets from B at each iteration are presented below:

$$
\begin{aligned}
g_1(x) &= 0.06(x_1 - 14.5) + 0.11(x_2 - 2) + 0.34(|x_1 - 14.5| + |x_2 - 2|) - 1,\\
g_2(x) &= -0.06(x_1 + 0.5) + 0.11(x_2 - 2) + 0.34(|x_1 + 0.5| + |x_2 - 2|) - 1,\\
g_3(x) &= -0.05(x_1 - 2) + 0.23(x_2 - 0.5) + 0.46(|x_1 - 2| + |x_2 - 0.5|) - 1,\\
g_4(x) &= 0.11(x_1 - 16) + 0.02(x_2 + 0.5) + 0.34(|x_1 - 16| + |x_2 + 0.5|) - 1,\\
g_5(x) &= -0.02(x_1 - 0.5) - 0.11(x_2 + 2) + 0.34(|x_1 - 0.5| + |x_2 + 2|) - 1,\\
g_6(x) &= -0.11(x_1 + 2) - 0.06(x_2 + 0.5) + 0.34(|x_1 + 2| + |x_2 + 0.5|) - 1,\\
g_7(x) &= -0.07(x_1 - 12) - 0.26(x_2 + 2) + 0.53(|x_1 - 12| + |x_2 + 2|) - 1.7.
\end{aligned}
$$

$$\tilde{A}_1(x) = \{(16,2),(14.5,2),(16,0.5),(12,2),(13.5,2),(14.5,-2)\},$$
$$\tilde{A}_2(x) = \{(2,2),(0.5,2),(-2,2),(-0.5,2),(-2,0.5),(-0.5,-2)\},$$
$$\tilde{A}_3(x) = \{(2,0.5),(2,-2),(2,-0.5)\},$$
$$\tilde{A}_4(x) = \{(16,-2),(16,-0.5),(12,-0.5)\},$$
$$\tilde{A}_5(x) = \{(0.5,-2),(-2,-2)\},$$
$$\tilde{A}_6(x) = \{(-2,-0.5)\},$$
$$\tilde{A}_7(x) = \{(12,0.5),(12,-2),(13.5,-2)\}.$$

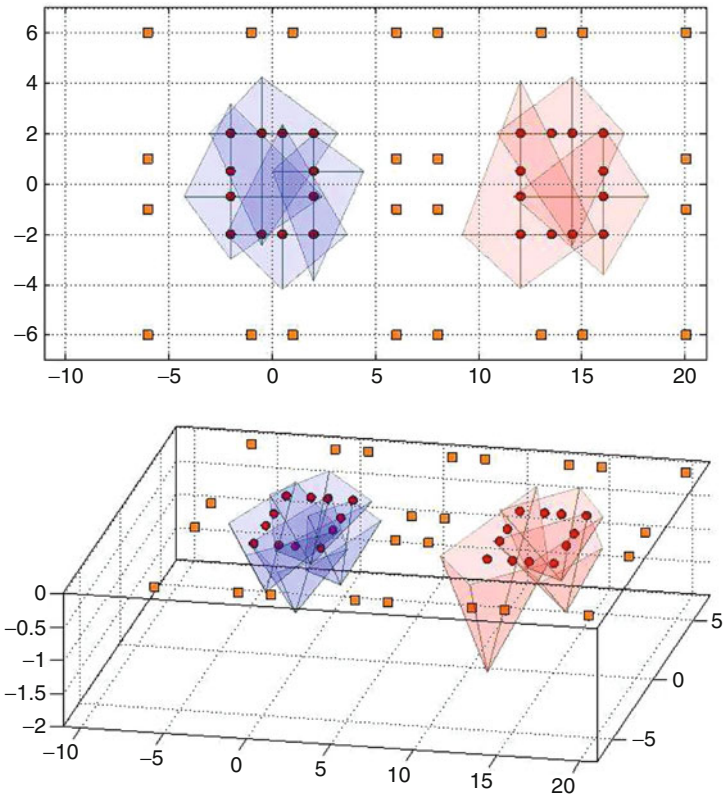


Fig. 2 Two-dimensional and three-dimensional views of polyhedral functions obtained by the first way

5 Incremental Max–Min Separability

In this section we describe an incremental algorithm for finding piecewise linear boundaries between finite sets. We assume that we are given a data set A with q classes: $A_1,\ldots,A_q$. At each iteration of the algorithm we solve problem (3) with a preset number of hyperplanes to find a piecewise linear boundary between a given class and the rest of the data set. This is done for all classes using the one vs all approach. After computing piecewise linear boundaries for all classes we define data points which can be easily classified using the piecewise linear boundaries from this iteration. Then all these points are removed before the next iteration.

The algorithm stops when there are no sets to separate (the remaining points, if any, belong to only one set). For each set the improvement in classification accuracy and the objective function value compared to the previous iteration is used as a stopping criterion for the final piecewise linear boundary between this set and the rest of the data set.

For the sake of simplicity we split the incremental algorithm into two parts: Algorithm 2 (outer) and Algorithm 3 (inner). Algorithm 2 contains the main steps of the method. These steps are the initialization of starting points, the number of hyperplanes, and the update of the set of undetermined points. Algorithm 3 is called at each iteration of Algorithm 2. It computes the piecewise linear boundaries for a given set; refines the set of undetermined points; updates starting points and the number of hyperplanes for the next iteration of Algorithm 2.

5.1 Algorithm

First, we describe the outer algorithm. Let $\varepsilon_0 > 0$ be a tolerance.

Algorithm 2. An incremental algorithm.

1: (*Initialization*) Set $A_u^1 := A_u,\ Q_u^1 := \emptyset,\ u = 1,\ldots,q$. Select any starting point (x,y) such that $x \in \mathbb{R}^n, y \in \mathbb{R}^1$, and set

$$X^{1u} := x, Y_{1u} := y, \forall\, u = 1,\ldots,q.$$

Set

$$C^1 := \{1,\ldots,q\},\ I_{1u} := \{1\}, J_1^{1u} := \{1\}, r_{1u} := 1, s_{1u}^1 := 1,\ u = 1,\ldots,q,$$

the number of hyperplanes for class u: $l_{1u} := 1$ and iteration counter $k := 1$.

2: (*Stopping criterion*) If $|C^k| \leq 1$ then stop. Otherwise go to Step 3.

3: (*Computation of piecewise linear functions*) For each $u \in C^k$ apply Algorithm 3. This algorithm generates a piecewise linear boundary (X^{ku*}, Y_{ku*}), the set of indices $I_{k+1,u}$, $J_i^{k+1,u}$, $i \in I_{k+1,u}$, a number of hyperplanes $l_{k+1,u}$, a starting point $(X^{k+1,u}, Y_{k+1,u}) \in \mathbb{R}^{(n+1)l_{k+1,u}}$ for class u, the set A_u^{k+1} containing "undetermined" points, and the set Q_u^k of easily separated points from class u.

4: (*Refinement of set C^k*) Refine the set C^k as follows:

$$C^{k+1} = \{u \in C^k : \ |A_u^{k+1}| > \varepsilon_0 |A_u|\}.$$

Set $k := k+1$ and go to Step 2.

We will now present the inner algorithm for separating class A_u, $u \in \{1,\ldots,q\}$ from the rest of the data set. At each iteration k of Algorithm 2 we get the subset $A_u^k \subseteq A_u$ of the set $u \in C^k$ which contains points from this class which are not easily separated using piecewise linear functions from previous iterations. Let

$$\overline{Q}_u^k = \bigcup_{j=1,\ldots,k} Q_u^j$$

be a set of all points removed from the set A_u during the first $k > 0$ iterations. We denote

$$D_k = \bigcup_{t=1,\ldots,q} \left(A_t \setminus \overline{Q}_t^k\right), \ \underline{A_u^k} = \bigcup_{t=1,\ldots,q, t\neq u} \left(A_t \setminus \overline{Q}_t^k\right).$$

Algorithm 3 finds a piecewise linear function separating the sets A_u^k and $\underline{A_u^k}$. Let $\varepsilon_1 > 0, \varepsilon_2 > 0, \varepsilon_3 > 0$ be given tolerances and $\sigma \geq 1$ be a given number.

Algorithm 3. Computation of piecewise linear functions.

Input: Starting points $(X^{ku}, Y_{ku}) \in \mathbb{R}^{(n+1)l_{ku}}$, the set of indices I_{ku}, J_i^{ku}, $i \in I_{ku}$, and the number of hyperplanes l_{ku} at iteration k of Algorithm 2.

Output: A piecewise linear boundary $(X^{ku*}, Y_{ku*}) \in \mathbb{R}^{(n+1)l_{ku}}$, the set of indices $I_{k+1,u}$, $J_i^{k+1,u}$, $i \in I_{k+1,u}$, a number of hyperplanes $l_{k+1,u}$, a starting point $(X^{k+1,u}, Y_{k+1,u}) \in \mathbb{R}^{(n+1)l_{k+1,u}}$ for class u, a set of undetermined points A_u^{k+1}, and a set Q_u^{k+1} of easily separated points from class u.

1: (*Finding a piecewise linear function*) Solve problem (3) over the set D_k starting from the point $(X^{ku}, Y_{ku}) \in \mathbb{R}^{(n+1)l_{ku}}$. Let (X^{ku*}, Y_{ku*}) be the solution to this problem, f_{ku}^* be the corresponding objective function value, and $f_{1,ku}^*$ and $f_{2,ku}^*$ be the values of functions f_1 and f_2, respectively. Let E_{ku} be the error rate for separating the sets A_u^k and $\underline{A_u}^k$ at iteration k over the set A, that is

$$E_{ku} = \frac{|\{a \in A_u^k : \varphi_u^k(a) > 0\} \cup \{b \in \underline{A_u^k} : \varphi_u^k(b) < 0\}|}{|A|},$$

where

$$\varphi_u^k(a) = \max_{i \in I_{ku}} \min_{j \in J_i^{ku}} \left(\langle x^{ij*}, a\rangle - y_{ij*}\right).$$

2: (*The first stopping criterion*) If $\max\{f_{1,ku}^*, f_{2,ku}^*\} \leq \varepsilon_1$ then set $A_u^{k+1} = \emptyset$, $Q_u^{k+1} = A_u \setminus \overline{Q}_u^k$ and stop. (X^{ku*}, Y_{ku*}) is the piecewise linear boundary for set A_u.

3: (*The second stopping criterion*) If $k \geq 2$ and $f^*_{k-1,u} - f^*_{ku} \leq \varepsilon_2$ then set $A^{k+1}_u = \emptyset$, $Q^{k+1}_u = \emptyset$, and stop. (X^{ku*}, Y_{ku*}) where $X^{ku*} = X^{k-1,u*}, Y_{ku*} = Y_{k-1,u*}$ is the piecewise linear boundary for the set A_u.

4: (*The third stopping criterion*) If $E_{ku} \leq \varepsilon_3$ then set $A^{k+1}_u = \emptyset$, $Q^{k+1}_u = A_u \setminus \overline{Q}^k_u$, and stop. (X^{ku*}, Y_{ku*}) is the piecewise linear boundary for the set A_u.

5: (*Refinement of sets of undetermined points*) Compute

$$f_{ku,\min} = \min_{a \in \underline{A^k_u}} \varphi^k_u(a)$$

and the following set of easily classified points by the function φ^k_u:

$$Q^{k+1}_u = \left\{ a \in A^k_u : \ \varphi^k_u(a) < \sigma f_{ku,\min} \right\}.$$

Refine the set of undetermined points from the set A_u as follows:

$$A^{k+1}_u = A^k_u \setminus Q^{k+1}_u.$$

6: (*Adding new hyperplanes*)

1. If $f^*_{1,ku} > \varepsilon_1$ then set

$$s^i_{k+1,u} = s^i_{ku} + 1, J^{k+1,u}_i = J^{ku}_i \cup \{s^i_{k+1,u}\}$$

for all $i \in I_{ku}$. Set

$$x^{ij} = x^{i,j-1,*}, y_{ij} = y_{i,j-1,*}, i \in I_{ku}, j = s^i_{k+1,u}.$$

2. If $f^*_{2,ku} > \varepsilon_1$ then set

$$r_{k+1,u} = r_{ku} + 1, I_{k+1,u} = I_{ku} \cup \{r_{k+1,u}\}, J^{k+1,u}_{r_{k+1,u}} = J^{ku}_{r_{ku}}.$$

Set

$$x^{ij} = x^{i-1,j,*}, y_{ij} = y_{i-1,j,*}, i = r_{k+1,u}, j \in J^{ku}_{r_{ku}}.$$

7: (*New starting point*) Set

$$X^{k+1,u} = (X^{ku*}, x_{ij}, \ i \in I_{k+1,u}, \ j \in J^{k+1,u}_i),$$

$$Y_{k+1,u} = (Y_{ku*}, y_{ij}, \ i \in I_{k+1,u}, \ j \in J^{k+1,u}_i),$$

$$l_{k+1,u} = \sum_{i \in I_{k+1,u}} |J^{k+1,u}_i|.$$

and go to Step 1.

5.2 Explanations to the Algorithms

The following explains Algorithm 2 in more detail. In Step 1 we initialize the starting points, the number of hyperplanes for each class, and the collection C^1 of sets to be separated. Step 2 is the stopping criterion verifying that the collection C^k contains at least two sets to separate. In the third step Algorithm 3 is called and returns piecewise linear boundaries for each set, the subsets of points not yet separated by these piecewise linear boundaries, and updated starting points for the next iteration of Algorithm 2. In Step 4 we refine the set C^k by removing sets fully separated using piecewise linear boundaries from previous and current iterations.

The following explanations clarify Algorithm 3. In Step 1 we compute a piecewise linear function with a preselected number of hyperplanes using the starting point provided by Algorithm 2. It also computes the separation error rate between a given class u and the rest of the data set. The algorithm contains three stopping criteria which are given in Steps 2–4.

- The algorithm terminates if both values $f^*_{1,ku}, f^*_{2,ku}$ for class u is less than a given tolerance $\varepsilon_1 > 0$. The last piecewise linear function for this class is accepted as a boundary between this class and the rest of the data set (Step 2).
- If $k \geq 2$ and the difference between values of the error function (for class u) in two successive iterations is less than a given tolerance $\varepsilon_2 > 0$ then the algorithm terminates. The piecewise linear function from the previous iteration is accepted as the boundary between this class and the rest of the data set (Step 3).
- Finally, if the error rate is less than a threshold $\varepsilon_3 > 0$ then the algorithm terminates and the piecewise linear function from the last iteration is accepted as the boundary between this class and the rest of the data set (Step 4).

If none of these stopping criteria is met, then in Step 5 we refine the set of undetermined points by removing points easily separated using the piecewise linear functions from the current iteration. In Step 6, depending on the values of the error function on both sets, we may add new hyperplanes. Finally in Step 7 we update the starting point and the number of hyperplanes.

As an illustration Fig. 3 shows the result of the first iteration of Algorithm 2 for a data set with three classes A_1, A_2, and A_3. At this iteration we compute one hyperplane for each set. The data set in its original form is illustrated in Fig. 3a. We select any starting point in Step 1 of Algorithm 2 and then call Algorithm 3 in Step 3. Algorithm 3 computes one linear function for each class using one vs all strategy. A hyperplane given in Fig. 3b presents the linear function separating the class A_1 from the rest of the data set with the minimum error function value. This hyperplane is computed in Step 1 of Algorithm 3. Then in Step 5 of Algorithm 3 we compute a hyperplane (with dashed lines in Fig. 3c, here $\sigma = 1$) by translating the best hyperplane so that beyond this dashed line only points from the class A_1 lie. We remove all points from the class A_1 which lie beyond this line before the next iteration (Step 5 of Algorithm 3) and do not consider them in the following iterations. These data points can be easily classified using linear separation. We repeat the

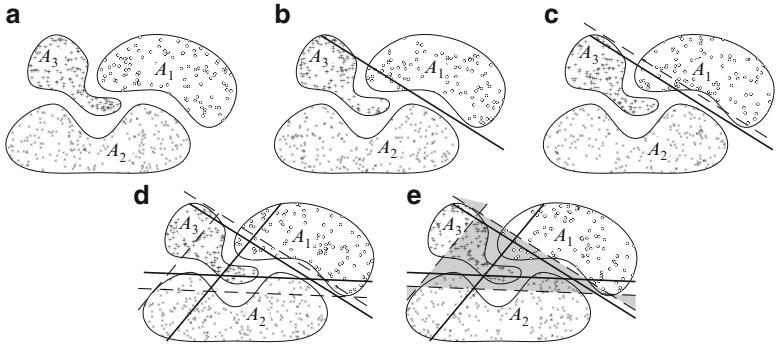


Fig. 3 The first iteration of Algorithm 2 for three sets A_1, A_2, and A_3

same computation for other classes A_2 and A_3 and remove all data points which can be classified using linear functions (see Fig. 3d). Then we compute all data points which lie in the grey area in Fig. 3e. These points cannot be determined by linear separators and we use only these points to compute piecewise linear boundaries in the next iteration of Algorithm 2.

6 The HPCAMS Algorithm

Algorithm 2 allows one to find piecewise linear boundaries between pattern classes. At each iteration the function (2) is minimized. The complexity of the computation of this function depends on the number of data points. However, in large data sets many data points lie far away from other classes. Therefore they are not relevant to the computation of the boundary between their class and other classes. In this section we propose Algorithm 4 where one PCF is used for each class in order to eliminate those irrelevant data points before applying Algorithm 2. First, we will explain each step of Algorithm 4 and then formulate it at the end of this section.

If we fix the point c, then the finding one PCF is a linear programming problem. Furthermore, if this point is not fixed, then the PCF may eliminate points which are close to other classes. It is preferable to select a data point which is far away from the boundary of its associated class as c. In order to find such a point we propose to use hyperboxes approximating classes. We select c lying inside only one hyperbox when possible (Step 1) and then we solve the problem (15)–(18). As a result we find a PCF approximating the interior of the classes (Step 2). Then we eliminate those points from the data set and apply Algorithm 2 to the remaining points (Step 3).

In the sequel we explain each step of Algorithm 4 in more detail.

6.1 Computation of Vertices of Polyhedral Conic Sets

In Step 1 of Algorithm 4 we approximate each class by one hyperbox.

Assume that we are given data set A with $q \geq 2$ classes $A_1, \ldots, A_q$. For each class A_i we compute:

$$\bar{\alpha}_j^i = \min_{a \in A_i} a_j, \;\; \bar{\beta}_j^i = \max_{a \in A_i} a_j, \;\; j = 1, \ldots, n, \;\; i = 1, \ldots, q$$

and define vectors $\bar{\alpha}^i = (\bar{\alpha}_1^i, \ldots, \bar{\alpha}_n^i)$, $\bar{\beta}^i = (\bar{\beta}_1^i, \ldots, \bar{\beta}_n^i)$, $i = 1, \ldots, q$ which in turn define the following hyperboxes in n-dimensional space $\mathbb{R}^n$ for $i = 1, \ldots, q$:

$$\bar{H}(A_i) = [\bar{\alpha}^i, \bar{\beta}^i] \equiv \left\{ x \in \mathbb{R}^n : \;\; \bar{\alpha}_j^i \leq x_j \leq \bar{\beta}_j^i, \;\; j = 1, \ldots, n \right\}.$$

All points from the i-th class belong to the hyperbox $\bar{H}(A_i)$.

To ensure that the first components of the vertices of the polyhedral conic sets lie inside the classes we take a sufficiently small $\eta > 0$ and consider the following extended hyperbox for each class A_i, $i = 1, \ldots, q$:

$$H(A_i) = [\alpha^i, \beta^i] \equiv \left\{ x \in \mathbb{R}^n : \;\; \alpha_j^i \leq x_j \leq \beta_j^i, \;\; j = 1, \ldots, n \right\},$$

where for $j = 1, \ldots, n$

$$\alpha_j^i = \bar{\alpha}_j^i - \eta(\bar{\beta}_j^i - \bar{\alpha}_j^i), \quad \beta_j^i = \bar{\beta}_j^i + \eta(\bar{\beta}_j^i - \bar{\alpha}_j^i).$$

The hyperbox $H(A_i)$ can be described as

$$H(A_i) = \{ x \in \mathbb{R}^n : \;\; \psi_i(x) \leq 0 \},$$

where the piecewise linear function $\psi_i(x)$ is defined as follows:

$$\psi_i(x) = \max \left\{ \alpha_j^i - x_j, x_j - \beta_j^i, \;\; j = 1, \ldots, n \right\}.$$

In order to find the vertex for the i-th polyhedral conic set we define the set

$$R_i = \left\{ a \in A_i : \min_{k=1,\ldots,q, \;\; k \neq i} \psi_k(a) > 0 \right\}.$$

This set contains all points from the i-th class which are outside hyperboxes of all other classes. First we consider the case when the set $R_i \neq \emptyset$. Figures 4 and 5 illustrate this case. We compute

$$\bar{Q}_1 = \min_{a \in R_i} \psi_i(a)$$

and choose c^i as follows:

$$c^i \in R_i, \psi_i(c^i) = \bar{Q}_1.$$

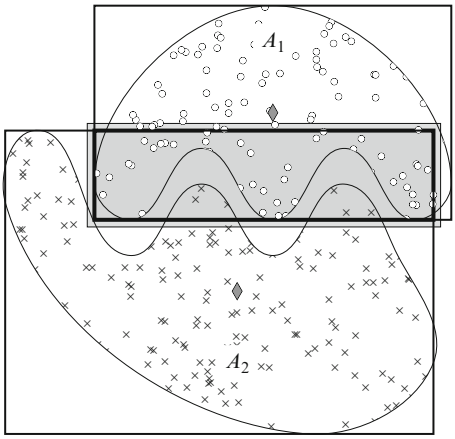


Fig. 4 Identification of the vertices of the polyhedral conic sets for the two classes A_1 and A_2 using hyperboxes

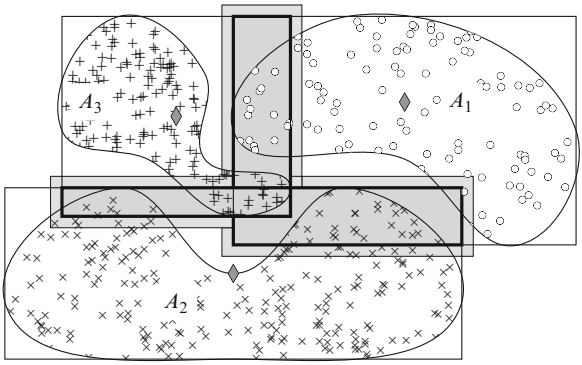


Fig. 5 Identification of the vertices of the polyhedral conic sets for the three classes A_1, A_2, and A_3 using hyperboxes

If $R_i = \emptyset$ then for any $a \in A_i$

$$\min_{k=1,\ldots,q,\ k\neq i} \psi_k(a) \leq 0.$$

In this case we compute

$$\bar{Q}_2 = \max_{a\in A_i} \min_{k=1,\ldots,q,\ k\neq i} \psi_k(a)$$

and choose c^i as follows:

$$c^i \in A_i, \min_{k=1,\ldots,q,\ k\neq i} \psi_k(c^i) = \bar{Q}_2.$$

6.2 Identification of Boundary Points

To identify boundary points we solve the problem (15)–(18). After solving it we find the values for the vector w^i and scalars $\xi_i,\ \gamma_i$ which define PCFs g_i for each class $i = 1,\dots,q$. Then for a given class $i \in \{1,\dots,q\}$ we compute the following:

$$\bar{\delta}_i = \min_{j=1,\dots,q, j\neq i} \min_{b\in A_j} g_i(b).$$

Consider the level sets of the function $g_i,\ i = 1,\dots,q$:

$$S_i(\delta) = \{x \in \mathbb{R}^n :\ g_i(x) \leq \delta\},\ \delta \in \mathbb{R}.$$

If $\bar{\delta}_i > 0$ then the set $S_i(0)$ does not contain points from any other classes, it only contains points from the i-th class. If $\bar{\delta}_i \leq 0$ then the set $S_i(0)$ also contains points from other classes. Therefore we replace $\bar{\delta}_i$ by $\hat{\delta}_i = \min\{0, \bar{\delta}_i\}$. To ensure that boundary points are not removed, $\hat{\delta}_i$ is again replaced by the following number:

$$\delta_i = \hat{\delta}_i - \theta(|\gamma_i| - \hat{\delta}_i)$$

where $\theta > 0$ is a sufficiently small number. For each class i we can define the following sets:

$$D_i = \{a \in A_i : g_i(a) \leq \delta_i\},\ i = 1,\dots,q. \tag{23}$$

The set D_i approximates the interior of the i-th class and it does not contain points from other classes. We then define the set of boundary points as follows:

$$B_i = A_i \setminus D_i,\ i = 1,\dots,q. \tag{24}$$

Let $\sigma \in (0,1)$ be a sufficiently small number. For each class $i = 1,\dots,q$ we introduce the following number:

$$r_i = |B_i|/|A_i|$$

and then we consider the set

$$P = \{i = 1,\dots,q :\ r_i > \sigma\}. \tag{25}$$

If $P = \emptyset$ then classes $A_i,\ i = 1,\dots,q$ can be approximated by their corresponding sets D_i with the sufficiently small error σ. Otherwise we can apply Algorithm 2 over sets $B_i,\ i \in P$ to find piecewise linear boundaries between classes.

6.3 Outline of the Algorithm

In summary an algorithm for finding piecewise linear boundaries between classes $A_i,\ i = 1,\dots,q$ can be formulated as follows:

Algorithm 4. Computation of piecewise linear boundaries.

1: (*Finding a vertex of a polyhedral conic set*) Approximate each class $i = 1, \ldots, q$ with the hyperbox $H(A_i)$ and compute the point c^i of the corresponding polyhedral conic set (see Sect. 6.1).

2: (*Identifying boundary points*) Compute polyhedral conic sets by solving the problem (15)–(18) and using the points $c^i,\ i = 1, \ldots, q$. Find the sets $D_i,\ i = 1, \ldots, q$ using (23) and the sets $B_i,\ i = 1, \ldots, q$ of boundary points using (24) (see Sect. 6.2). Compute the set P using (25). If $|P| \leq 1$ then stop. Otherwise go to Step 3.

3: (*Finding piecewise linear boundaries*) Apply Algorithm 2 over sets $B_i,\ i \in P$ to find piecewise linear boundaries between classes (see Sect. 5).

Algorithm 4 is illustrated in Fig. 6.

We call this algorithm the HPCAMS algorithm. This algorithm generates one PCF g_i for each class $i = 1, \ldots, q$. It also generates piecewise linear functions φ_i for classes $i \in P$ when $|P| > 1$. If $i \notin P$ then we set $\varphi_i(x) \equiv +\infty$. If $|P| \leq 1$ then we set $\varphi_i(x) \equiv +\infty$ for all $i = 1, \ldots, q$. Then the function Ψ_i separating the i-th class from the rest of the data set can be computed as follows:

$$\Psi_i(x) = \min\{g_i(x),\ \varphi_i(x)\},\ i = 1, \ldots, q. \tag{26}$$

6.4 Implementation of the Algorithm

In this subsection we describe the conditions for the implementation of Algorithm 4. As mentioned earlier this algorithm consists of two stages. In the first stage we compute PCFs approximating classes (Steps 1 and 2). There are three tolerances in this stage, in Step 1 $\eta > 0$ and in Step 2 $\theta > 0$ and $\sigma > 0$. We take $\eta = 0.1,\ \theta = 0.05$, and $\sigma = 0.01$.

In the second stage we apply Algorithm 2 to find piecewise linear boundaries (Step 3). This algorithm contains one tolerance $\varepsilon_0 \geq 0$. We choose $\varepsilon_0 = 0.01$. The following conditions have been chosen for the implementation of Algorithm 3.

1. The values of tolerances $\varepsilon_1 > 0, \varepsilon_2 > 0$, and $\varepsilon_3 > 0$ are:

$$\varepsilon_1 = 0.005,\ \varepsilon_2 = f_1^*/100,\ \varepsilon_3 = 0.001,$$

   where f_1^* is the optimal value of the objective function for linear separation.
2. We restrict the number of hyperplanes to 10.
3. In Step 1 of Algorithm 3 we use the discrete gradient method of [8, 9] as modified in [11] to solve minimization problem (3).

We implemented the algorithm in Fortran 95 and compiled it using the Lahey Fortran compiler on a 1.83 GHz Intel Pentium IV CPU with 1 GB of RAM running Windows XP.

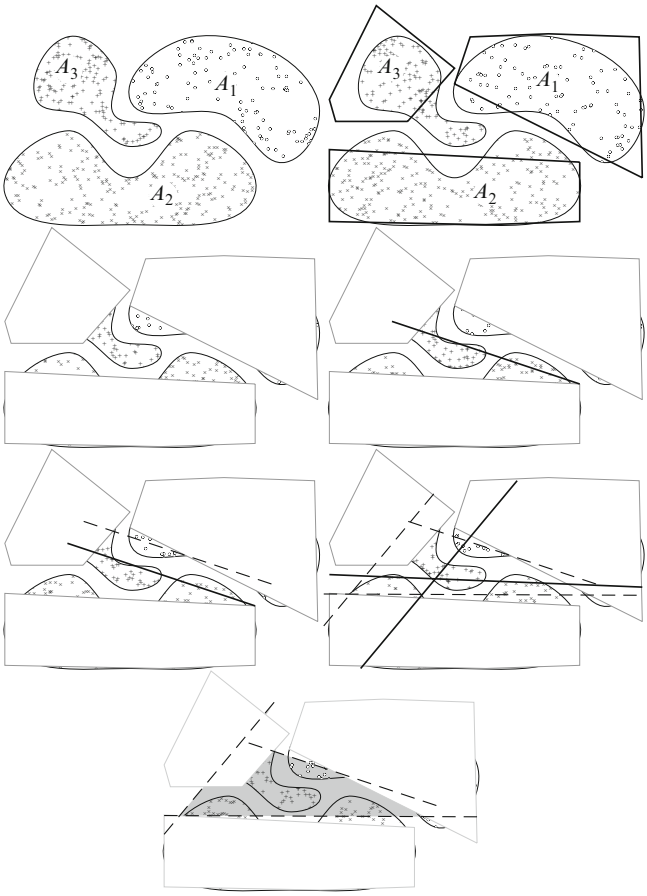


Fig. 6 Algorithm 4 for three sets A_1, A_2, and A_3

6.5 Classification Rules

To compute piecewise linear boundaries between classes we use the one vs all strategy, that is, for each class i we consider this class as one class and the rest of the data set as a second class. Then we apply Algorithm 4 to separate the i-th class from the rest of the data set. This means that for each class i Algorithm 4 generates the separating function Ψ_i defined in (26). Then we can apply the following classification rule to classify new data points (observations). If the new point v belongs to the set D_i then we classify it to the i-th class. If this point does not belong to any of the sets D_i, $i = 1,\ldots,q$ then we compute the values $\Psi_1(v),\ldots,\Psi_q(v)$ and classify this point to the class i associated with the minimum function value: $i = \text{argmin}\{\Psi_1(v),\ldots,\Psi_q(v)\}$.

7 Computational Results

We tested the HPCAMS algorithm on medium sized and large-scale real-world data sets readily available from the UCI machine learning repository [5]. The selected data sets contain either continuous or integer attributes and have no missing values. Table 2 contains a brief description of the characteristics of the data sets. This table contains the number of data points in training and test sets. The class attribute is included in the number of attributes in this table.

Table 2 Brief description of data sets

Data sets	(Train, test)	No. of attributes	No. of classes
Shuttle control (SH)	(43,500, 14,500)	10	7
Letter recognition (LET)	(15,000, 5,000)	17	26
Landsat satellite image (LSI)	(4,435, 2,000)	37	6
Pen-based recognition of handwritten digits (PD)	(7,494, 3,498)	17	10
Page blocks (PB)	(4,000, 1,473)	11	5
Optical recognition of handwritten digits (OD)	(3,823, 1,797)	65	10
Spambase (SB)	(3,682, 919)	58	2
Abalone (AB)	(3,133, 1,044)	9	3
DNA	(2,000, 1,186)	180	3
Isolet (ISO)	(6,238, 1,559)	618	26
Phoneme_CR (PHON)	(4,322, 1,082)	6	2
Texture_CR (TEXT)	(4,400, 1,100)	41	11

In our experiments we used some classifiers from WEKA (Waikato Environment for Knowledge Analysis) for a comparison. WEKA is a popular machine learning suite for data mining tasks written in Java and developed at the University of Waikato, New Zealand (see for details [42]). We chose representatives of each type of classifier from WEKA: Naive Bayes (with kernel) (NB kernel), Logistic, Multi-Layer Perceptron (MLP), Linear LIBSVM (LIBSVM (LIN)), support vector machines classifier SMO with normalized polynomial kernel (SMO (NPOL)), SMO (PUK), a decision tree classifier J48 (which is an implementation of the C4.5 algorithm), and a rule-based classifier PART. The classifiers chosen produced an overall better accuracy than other classifiers. We also include the original incremental max–min separability algorithm (CIMMS) from Sect. 5 in our experiments.

We apply all algorithms from WEKA with the default parameter values. We put the following limits: 3 h of CPU time (for training and testing) and 1 GB of memory usage. In the tables a dash line shows that an algorithm exceeded one of these limits.

Tables 3 and 4 contain test set accuracy on different data sets using different classifiers. One can see that in most of the data sets (except Optical recognition of handwritten digits, Phoneme_CR, Landsat satellite image, Isolet and Page blocks) the classification accuracy achieved over the test set by the HPCAMS algorithm is either the best or comparable with the best accuracy.

Table 3 Test set accuracy for different classifiers

Data set	AB	DNA	LSI	LET	OD	PD
Classifier						
NB(kernel)	57.85	93.34	82.10	74.12	90.32	84.13
Logistic	64.27	88.36	83.75	77.40	92.21	92.85
MLP	63.51	93.68	88.50	83.20	96.55	89.85
LIBSVM (LIN)	60.73	93.09	85.05	82.40	96.55	95.00
SMO (NPOL)	60.25	95.36	79.60	82.34	96.66	96.86
SMO (PUK)	64.18	57.93	91.45	–	96.61	97.88
J48	60.15	92.50	85.35	87.70	85.75	92.05
PART	57.95	91.06	85.25	87.32	89.54	93.65
CIMMS	65.80	93.42	88.15	91.90	94.27	96.63
HPCAMS	66.09	94.18	87.15	91.04	93.10	96.57

Table 4 Test set accuracy for different classifiers (cont.)

Data set	PHON	SH	TEXT	ISO	PB	SB
Classifier						
NB(kernel)	76.53	98.32	81.00	–	88.39	76.17
Logistic	74.58	96.83	99.64	–	91.72	92.06
MLP	81.52	99.75	99.91	–	92.80	92.06
LIBSVM (LIN)	77.54	–	99.18	96.02	87.03	90.97
SMO (NPOL)	78.74	96.81	97.27	–	89.48	92.60
SMO (PUK)	83.27	99.50	99.55	–	88.53	93.04
J48	85.67	99.95	93.91	83.45	93.55	92.93
PART	82.72	99.98	93.82	82.81	92.46	91.40
CIMMS	81.05	99.84	99.82	95.19	87.10	93.80
HPCAMS	80.13	99.86	99.36	93.52	89.55	93.47

Table 5 presents pairwise comparison of the HPCAMS classifier with other classifiers using test set accuracy. The table contains the number of data sets and their proportion where the HPCAMS algorithm achieved better testing accuracy. These results demonstrate that the HPCAMS algorithm performs well on test set in comparison with other classifiers. Comparison with the CIMMS algorithm shows that on some data sets boundary between classes is highly nonlinear and application of PCFs may remove some points from the boundary. In such cases CIMMS algorithm achieves better accuracy than the HPCAMS algorithm. However, in most cases the difference in accuracy is less than 1 %.

Table 4 presents training and testing time for the HPCAMS algorithm and training time for the CIMMS algorithm. Results demonstrate that the use of PCFs allows us to significantly reduce training time on data sets with a large number of data points. However, the HPCAMS algorithm still requires a longer training time than most of the other tested classifiers. The proposed algorithm is very fast in testing phase for all data sets. Results show that testing of the new algorithm is similar to that of Neural Network classifier MLP, Logistic classifier, and CIMMS. Decision tree and rule-based classifiers use more testing time than the proposed algorithm. SVM algorithms use 1–2 order more testing time than the HPCAMS algorithm.

Table 5 Pairwise comparison of the HPCAMS classifier with others using testing accuracy

Classifier	No. of data sets	Proportion (%)
NB(kernel)	12	100
Logistic	10	83.33
MLP	7	58.33
LIBSVM(LIN)	10	83.33
SMO (NPOL)	9	75.00
SMO (PUK)	7	58.33
J48	9	75.00
PART	8	75.00
CIMMS	4	33.33

Table 6 Training and testing time (in seconds)

Data set	Training time CIMMS	Training time HPCAMS	Testing time HPCAMS
AB	27.22	37.03	0.00
DNA	32.06	42.27	0.03
LSI	523.28	451.67	0.03
LET	9,941.34	7,389.73	0.16
OD	81.88	106.31	0.05
PD	203.02	158.91	0.03
PHON	34.75	39.13	0.00
SH	782.47	731.70	0.03
TEXT	47.28	55.33	0.02
ISO	3,927.3	2,994.64	1.86
PB	27.63	89.55	0.02
SB	295.23	240.20	0.02

It should be noted that in order to implement the HPCAMS classifier it is sufficient to save in memory one polyhedral conic and one piecewise linear functions for each class. Therefore the memory usage of the HPCAMS classifier is very low.

8 Conclusion

In this chapter we have developed a new algorithm for the computation of piecewise linear boundaries between pattern classes. This algorithm consists of two main stages. In the first stage we compute one PCF for each class in order to identify data points which lie on or close to the boundaries between classes. In the second stage we apply the max–min separability algorithm to find piecewise linear boundaries using only those data points. Such an approach allows us to reduce the training time of the max–min separability algorithm from [10] on large data sets by 3–10 times. The new algorithm provides almost instantaneous testing and has a low memory usage. We present the results of numerical experiments. These results demonstrate that the proposed algorithm consistently produces a good test set accuracy on most data sets when comparing with a number of other mainstream classifiers. However, the proposed algorithm requires more training time than most of the other classifiers.

Acknowledgements R. Kasimbeyli and G. Özturk are the recipients of a grant of the Scientific and Technological Research Council of Turkey—TUBITAK (Project number:107M472).

References

1. Astorino, A., Fuduli, A.: Nonsmooth optimization techniques for semisupervised classification. IEEE Trans. Pattern Anal. Mach. Intell. **29**(12), 2135–2142 (2007)
2. Astorino, A., Gaudioso, M.: Polyhedral separability through successive LP. J. Optim. Theory Appl. **112**(2), 265–293 (2002)
3. Astorino, A., Fuduli, A., Gorgone, E.: Non-smoothness in classification problems. Optim. Methods Softw. **23**(5), 675–688 (2008)
4. Astorino, A., Fuduli, A., Gaudioso, M.: DC models for spherical separation. J. Glob. Optim. **48**(4), 657–669 (2010)
5. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine. http://www.ics.uci.edu/~mlearn/MLRepository.html (2007)
6. Azimov, A.Y., Gasimov, R.N.: On weak conjugacy, weak subdifferentials and duality with zero gap in nonconvex optimization. Int. J. Appl. Math. **1**, 171–192 (1999)
7. Azimov, A.Y., Gasimov, R.N.: Stability and duality of nonconvex problems via augmented Lagrangian. Cybern. Syst. Anal. **3**, 120–130 (2002)
8. Bagirov, A.M.: Minimization methods for one class of nonsmooth functions and calculation of semi-equilibrium prices. In: Eberhard, A., et al. (eds.) Progress in Optimization: Contribution from Australasia, pp. 147–175. Kluwer, Boston (1999)
9. Bagirov, A.M.: A method for minimization of quasidifferentiable functions. Optim. Methods Softw. **17**(1), 31–60 (2002)
10. Bagirov, A.M.: Max-min separability. Optim. Methods Softw. **20**(2–3), 271–290 (2005)
11. Bagirov, A.M., Ugon, J.: Supervised data classification via max-min separability. In: Jeyakumar, V., Rubinov, A.M. (eds.) Continuous Optimisation: Current Trends and Modern Applications, Chap. 6, pp. 175–208. Springer, Berlin (2005)
12. Bagirov, A.M., Ugon, J.: Piecewise partially separable functions and a derivative-free algorithm for large scale nonsmooth optimization. J. Glob. Optim. **35**(2), 163–195 (2006)

13. Bagirov, A.M., Yearwood, J.: A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems. Eur. J. Oper. Res. **170**(2), 578–596 (2006)
14. Bagirov, A.M., Ugon, J., Webb, D.: An efficient algorithm for the incremental construction of a piecewise linear classifier. Inf. Syst. **36**(4), 782–790 (2011)
15. Bagirov, A.M., Ugon, J., Webb, D., Karasozen, B.: Classification through incremental max-min separability. Pattern Anal. Appl. **14**, 165–174 (2011)
16. Bagirov, A.M., Ugon, J., Webb, D., Ozturk, G., Kasimbeyli, R.: A novel piecewise linear classifier based on polyhedral conic and max-min separabilities. TOP **21**(1), 3–24 (2013)
17. Bennett, K.P., Mangasarian, O.L.: Bilinear separation of two sets in *n*-space. Comput. Optim. Appl. **2**, 207–227 (1993)
18. Bobrowski, L.: Design of piecewise linear classifiers from formal neurons by a basis exchange technique. Pattern Recognit. **24**(9), 863–870 (1991)
19. Chai, B., Huang, T., Zhuang, X., Zhao, Y., Sklansky, J.: Piecewise linear classifiers using binary tree structure and genetic algorithm. Pattern Recognit. **29**(11), 1905–1917 (1996)
20. Gasimov, R.N.: Characterization of the Benson proper efficiency and scalarization in nonconvex vector optimization. Multiple Criteria Decision Making in the New Millennium. In: Lecture Notes in Economics and Mathematical Systems, vol. 507, pp. 189–198. Springer, New York (2001)
21. Gasimov, R.N.: Augmented Lagrangian duality and nondifferentiable optimization methods in nonconvex programming. J. Glob. Optim. **24**, 187–203 (2002)
22. Gasimov, R.N., Ozturk, G.: Separation via polyhedral conic functions. Optim. Methods Softw. **21**(4), pp. 527–540 (2006)
23. Gasimov, R.N., Rubinov, A.M.: On augmented Lagrangians for optimization problems with a single constraint. J. Glob. Optim. **28**(2), 153–173 (2004)
24. Herman, G.T., Yeung, K.T.D.: On piecewise-linear classification. IEEE Trans. Pattern Anal. Mach. Intell. **14**(7), 782–786 (1992)
25. Kasimbeyli, R.: A nonlinear cone separation theorem and scalarization in nonconvex vector optimization. SIAM J. Optim. **20**(3), 1591–1619 (2010)
26. Kasimbeyli, R., Mammadov, M.: On weak subdifferentials, directional derivatives and radial epiderivatives for nonconvex functions. SIAM J. Optim. **20**(2), 841–855 (2009)
27. Kostin, A.: A simple and fast multi-class piecewise linear pattern classifier. Pattern Recognit. **39**, 1949–1962 (2006)
28. Lo, Z.-P., Bavarian, B.: Comparison of a neural network and a piecewise linear classifier. Pattern Recognit. Lett. **12**(11), 649–655 (1991)
29. Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds.) Machine Learning, Neural and Statistical Classification. Ellis Horwood, London (1994)
30. Pallaschke, D., Rolewicz, S.: Foundations of Mathematical Optimization (Convex Analysis Without Linearity). Kluwer, Dordrecht (1997)
31. Palm, H.C.: A new piecewise linear classifier. In: Pattern Recognition, Proceedings of the 10th International Conference on Machine Learning, vol. 1. pp. 742–744 (1990)
32. Park, Y., Sklansky, J.: Automated design of multiple-class piecewise linear classifiers. J. Classif. **6**, 195–222 (1989)
33. Rubinov, A.M.: Abstract Convexity and Global Optimization. Kluwer, Dordrecht (2000)
34. Rubinov, A.M., Gasimov, R.N.: The nonlinear and augmented Lagrangians for nonconvex optimization problems with a single constraint. Appl. Comput. Math. **1**(2), 142–157 (2002)
35. Rubinov, A.M., Gasimov, R.N.: Strictly increasing positively homogeneous functions with application to exact penalization. Optimization **52**(1), 1–28 (2003)
36. Rubinov, A.M., Yang, X.Q., Bagirov, A.M., Gasimov, R.N.: Lagrange-type functions in constrained optimization. J. Math. Sci. **115**(4), 2437–2505 (2003)
37. Schulmeister, B., Wysotzki, F.: The piecewise linear classifier DIPOL92. In: Bergadano, F., De Raedt, L. (eds.) Proceedings of the European Conference on Machine Learning (Catania, Italy), pp. 411–414. Springer, New York/Secaucus (1994)
38. Singer, I.: Abstract Convex Analysis. Wiley, New York (1997)

39. Sklansky, J., Michelotti, L.: Locally trained piecewise linear classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **2**(2), 101–111 (1980)
40. Sklansky, J., Wassel, G.S.: Pattern Classifiers and Trainable Machines. Springer, Berlin (1981)
41. Tenmoto, H., Kudo, M., Shimbo, M.: Piecewise linear classifiers with an appropriate number of hyperplanes. Pattern Recognit. **31**(11), 1627–1634 (1998)
42. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

Variational Inequality Models Arising in the Study of Viscoelastic Materials

Oanh Chau, Daniel Goeleven, and Rachid Oujja

1 Thermo-Viscoelastic Models

Because of their considerable impact in everyday life and their multiple open problems, contact mechanics still remain a rich and fascinating domain of challenge. The literature devoted to various aspects of the subject is considerable, it concerns the modelling, the mathematical analysis as well as the numerical approximation of the related problems.

For example, many food materials used in process engineering are viscoelastic [18] and consequently, mathematical models can be very helpful in understanding various problems related to the product development, packing, transport, shelf life testing, thermal effects, and heat transfer. It is thus important to study mathematical models that can be used to describe the dynamical behavior of a given viscoelastic material subjected to various highly nonlinear and even non-smooth phenomena like contact, friction, and thermal effects.

A panoply of tools and approaches are needed to face the multiple difficulties of the problems. Abstract functional nonlinear analysis could be found in [5, 7, 14, 20, 26]. An early attempt at the study of contact problems for elastic and viscoelastic materials within the applied mathematical analysis framework was introduced in the pioneering reference works [12, 13, 19, 22]. For the error estimates analysis and numerical approximation, the reader can refer to [11, 15, 17]. Further extensions to non-convex contact conditions with non-monotone and possible multi-valued constitutive laws led to the active domain of non-smooth mechanic within the framework of the so-called hemivariational inequalities, for a mathematical as well as mechanical treatment we refer to [16, 23]. Finally, the mathematical, mechanical, and numerical state of the art can be found in the proceedings [25].

O. Chau • D. Goeleven • R. Oujja (✉)
University of La Réunion, PIMENT EA4518, 97715 Saint-Denis Messag, cedex 9 La Réunion, France
e-mail: oanh.chau@univ-reunion.fr; daniel.goeleven@univ-reunion.fr; rachid.oujja@univ-reunion.fr

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_2,

The basic mechanical contact problem is the following. We consider a deformable body which occupies a bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 1$ or 2 or 3), with a Lipschitz boundary Γ and let ν denote the unit outer normal on Γ. The body is acted upon by given forces and tractions. As a result, its mechanical state evolves over the time interval $[0,T]$, $T > 0$. We assume that the boundary Γ of Ω is partitioned into three disjoint measurable parts Γ_1, Γ_2, and Γ_3. The body is clamped on $\Gamma_1 \times (0,T)$. Here, we are interested in various natures of Γ_1. In case of classical fixed condition, the property $\text{meas}(\Gamma_1) > 0$ holds (see [1, 3, 8, 10]), which allows to use the well-known Korn's inequality. In case of a free boundary, we consider $\text{meas}(\Gamma_1) = 0$, where Γ_1 is reduced to one point or eventually may be an empty set. This last case presents a source of additional difficulties and new approach is necessary (see [2, 10]). We suppose also that surface tractions of density f_2 act on $\Gamma_2 \times (0,T)$. The solid is in frictional contact with a rigid obstacle on $\Gamma_3 \times (0,T)$, where various contact conditions may be considered. Moreover, a volume force of density f_0 acts on the body in $\Omega \times (0,T)$ (see Fig. 1).

In this paper, $u = (u_i)$ denotes the displacement field, $\sigma = (\sigma_{ij})$ is the stress field, and $\varepsilon(u) = (\varepsilon_{ij}(u))$ denotes the linearized strain tensor.

In what follows, for simplification, we don't indicate explicitly the dependence of functions with respect to $x \in \Omega \cup \Gamma$ and $t \in [0,T]$. Everywhere in the sequel, the indexes i and j run from 1 to d, summation over repeated indices is implied, and the index that follows a comma represents the partial derivative with respect to the corresponding component of the independent variable. Moreover the dot above represents the time derivative, i.e.,

$$\dot{u} = \frac{du}{dt}, \qquad \ddot{u} = \frac{d^2u}{dt^2}.$$

Let us denote the mass density by $\rho : \Omega \longrightarrow \mathbb{R}_+$. The dynamical evolution of the body is described by the following equation of motion

$$\rho\,\ddot{u} = \text{Div}\ \sigma + f_0 \quad \text{in} \quad \Omega \times (0,T).$$

Here $\ddot{u}$ represents the acceleration of the dynamical process.

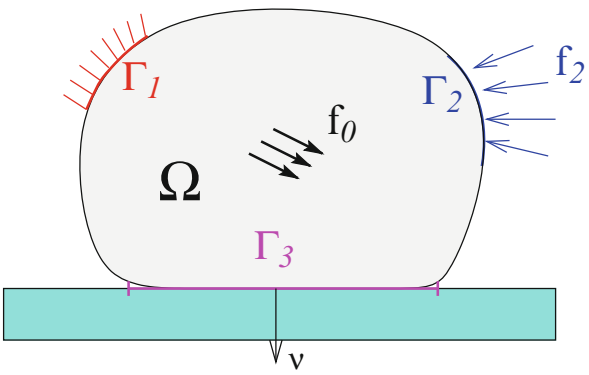


Fig. 1 The mechanical contact problem

1.1 Thermo-Viscoelastic Constitutive Law

For viscoelastic materials, the body follows a constitutive law of Kelvin–Voigt's type in the form

$$\sigma(t) = \mathcal{A}\varepsilon(\dot{u}(t)) + \mathcal{G}\varepsilon(u(t)),$$

where $\mathcal{A}$ and $\mathcal{G}$ are generally nonlinear functions, $\mathcal{A}$ represents the viscosity operator, and $\mathcal{G}$ the elasticity operator.

In the case of linear Kelvin–Voigt constitutive law, we have

$$\sigma_{ij} = a_{ijkl}\,\varepsilon_{kl}(\dot{u}) + g_{ijkl}\,\varepsilon_{kl}(u),$$

where $\mathcal{A} = (a_{ijkl})$ is the viscosity tensor and $\mathcal{G} = (g_{ijkl})$ the elasticity tensor.

This last law is qualified as of short memory, for it is instantaneous and takes place at each time t.

The long memory viscoelastic constitutive law is defined by

$$\sigma(t) = \mathcal{A}\varepsilon(\dot{u}(t)) + \mathcal{G}\varepsilon(u(t)) + \int_0^t \mathcal{B}(t-s)\,\varepsilon(u(s))\,ds \quad \text{in} \quad \Omega.$$

Here $\mathcal{B}$ is the so-called tensor of relaxation which defines the long memory behavior of the material. The above convolution term represents a kind of sum of all the elasticity of the body through the past, from the initial time to the present time. Of course, as a particular case, when $\mathcal{B} \equiv 0$, we recover the usual visco-elasticity of short memory.

In order to complete the last law with some additional thermal effects, we consider the following Kelving–Voigt's long memory thermo-viscoelastic constitutive law

$$\sigma(t) = \mathcal{A}\varepsilon(\dot{u}(t)) + \mathcal{G}\varepsilon(u(t)) + \int_0^t \mathcal{B}(t-s)\,\varepsilon(u(s))\,ds - \theta(t)C_e \quad \text{in} \quad \Omega,$$

where $C_e := (c_{ij})$ represents the thermal expansion tensor and θ is the temperature field.

1.2 The Temperature Field

We suppose that the evolution of the temperature field θ is governed by the heat equation (see [7, 8]) obtained from the conservation of energy and defined by the following differential equation

$$\dot{\theta} - \operatorname{div}(K\nabla\theta) = r(\dot{u}(t)) + q(t),$$

where $K = (k_{ij})$ represents the thermal conductivity tensor, $\mathrm{div}(K\nabla\theta) = (k_{ij}\theta_{,j})_{,i}$, $q(t)$ the density of volume heat sources, and $r(\dot{u}(t))$ a nonlinear function of the velocity. Usually, the following linear function is used

$$r(\dot{u}(t)) = -c_{ij}\dot{u}_{i,j}(t).$$

The associated temperature boundary condition on Γ_3 is described by

$$k_{ij}\,\theta_{,i}\,n_j = -k_e\,(\theta - \theta_R) \quad \text{on } \Gamma_3 \times (0,T),$$

where θ_R is the temperature of the foundation and k_e is the heat exchange coefficient between the body and the obstacle.

1.3 Subdifferential Contact Condition

Let us here describe the surface contact condition on Γ_3. We model the frictional contact with a general subdifferential boundary condition of the form

$$u \in U, \quad \varphi(v) - \varphi(\dot{u}) \geq -\sigma\nu\,(v - \dot{u}) \qquad \forall v \in U. \tag{1}$$

In this condition, U represents the set of contact admissible test functions, $\sigma\nu$ denotes the Cauchy stress vector on the contact boundary, and $\varphi : \Gamma_3 \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is a given convex function. The inequality in (1) holds almost everywhere on the contact surface. Various situations may be modeled by such a condition. Examples and detailed explanations of inequality problems in contact mechanics which lead to boundary conditions of this form can be found in [9, 23].

Here we present some examples of contact and dry friction laws which lead to such subdifferential inequality.

Example 1. Bilateral contact with Tresca's friction law. This contact condition can be found in [13, 23]. It is written in the form of the following boundary condition:

$$\begin{cases} u_\nu = 0, \quad |\sigma_\tau| \leq g, \\ |\sigma_\tau| < g \implies \dot{u}_\tau = 0, \\ |\sigma_\tau| = g \implies \dot{u}_\tau = -\lambda\sigma_\tau, \ \ \lambda \geq 0 \end{cases} \qquad \text{on } \Gamma_3 \times (0,T). \tag{2}$$

Here $g \geq 0$ represents the friction bound, i.e., the magnitude of the limiting friction traction at which slip begins. The contact is assumed to be bilateral, i.e., there is no loss of contact during the process.

The set of admissible test functions U consists of those elements of H_1 whose normal component vanishes on Γ_3.

Moreover, it is straightforward to show that if $\{u, \sigma\}$ is a pair of regular functions satisfying (1) then

$$\sigma\,\nu(v - \dot{u}) \geq g|\dot{u}_\tau| - g|v_\tau| \qquad \forall v \in U,$$

a.e. on $\Gamma_3 \times (0,T)$. We get the following contact functional

$$\varphi(v) = g|v_\tau|.$$

Example 2. Viscoelastic contact with Tresca's friction law. We consider the contact problem with the boundary conditions

$$\begin{cases} -\sigma_\nu = k|\dot{u}_\nu|^{r-1}\dot{u}_\nu, \ \ |\sigma_\tau| \le g, \\ |\sigma_\tau| < g \implies \dot{u}_\tau = 0, \\ |\sigma_\tau| = g \implies \dot{u}_\tau = -\lambda\sigma_\tau, \ \ \lambda \ge 0 \end{cases} \qquad \text{on } \Gamma_3 \times (0,T). \tag{3}$$

Here $g, k \ge 0$ and the normal contact stress depends on a power of the normal speed (this condition may describe the motion of a body, say a wheel, on a fine granular material, say the sand on a beach). We have $U = H_1$, $0 < r \le 1$, and

$$\varphi(v) = \frac{k}{r+1}|v_\nu|^{r+1} + g|v_\tau|.$$

Example 3. Viscoelastic contact with friction. Here, the body is moving on sand or a granular material and the normal stress is proportional to a power of the normal speed, while the tangential shear is proportional to a power of the tangential speed. We choose the following boundary conditions:

$$-\sigma_\nu = k|\dot{u}_\nu|^{r-1}\dot{u}_\nu, \ \ \sigma_\tau = -\mu|\dot{u}_\tau|^{p-1}\dot{u}_\tau \qquad \text{on } \Gamma_3 \times (0,T). \tag{4}$$

Here $\mu \in L^\infty(\Gamma_3)$ and $k \in L^\infty(\Gamma_3)$ are positive functions and $0 < p, r \le 1$. We choose $U = H_1$, $V = \{\, v \in H_1 \mid v = 0 \ \text{ on } \ \Gamma_1 \,\}$, and

$$\varphi(v) = \frac{k}{r+1}|v_\nu|^{r+1} + \frac{\mu}{p+1}|v_\tau|^{p+1}.$$

Remark 1. In the examples above, the normal pressure as well as the tangential stress is related to powers of the normal and tangential speeds. This is dictated by the structure of the functional φ which depends only on the surface velocity.

1.4 Notation and Functional Spaces

In this short section, we present the notations we shall use and some preliminary materials for functional spaces. For further details, we refer the reader to [13].

We denote by S_d the space of second order symmetric tensors on $\mathbb{R}^d$ $(d = 2,3)$, while " $\cdot$ " and $|\cdot|$ will represent the inner product and the Euclidean norm on S_d and $\mathbb{R}^d$. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary Γ and let ν denote the unit outer normal on Γ. We also use the following notation:

$$H = \{ u = (u_i) \mid u_i \in L^2(\Omega) \}, \qquad \mathcal{H} = \{ \sigma = (\sigma_{ij}) \mid \sigma_{ij} = \sigma_{ji} \in L^2(\Omega) \},$$

$$H_1 = \{ u \in H \mid \varepsilon(u) \in \mathcal{H} \}, \qquad \mathcal{H}_1 = \{ \sigma \in \mathcal{H} \mid \mathrm{Div}\ \sigma \in H \}.$$

Here $\varepsilon : H_1 \longrightarrow \mathcal{H}$ and $\mathrm{Div} : \mathcal{H}_1 \longrightarrow H$ are the deformation and the divergence operators respectively defined by:

$$\varepsilon(u) = (\varepsilon_{ij}(u)), \quad \varepsilon_{ij}(u) = \frac{1}{2}(u_{i,j} + u_{j,i}), \quad \mathrm{Div}\ \sigma = (\sigma_{ij,j}).$$

The spaces H, $\mathcal{H}$, H_1, and $\mathcal{H}_1$ are real Hilbert spaces endowed with the cannonical inner products given by:

$$\langle u, v \rangle_H = \int_\Omega u_i v_i \, dx, \qquad \langle \sigma, \tau \rangle_{\mathcal{H}} = \int_\Omega \sigma_{ij} \tau_{ij} \, dx,$$

$$\langle u, v \rangle_{H_1} = \langle u, v \rangle_H + \langle \varepsilon(u), \varepsilon(v) \rangle_{\mathcal{H}}, \qquad \langle \sigma, \tau \rangle_{\mathcal{H}_1} = \langle \sigma, \tau \rangle_{\mathcal{H}} + \langle \mathrm{Div}\ \sigma, \mathrm{Div}\ \tau \rangle_H.$$

The associated norms on the spaces H, $\mathcal{H}$, H_1, and $\mathcal{H}_1$ are denoted by $|\cdot|_H$, $|\cdot|_{\mathcal{H}}$, $|\cdot|_{H_1}$ and $|\cdot|_{\mathcal{H}_1}$, respectively.

Let $H_\Gamma = H^{\frac{1}{2}}(\Gamma)^d$ and let $\gamma : H_1 \longrightarrow H_\Gamma$ be the trace map. For every element $u \in H_1$, we also use the notation u to denote the trace γu of u on Γ and we denote by u_ν and u_τ the normal and the tangential components of u on Γ given by:

$$u_\nu = u \cdot \nu, \qquad u_\tau = u - u_\nu \nu. \tag{5}$$

Let H'_Γ be the dual of H_Γ and let $< \cdot, \cdot >$ denote the duality pairing between H'_Γ and H_Γ. For every $\sigma \in \mathcal{H}_1$, $\sigma\nu$ can be defined as the element in H'_Γ which satisfies:

$$\langle \sigma\nu, \gamma u \rangle = \langle \sigma, \varepsilon(u) \rangle_{\mathcal{H}} + \langle \mathrm{Div}\ \sigma, u \rangle_H \qquad \forall u \in H_1. \tag{6}$$

Let also σ_ν and σ_τ denote the normal and tangential traces of σ, respectively. If σ is a smooth function, e.g., $\sigma \in C^1$, then

$$\langle \sigma\nu, \gamma u \rangle = \int_\Gamma \sigma\nu \cdot u \, da \qquad \forall u \in H_1 \tag{7}$$

where da is the surface measure element and

$$\sigma_\nu = (\sigma\nu) \cdot \nu, \qquad \sigma_\tau = \sigma\nu - \sigma_\nu \nu. \tag{8}$$

Finally, we recall that $C([0,T];X)$ is the space of continuous functions from $[0,T]$ to X; while $C^m([0,T];X)$ ($m \in \mathbb{N}^*$) is the set of m times differentiable functions. Then $\mathcal{D}(\Omega)$ denotes the set of infinitely differentiable real functions with compact support in Ω. We will also use the Lesbesgue spaces $L^p(0,T;X)$; and the Sobolev spaces:

$$W^{m,p}(0,T;X), \quad H_0^m(\Omega) := \{ w \in W^{m,2}(\Omega),\ w = 0 \text{ on } \Gamma \},$$

where $m \geq 1$ and $1 \leq p \leq +\infty$.

2 Dynamic Contact Problems with Clamped Condition

In this section, we present the results obtained in [1, 3, 8, 10] with the usual fixed condition. In [8], the authors (Chau and Awbi) analyze a problem which describes the frictional contact between a short memory thermo-viscoelastic body and a rigid foundation. The process is assumed to be quasistatic and the contact is modeled by a normal damped response condition with friction law. Moreover, heat exchange condition has been taken into account on the contact surface. The mechanical model is described as a coupled system of a variational elliptic equality for the displacements and a differential heat equation for the temperature. Then the authors present a variational formulation of the problem and establish the existence and uniqueness of weak solution in using general results on evolution equations with monotone operators and fixed point arguments. In [1], the constitutive law has been extended to a long memory viscoelastic type and the contact has been modeled by a general subdifferential condition on the velocity. The authors (K. Addi, O. Chau, and D. Goeleven) derive weak formulations for the models and establish existence and uniqueness results. The proofs are based on evolution variational inequalities, in the framework of monotone operators and fixed point methods. The quasistatic evolution in these two latter works has been then extended to dynamic process for long memory thermo-viscoelastic materials in [3]. Finally, the authors (O. Chau, D. Goeleven, and R. Oujja) complete the study by numerical approximations in [10], where analysis of error order estimate and various simulations have been provided. The dynamic mechanical problem for long memory thermo-viscoelastic materials subjected to subdifferential contact condition and to clamped condition is then formulated as follows.

Problem Q **:** Find a displacement field $u : \Omega \times [0,T] \longrightarrow \mathbb{R}^d$, a stress field $\sigma : \Omega \times [0,T] \longrightarrow S_d$, and a temperature field $\theta : \Omega \times [0,T] \longrightarrow \mathbb{R}_+$ such that for a.e. $t \in (0,T)$:

$$\sigma(t) = \mathcal{A}\varepsilon(\dot{u}(t)) + \mathcal{G}\varepsilon(u(t)) + \int_0^t \mathcal{B}(t-s)\,\varepsilon(u(s))\,ds - \theta(t)C_e \quad \text{in} \quad \Omega, \tag{9}$$

$$\ddot{u}(t) = \mathrm{Div}\,\sigma(t) + f_0(t) \quad \text{in} \quad \Omega, \tag{10}$$

$$u(t) = 0 \quad \text{on} \quad \Gamma_1, \tag{11}$$

$$\sigma(t)\nu = f_2(t) \quad \text{on} \quad \Gamma_2, \tag{12}$$

$$u(t) \in U, \quad \varphi(w) - \varphi(\dot{u}(t)) \geq -\sigma(t)\nu \cdot (w - \dot{u}(t)) \quad \forall w \in U \quad \text{on} \quad \Gamma_3, \tag{13}$$

$$\dot{\theta}(t) - \mathrm{div}(K_c \nabla\theta(t)) = -c_{ij}\frac{\partial \dot{u}_i}{\partial x_j}(t) + q(t) \quad \text{on} \quad \Omega, \tag{14}$$

$$-k_{ij}\frac{\partial \theta}{\partial x_j}(t)\,n_i = k_e\,(\theta(t) - \theta_R) \quad \text{on} \quad \Gamma_3, \tag{15}$$

$$\theta(t) = 0 \quad \text{on} \quad \Gamma_1 \cup \Gamma_2, \tag{16}$$

$$\theta(0) = \theta_0 \quad \text{in} \quad \Omega, \tag{17}$$

$$u(0) = u_0, \quad \dot{u}(0) = v_0 \quad \text{in} \quad \Omega. \tag{18}$$

Here, we suppose that $\text{meas}(\Gamma_1) > 0$ and the mass density $\rho \equiv 1$. We suppose also that the set of contact admissible test functions verifies

$$\mathcal{D}(\Omega)^d \subset U \subset H_1.$$

Finally, u_0, v_0, θ_0 represent, respectively, the initial displacement, velocity, and temperature.

To obtain the variational formulation of the mechanical problems (9)–(18) we need additional notations. Thus, let V denote the closed subspace of H_1 defined by

$$\mathcal{D}(\Omega)^d \subset V = \{ v \in H_1 \mid v = 0 \quad \text{on} \quad \Gamma_1 \} \cap U.$$

We set

$$E = \{\eta \in H^1(\Omega),\ \eta = 0 \quad \text{on} \quad \Gamma_1 \cup \Gamma_2\}, \quad F = L^2(\Omega).$$

Since $\text{meas}\,\Gamma_1 > 0$, Korn's inequality holds, i.e., there exists $C_K > 0$ which depends only on Ω and Γ_1 such that

$$\|\varepsilon(v)\|_{\mathcal{H}} \geq C_K \|v\|_{H_1} \qquad \forall v \in V.$$

A proof of Korn's inequality may be found in [21, p. 79].
On V we consider the inner product given by

$$(u, v)_V = (\varepsilon(u), \varepsilon(v))_{\mathcal{H}} \qquad \forall u, v \in V,$$

and let $\|\cdot\|_V$ be the associated norm, i.e.,

$$\|v\|_V = \|\varepsilon(v)\|_{\mathcal{H}} \qquad \forall v \in V.$$

It follows that $\|\cdot\|_{H_1}$ and $\|\cdot\|_V$ are equivalent norms on V and therefore $(V, \|\cdot\|_V)$ is a real Hilbert space. Moreover, by the Sobolev's trace theorem, we have a constant $C_0 > 0$ depending only on Ω, Γ_1, and Γ_3 such that

$$\|v\|_{L^2(\Gamma_3)} \leq C_0 \|v\|_V \qquad \forall v \in V.$$

In the study of the mechanical problem (9)–(18), we assume the following conditions (see e.g. [3, 20]).

The viscosity operator $\mathcal{A} : \Omega \times S_d \longrightarrow S_d$ satisfies:

$$\begin{cases} \text{(a)} & \text{there exists } L_{\mathcal{A}} > 0 \text{ such that} \\ & |\mathcal{A}(x,\varepsilon_1) - \mathcal{A}(x,\varepsilon_2)| \le L_{\mathcal{A}} |\varepsilon_1 - \varepsilon_2| \\ & \forall \varepsilon_1, \varepsilon_2 \in S_d, \text{ a.e. } x \in \Omega, \\ \text{(b)} & \text{there exists } m_{\mathcal{A}} > 0 \text{ such that} \\ & (\mathcal{A}(x,\varepsilon_1) - \mathcal{A}(x,\varepsilon_2)) \cdot (\varepsilon_1 - \varepsilon_2) \ge m_{\mathcal{A}} |\varepsilon_1 - \varepsilon_2|^2 \\ & \forall \varepsilon_1, \varepsilon_2 \in S_d, \text{ a.e. } x \in \Omega, \\ \text{(c)} & x \longmapsto \mathcal{A}(x,\varepsilon) \text{ is Lebesgue measurable on } \Omega, \forall \varepsilon \in S_d, \\ \text{(d)} & \text{the mapping } x \longmapsto \mathcal{A}(x,0) \in \mathcal{H}. \end{cases} \tag{19}$$

The elasticity operator $\mathcal{G} : \Omega \times S_d \longrightarrow S_d$ satisfies:

$$\begin{cases} \text{(a)} & \text{there exists } L_{\mathcal{G}} > 0 \text{ such that} \\ & |\mathcal{G}(x,\varepsilon_1) - \mathcal{G}(x,\varepsilon_2)| \le L_{\mathcal{G}} |\varepsilon_1 - \varepsilon_2| \\ & \forall \varepsilon_1, \varepsilon_2 \in S_d, \text{ a.e. } x \in \Omega, \\ \text{(b)} & x \longmapsto \mathcal{G}(x,\varepsilon) \text{ is Lebesgue measurable on } \Omega, \forall \varepsilon \in S_d, \\ \text{(c)} & \text{the mapping } x \longmapsto \mathcal{G}(x,0) \in \mathcal{H}. \end{cases} \tag{20}$$

The relaxation tensor $\mathcal{B} : [0,T] \times \Omega \times S_d \longrightarrow S_d$, $(t,x,\tau) \longmapsto (B_{ijkh}(t,x)\,\tau_{kh})$ satisfies

$$\begin{cases} \text{(i) } B_{ijkh} \in W^{1,\infty}(0,T;L^\infty(\Omega)), \\ \text{(ii) } \mathcal{B}(t)\sigma \cdot \tau = \sigma \cdot \mathcal{B}(t)\tau \\ \qquad \forall \sigma, \tau \in S_d, \text{ a.e. } t \in (0,T), \text{ a.e. in } \Omega. \end{cases} \tag{21}$$

We suppose that the body forces and surface tractions satisfy

$$f_0 \in W^{1,2}(0,T;H), \qquad f_2 \in W^{1,2}(0,T;L^2(\Gamma_2)^d). \tag{22}$$

We assume that the thermal tensor and the heat source density satisfy the conditions:

$$C_e = (c_{ij}), \quad c_{ij} = c_{ji} \in L^\infty(\Omega), \quad q \in W^{1,2}(0,T;L^2(\Omega)). \tag{23}$$

The boundary thermic data are supposed to satisfy the regularity condition:

$$k_e \in L^\infty(\Omega;\mathbb{R}^+), \quad \theta_R \in W^{1,2}(0,T;L^2(\Gamma_3)). \tag{24}$$

We suppose that the thermal conductivity tensor verifies the usual symmetry and ellipticity properties, i.e., for some $c_k > 0$ and for all $(\xi_i) \in \mathbb{R}^d$:

$$K_c = (k_{ij}), \quad k_{ij} = k_{ji} \in L^\infty(\Omega), \quad k_{ij}\,\xi_i \xi_j \ge c_k\,\xi_i \xi_i. \tag{25}$$

We assume that the initial data satisfy the conditions

$$u_0 \in V, \quad v_0 \in V, \quad \theta_0 \in E. \tag{26}$$

On the contact surface, the following frictional contact function

$$\psi(w) := \int_{\Gamma_3} \varphi(w)\,da$$

is assumed to verify

$$\begin{cases} \text{(i)}\ \ \psi : V \longrightarrow \mathbb{R} \quad \text{is well defined, continuous, and convex,} \\ \text{(ii)}\ \ \text{there exists a sequence of differentiable convex functions} \\ \qquad (\psi_n) : V \longrightarrow \mathbb{R} \quad \text{such that} \quad \forall w \in L^2(0,T;V), \\ \qquad \displaystyle\int_0^T \psi_n(w(t))\,dt \longrightarrow \int_0^T \psi(w(t))\,dt,\ n \longrightarrow +\infty, \\ \text{(iii)}\ \ \text{for all sequence } (w_n) \text{ and } w \text{ in } W^{1,2}(0,T;V) \text{ such that} \\ \qquad w_n \rightharpoonup w,\ w_n' \rightharpoonup w' \ \text{weakly in } L^2(0,T;V), \\ \qquad \text{then } \displaystyle\liminf_{n\longrightarrow+\infty} \int_0^T \psi_n(w_n(t))\,dt \geq \int_0^T \psi(w(t))\,dt, \\ \text{(iv) if } w \in V,\ w = 0 \ \text{ on } \Gamma_3, \text{ then } \forall n \in \mathbb{N},\ \psi_n'(w) = 0_{V'}. \end{cases} \tag{27}$$

Here $\psi_n'(v)$ denotes the Fréchet derivative of ψ_n at v.

Using Green's formula, we obtain the variational formulation of the mechanical problem Q as follows.

Problem QV **:** Find $u : [0,T] \to V$, $\theta : [0,T] \to E$ satisfying a.e. $t \in (0,T)$:

$$\begin{cases} \langle \ddot{u}(t) + A\,\dot{u}(t) + B\,u(t) + C\,\theta(t), w - \dot{u}(t)\rangle_{V'\times V} \\ + \left(\displaystyle\int_0^t B(t-s)\,\varepsilon(u(s))\,ds, \varepsilon(w) - \varepsilon(\dot{u}(t))\right)_{\mathcal{H}} + \psi(w) - \psi(\dot{u}(t)), \\ \qquad \geq \langle f(t), w - \dot{u}(t)\rangle_{V'\times V} \quad \forall w \in V, \\ \dot{\theta}(t) + K\,\theta(t) = R\dot{u}(t) + Q(t) \qquad \text{in} \quad E', \\ u(0) = u_0, \quad \dot{u}(0) = v_0, \quad \theta(0) = \theta_0. \end{cases}$$

Here the operators and functions $A, B : V \longrightarrow V'$, $C : E \longrightarrow V'$, $\psi : V \longrightarrow \mathbb{R}$, $K : E \longrightarrow E'$, $R : V \longrightarrow E'$, $f : [0,T] \longrightarrow V'$, and $Q : [0,T] \longrightarrow E'$ are defined by $\forall v \in V$, $\forall w \in V$, $\forall \tau \in E$, $\forall \eta \in E$:

$$\begin{aligned} &\langle A\,v, w\rangle_{V'\times V} = (\mathcal{A}(\varepsilon v), \varepsilon w)_{\mathcal{H}}, \\ &\langle B\,v, w\rangle_{V'\times V} = (\mathcal{G}(\varepsilon v), \varepsilon w)_{\mathcal{H}}, \\ &\langle C\tau, w\rangle_{V'\times V} = -(\tau C_e, \varepsilon w)_{\mathcal{H}}, \\ &\langle f(t), w\rangle_{V'\times V} = (f_0(t), w)_H + (f_2(t), w)_{(L^2(\Gamma_2))^d}, \\ &\langle Q(t), \eta\rangle_{E'\times E} = \int_{\Gamma_3} k_e\,\theta_R(t)\,\eta\,dx + \int_\Omega q(t)\,\eta\,dx, \\ &\langle K\,\tau, \eta\rangle_{E'\times E} = \sum_{i,j=1}^{d} \int_\Omega k_{ij}\,\frac{\partial \tau}{\partial x_j}\,\frac{\partial \eta}{\partial x_i}\,dx + \int_{\Gamma_3} k_e\,\tau\cdot\eta\,da, \\ &\langle R\,v, \eta\rangle_{E'\times E} = -\int_\Omega c_{ij}\,\frac{\partial v_i}{\partial x_j}\,\eta\,dx. \end{aligned}$$

Let us recall now the following main mathematical result (see for details [3]):

Theorem 1. *Assume that (19)–(27) hold. Then there exists an unique solution* $\{u,\theta\}$ *to problem QV with the regularity:*

$$\begin{cases} u \in W^{2,2}(0,T;V) \cap W^{2,\infty}(0,T;H), \\ \theta \in W^{1,2}(0,T;E) \cap W^{1,\infty}(0,T;F). \end{cases}$$

2.1 Analysis of a Numerical Scheme

In this section, we study a fully discrete numerical approximation scheme of the variational problem *QV* (see [10]). For this purpose, we suppose in the following that the conditions on the data (19)–(27) of Theorem 1 are satisfied. In particular, we have

$$f \in C([0,T];V'), \quad Q \in C([0,T];E').$$

Let $\{u,\theta\}$ be the unique solution of the problem *QV* and let us introduce the velocity variable

$$v(t) = \dot{u}(t), \quad \forall t \in [0,T].$$

Then

$$u(t) = u_0 + \int_0^t v(s)\,ds, \quad \forall t \in [0,T].$$

From Theorem 1 we see that $\{v,\theta\}$ verify for all $t \in [0,T]$:

$$\begin{cases} \langle \dot{v}(t) + A\,v(t) + B\,u(t) + C\,\theta(t), w - v(t)\rangle_{V'\times V} \\ + \left(\displaystyle\int_0^t B(t-s)\,\varepsilon(u(s))\,ds, \varepsilon(w) - \varepsilon(\dot{u}(t))\right)_{\mathcal{H}} + \psi(w) - \psi(v(t)) \\ \qquad \geq \langle f(t), w - v(t)\rangle_{V'\times V}, \quad \forall w \in V. \end{cases} \tag{28}$$

$$\langle \dot{\theta}(t), \eta\rangle_F + \langle K\,\theta(t), \eta\rangle_{E'\times E} = \langle R v(t), \eta\rangle_{E'\times E} + \langle Q(t), \eta\rangle_{E'\times E}, \quad \forall \eta \in E. \tag{29}$$

$$u(0) = u_0, \quad v(0) = v_0, \quad \theta(0) = \theta_0, \tag{30}$$

with the regularity:

$$\begin{cases} v \in W^{1,2}(0,T;V) \cap W^{1,\infty}(0,T;H), \\ \theta \in W^{1,2}(0,T;E) \cap W^{1,\infty}(0,T;F). \end{cases} \tag{31}$$

In this section, we make the following additional assumptions on the solution and contact function:

$$v \in W^{2,2}(0,T;H), \tag{32}$$

$$\theta \in W^{2,2}(0,T;F), \tag{33}$$

$$\psi \text{ is Lipschitz continuous on } V. \tag{34}$$

Let $V^h \subset V$ and $E^h \subset E$ be a family of finite dimensional subspaces, with $h > 0$ a discretization parameter. We divide the time interval $[0,T]$ into N equal parts: $t_n = nk$, $n = 0,1,\ldots,N$, with the time step $k = T/N$. For a continuous function $w \in C([0,T];X)$ with values in a space X, we use the notation $w_n = w(t_n) \in X$. Then from (28) and (29) we introduce the following fully discrete scheme.

Problem P^{hk}. Find $v^{hk} = \{v_n^{hk}\}_{n=0}^N \subset V^h$, $\theta^{hk} = \{\theta_n^{hk}\}_{n=0}^N \subset E^h$ such that

$$v_0^{hk} = v_0^h, \quad \theta_0^{hk} = \theta_0^h \tag{35}$$

and for $n = 1,\ldots,N$,

$$\begin{aligned} &\left(\frac{v_n^{hk} - v_{n-1}^{hk}}{k}, w^h - v_n^{hk}\right)_H + \langle A\, v_n^{hk}, w^h - v_n^{hk}\rangle_{V'\times V} + \langle B\, u_{n-1}^{hk}, w^h - v_n^{hk}\rangle_{V'\times V} \\ &\quad + \langle C\, \theta_{n-1}^{hk}, w^h - v_n^{hk}\rangle_{V'\times V} + \psi(w^h) - \psi(v_n^{hk}) \\ &\quad + \left(k\sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m)\,\varepsilon(u_m^{hk}), \varepsilon(w^h) - \varepsilon(v_n^{hk})\right)_{\mathcal{H}} \\ &\quad \geq \langle f_n, w^h - v_n^{hk}\rangle_{V'\times V}, \quad \forall\, w^h \in V^h, \end{aligned} \tag{36}$$

$$\begin{aligned} &\left(\frac{\theta_n^{hk} - \theta_{n-1}^{hk}}{k}, \eta^h\right)_F + \langle K\,\theta_n^{hk}, \eta^h\rangle_{E'\times E} \\ &\quad = \langle R\, v_n^{hk}, \eta^h\rangle_{E'\times E} + \langle Q_n, \eta^h\rangle_{E'\times E}, \quad \forall\, \eta^h \in E^h, \end{aligned} \tag{37}$$

where

$$u_n^{hk} = u_{n-1}^{hk} + k\, v_n^{hk}, \quad u_0^{hk} = u_0^h. \tag{38}$$

Here $u_0^h \in V^h$, $v_0^h \in V^h$, $\theta_0^h \in E^h$ are suitable approximations of the initial values u_0, v_0, θ_0.

For $n = 1,\ldots,N$, suppose that u_{n-1}^{hk}, v_{n-1}^{hk}, θ_{n-1}^{hk} are known. We may then calculate v_n^{hk} by (36), θ_n^{hk} by (37), and u_n^{hk} by (38). Hence the discrete solution $v^{hk} \subset V^h$, $\theta^{hk} \subset E^h$ exists and is unique.

We now turn to an error analysis of the numerical solution. The main result of this section is the following one (see for details [10]).

Theorem 2. *We keep the assumptions of Theorem 1. Under the additional assumptions (32)–(33), then for the unique solution $v^{hk} \subset V^h$, $\theta^{hk} \subset E^h$ of the discrete problem P^{hk}, we have the following error estimate*

$$\begin{aligned}
&\max_{1\le n\le N}\|v_n-v_n^{hk}\|_H^2+\left(k\sum_{n=1}^{N}\|v_n-v_n^{hk}\|_V^2\right)+\max_{1\le n\le N}\|\theta_n-\theta_n^{hk}\|_F^2+\left(k\sum_{n=1}^{N}\|\theta_n-\theta_n^{hk}\|_E^2\right)\\
&\quad\le c\|u_0-u_0^h\|_V^2+c\|v_0-v_0^{hk}\|_H^2+c\|\theta_0-\theta_0^h\|_F^2+c\max_{1\le n\le N}\|v_n-w_n^h\|_H\\
&\qquad+c\max_{1\le n\le N}\|\theta_n-\eta_n^h\|_F^2+ck\sum_{j=1}^{N}\|v_j-w_j^h\|_V^2+ck\sum_{j=1}^{N}\|\theta_j-\eta_j^h\|_E^2\\
&\qquad+c\left(\sum_{j=1}^{N-1}\|(v_j-w_j^h)-(v_{j+1}-w_{j+1}^h)\|_H\right)^2\\
&\qquad+c\left(\sum_{j=1}^{N-1}\|\theta_j-\eta_j^h-(\theta_{j+1}-\eta_{j+1}^h)\|_F\right)^2+ck^2+ck\sum_{j=1}^{N}\|v_j-w_j^h\|_V,
\end{aligned}\tag{39}$$

where for $j=1,\dots,N$, $w_j^h\in V^h$, $\eta_j^h\in E^h$ are arbitrary.

The inequality (39) is a basis for error estimates for particular choice of the finite-dimensional subspace V^h and under additional data and solution regularities.

As a typical example, let us consider $\Omega\subset\mathbb{R}^d$, $d\in\mathbb{N}^*$, a polygonal domain. Let $\mathcal{T}^h$ be a regular finite element partition of Ω. Let $V^h\subset V$ and $E^h\subset E$ be the finite element space consisting of piecewise polynomials of degree $\le m-1$, with $m\ge 2$, according to the partition $\mathcal{T}^h$. Denote by $\Pi_V^h: H^m(\Omega)^d\to V^h$ and $\Pi_E^h: H^m(\Omega)\to E^h$ the corresponding finite element interpolation operator. Recall (see, e.g., [11]) that:

$$\begin{cases}\|w-\Pi_V^h w\|_{H^l(\Omega)^d}\le ch^{m-l}|w|_{H^m(\Omega)^d}, & \forall w\in H^m(\Omega)^d,\\ \|\eta-\Pi_E^h\eta\|_{H^l(\Omega)}\le ch^{m-l}|\eta|_{H^m(\Omega)}, & \forall\eta\in H^m(\Omega),\end{cases}$$

where $l=0$ (for which $H^0=L^2$) or $l=1$.

In the following we assume the additional data and solution regularities

$$\begin{cases}u_0\in H^{\alpha+1}(\Omega)^d,\\ v\in C([0,T];H^{2\alpha+1}(\Omega)^d),\quad \dot v\in L^1(0,T;H^\alpha(\Omega)^d),\\ \theta\in C([0,T];H^{\alpha+1}(\Omega)),\quad \dot\theta\in L^1(0,T;H^\alpha(\Omega)).\end{cases}\tag{40}$$

Here

$$\alpha=m-1\ge 1.$$

We remark that the previous properties already hold for $\alpha=1$, except for

$$v\in C([0,T];H^3(\Omega)^d)\quad\text{and}\quad\theta\in C([0,T];H^2(\Omega)).$$

Then we choose in (39) the elements

$$u_0^h=\Pi_V^h u_0,\quad v_0^h=\Pi_V^h v_0,\quad \theta_0^h=\Pi_E^h\theta_0,$$

and

$$w_j^h=\Pi_V^h v_j,\quad \eta_j^h=\Pi_E^h\theta_j,\quad j=1,\dots,N.$$

From the assumptions (40), we have:

$$\begin{aligned} \|u_0 - u_0^h\|_V \le c\,h^\alpha, \quad \|e_0\|_H \le c\,h^\alpha, \quad \|\varepsilon_0\|_F \le c\,h^\alpha, \\ A_0 \le c\,h^\alpha, \quad B_0 \le c\,h^{2\alpha}, \\ A_3 \le c\,h^\alpha, \quad B_3 \le c\,h^\alpha, \\ kA_2 \le c\,h^{2\alpha}, \quad kB_2 \le c\,h^{2\alpha}, \quad k\hat{B}_2 \le c\,h^{2\alpha}. \end{aligned}$$

Using these estimates in (39), we conclude to the following error estimate result.

Theorem 3. *We keep the assumptions of Theorem 2. Under the additional assumptions (40), we obtain the error estimate for the corresponding discrete solution* v_n^{hk}, θ_n^{hk}, $n = 1,\ldots,N$:

$$\begin{aligned} &\max_{0\le n\le N} \|v_n - v_n^{hk}\|_H + \left(k\sum_{n=0}^{N} \|v_n - v_n^{hk}\|_V^2\right)^{1/2} \\ &\quad + \max_{0\le n\le N} \|\theta_n - \theta_n^{hk}\|_F + \left(k\sum_{n=0}^{N} \|\theta_n - \theta_n^{hk}\|_E^2\right)^{1/2} \le c\,(h^\alpha + k). \end{aligned}$$

In particular, for $\alpha = 1$, *we have*

$$\begin{aligned} &\max_{0\le n\le N} \|v_n - v_n^{hk}\|_H + \left(k\sum_{n=0}^{N} \|v_n - v_n^{hk}\|_V^2\right)^{1/2} \\ &\quad + \max_{0\le n\le N} \|\theta_n - \theta_n^{hk}\|_F + \left(k\sum_{n=0}^{N} \|\theta_n - \theta_n^{hk}\|_E^2\right)^{1/2} \le c\,(h + k). \end{aligned}$$

2.2 Numerical Computations

Here we consider two typical examples of thermal contact problems with Tresca's friction law, the first one is bilateral and the second one obeys a normal damped response condition (see [3, 10]). We provide numerical simulations for the discrete schemes in Sect. 3 in using Matlab computation codes.

Example 4. Thermal bilateral contact problem with Tresca's friction law.
The contact condition on Γ_3 is bilateral, and satisfies (see e.g. [13, 22]):

$$\begin{cases} u_\nu = 0, \quad |\sigma_\tau| \le g, \\ |\sigma_\tau| < g \Longrightarrow \dot{u}_\tau = 0, \\ |\sigma_\tau| = g \Longrightarrow \dot{u}_\tau = -\lambda\,\sigma_\tau, \text{ for some } \lambda \ge 0, \end{cases} \qquad \text{on } \Gamma_3 \times (0,T).$$

Here g represents the friction bound, i.e., the magnitude of the limiting friction traction at which slip begins, with $g \in L^\infty(\Gamma_3)$, $g \geq 0$ a.e. on Γ_3. The corresponding admissible displacement space is:

$$V := \{w \in H_1, \text{ with } w = 0 \text{ on } \Gamma_1,\ w_\nu = 0 \text{ on } \Gamma_3\},$$

and the subdifferential contact function is given by:

$$\varphi(x,y) = g(x)|y_{\tau(x)}| \quad \forall x \in \Gamma_3,\ y \in \mathbb{R}^d,$$

where $y_{\tau(x)} := y - y_{\nu(x)}\nu(x)$, $y_{\nu(x)} := y \cdot \nu(x)$, with $\nu(x)$ denoting the unit normal at $x \in \Gamma_3$. Then the function

$$\psi(v) := \int_{\Gamma_3} g\,|v_\tau|\,da, \quad \forall v \in V$$

is well defined on V and is Lipschitz continuous on V (see [3, 10]).

For our computations, we consider a rectangular open set, linear elastic and long memory viscoelastic operators. We set:

$$\begin{aligned}
&\Omega = (0,L_1) \times (0,L_2),\\
&\Gamma_1 = (\{0\} \times [0,L_2]), \quad \Gamma_2 = [0,L_1] \times \{L_2\} \cup (\{L_1\} \times [0,L_2]), \quad \Gamma_3 = [0,L_1] \times \{0\},\\
&(\mathcal{G}\,\tau)_{ij} = \frac{E\,\kappa}{1-\kappa^2}(\tau_{11}+\tau_{22})\,\delta_{ij} + \frac{E}{1+\kappa}\tau_{ij}, \quad 1 \leq i,j \leq 2,\ \tau \in S_2,\\
&(\mathcal{A}\,\tau)_{ij} = \mu\,(\tau_{11}+\tau_{22})\,\delta_{ij} + \eta\,\tau_{ij}, \quad 1 \leq i,j \leq 2,\ \tau \in S_2,\\
&(\mathcal{B}(t)\,\tau)_{ij} = B_1(t)\,(\tau_{11}+\tau_{22})\,\delta_{ij} + B_2(t)\,\tau_{ij}, \quad 1 \leq i,j \leq 2,\ \tau \in S_2,\ t \in [0,T].
\end{aligned}$$

Here E is the Young's modulus, κ is the Poisson's ratio of the material, δ_{ij} denotes the Kronecker symbol, and μ, η are viscosity constants.

We use spaces of continuous piecewise affine functions $V^h \subset V$ and $E^h \subset E$ as families of approximating subspaces. For our computations, we considered the following data (IS unity):

$$\begin{aligned}
&L_1 = L_2 = 1, \quad T = 1,\\
&\mu = 10, \quad \eta = 10, \quad E = 2, \quad \kappa = 0.1,\\
&f_0(x,t) = (0,-t), \quad f_2(x,t) = (1,0), \quad \forall t \in [0,T],\\
&c_{ij} = k_{ij} = k_e = 1,\ 1 \leq i,j \leq 2, \quad q = 1,\\
&B_1(t) = B_2(t) = 10^{-2}\,e^{-t}, \quad \forall t \in [0,T],\\
&u_0 = (0,0), \quad v_0 = (0,0), \quad \theta_0 = 0.
\end{aligned}$$

The initial configuration is represented in Fig. 2.

Then we show in Fig. 3 the deformed configurations at final time, where the relaxation coefficients are positive and decreasing for the two different types of Tresca's friction bounds. For small friction bound, where $g(x,0) = \frac{x}{2}$, $0 \leq x \leq 1$, we observe on the contact surface a slip phenomena in the direction of the surface

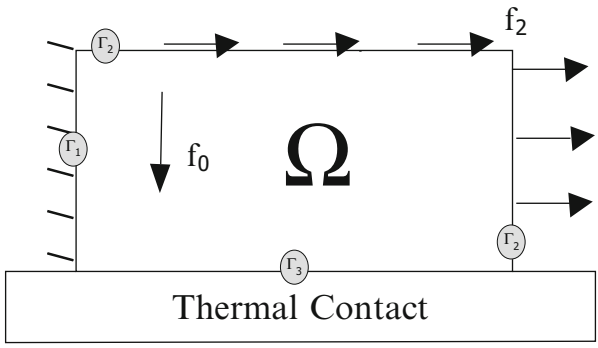
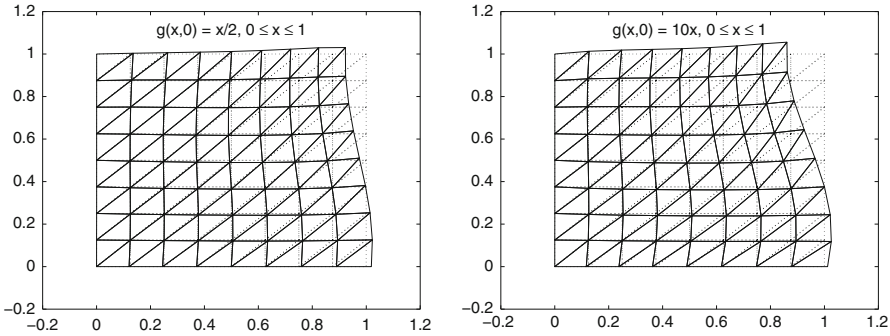


Fig. 2 Initial configuration



$g(x,0) = x/2,\ 0 \le x \le 1$ $\qquad$ $g(x,0) = 10x,\ 0 \le x \le 1$

Fig. 3 Deformed configurations at final time, $\theta_R(t) = 1,\ 0 \le t \le 1$

fraction on Γ_2. That means that the friction bound has been obtained in the zone of the values of x near to 1. Whereas for large friction bound, e.g., for $g(x,0) = 10x$, $0 \le x \le 1$, then slip in the direction of the traction could not be realized. In Fig. 4, we compute the Von Mise norm, which gives a global measure of the stress in the body. The maxima of the norm could be seen in the neighborhood of the point $(0,1)$ for small friction bounds and in the neighborhood of the point $(1,0)$ for large friction bounds. In Fig. 5, we show the influence of the different temperatures of the foundation on the temperature field of the body. We observe larger deformations of the body for greater temperature of the foundation.

Example 5. Thermal contact problem with normal damped response and Tresca's friction law.

The normal damped response contact condition with Tresca's friction law is defined by:

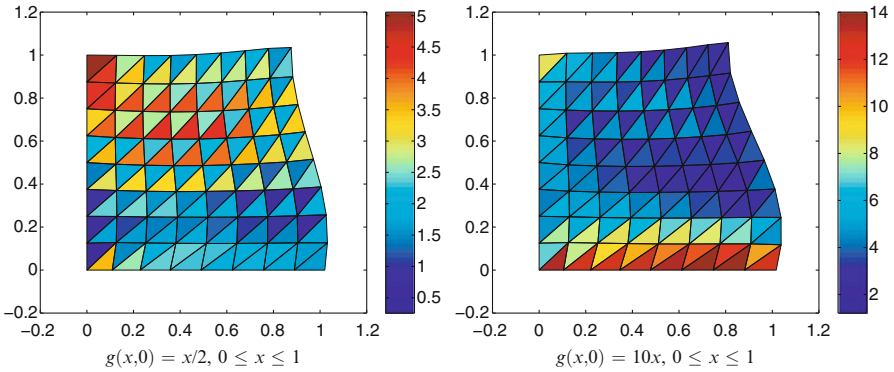


Fig. 4 Von Mise norm in deformed configurations, $\theta_R(t) = 1,\ 0 \le t \le 1$

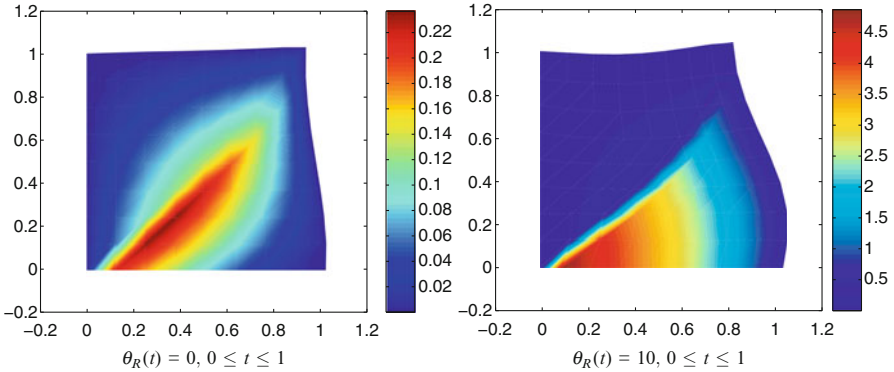


Fig. 5 Temperature field at final time, $g(x,0) = \frac{x}{2},\ 0 \le x \le 1$

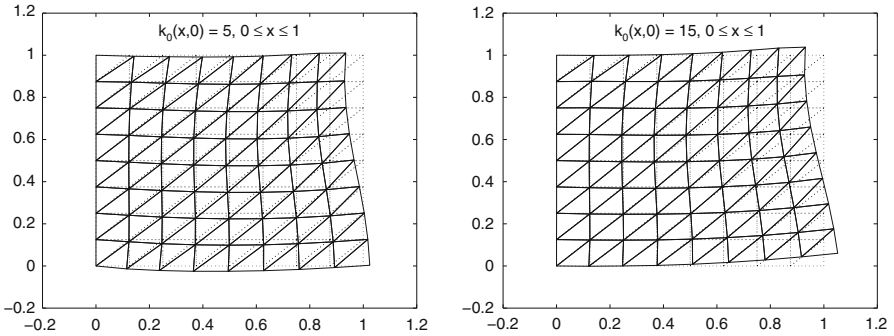


Fig. 6 Deformed configurations at final time, $\theta_R(t) = 1,\ 0 \le t \le 1$

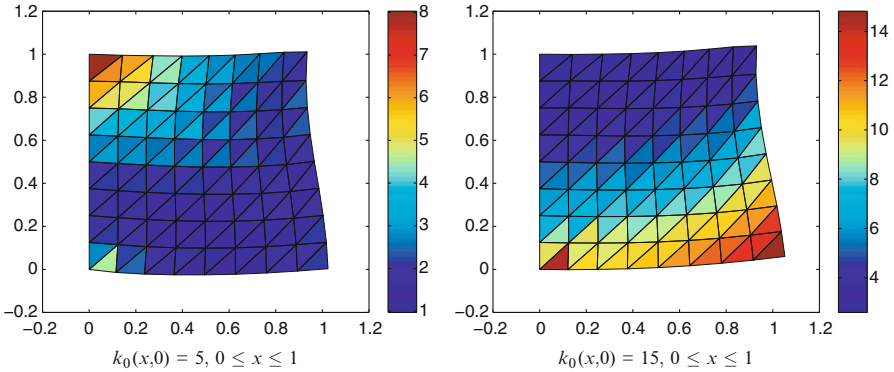


Fig. 7 Von Mise norm in deformed configurations, $\theta_R(t) = 1,\ 0 \le t \le 1$

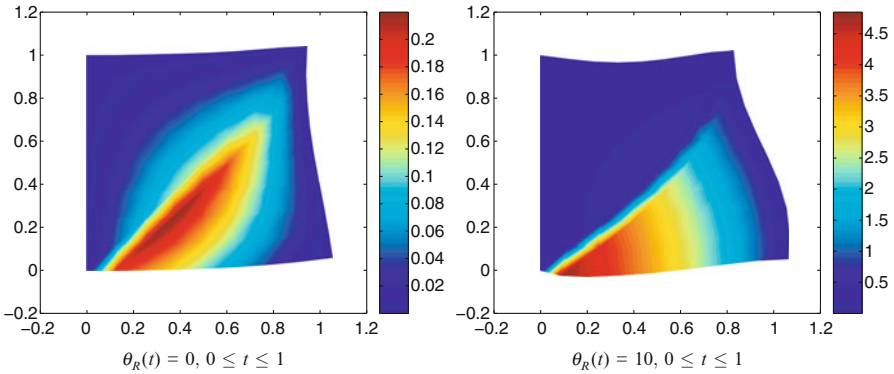


Fig. 8 Temperature field at final time, $k_0(x,0) = 15,\ 0 \le x \le 1$

$$\begin{cases} -\sigma_\nu = k_0\,|\dot{u}_\nu|^{r-1}\dot{u}_\nu, \quad |\sigma_\tau| \le g, \\ |\sigma_\tau| < g \implies \dot{u}_\tau = 0, \\ |\sigma_\tau| = g \implies \dot{u}_\tau = -\lambda\sigma_\tau, \quad \text{for some } \lambda \ge 0, \end{cases} \qquad \text{on } \Gamma_3 \times (0,T).$$

Here $0 < r < 1$ and $g, k_0 \in L^\infty(\Gamma_3)$, $g \ge 0$, $k_0 \ge 0$. The coefficient k_0 represents the hardness of the foundation and g is the friction threshold.

The admissible displacement space is given by:

$$V := \{w \in H_1,\ \text{with } w = 0 \text{ on } \Gamma_1\}$$

and the subdifferential contact function is:

$$\varphi(x,y) = \frac{1}{r+1}\,k_0(x)\,|y_{\nu(x)}|^{r+1} + g(x)|y_\tau(x)| \quad \forall x \in \Gamma_3,\ y \in \mathbb{R}^d.$$

Then setting $p := r+1$, we have the contact function well defined on V by

$$\psi(v) := \int_{\Gamma_3} \frac{k_0}{p} |v_\nu|^p \, da, + \int_{\Gamma_3} g\,|v_\tau| \, da, \quad \forall v \in V.$$

We verify also that ψ is Lipschitz continuous on V (see [3, 10]).

For our computations, we take again the previous rectangular open set, with linear elasticity and visco-elasticity, and used the following data (IS unity):

$$\begin{array}{l} L_1 = L_2 = 1, \quad T = 1, \\ \mu = 10, \quad \eta = 10, \quad E = 2, \quad \kappa = 0.1, \\ f_0(x,t) = (0,-t), \quad f_2(x,t) = (1,0), \quad \forall t \in [0,T], \\ c_{ij} = k_{ij} = k_e = 1,\ 1 \le i,j \le 2, \quad q = 1, \\ g(x,0) = \frac{x}{2},\ 0 \le x \le 1, \quad r = 0.5, \\ B_1(t) = B_2(t) = 10^{-2}\, e^{-t}, \quad \forall t \in [0,T], \\ u_0 = (0,0), \quad v_0 = (0,0), \quad \theta_0 = 0. \end{array}$$

We show in Fig. 6 the deformed configurations at final time, and through the body for different normal damped response coefficients k_0, we verify that the penetrability of the foundation depends on its coefficient of hardness. In Fig. 7 we compute the Von Mise norm. Larger stress near the contact surface is then observed for hard obstacle. Finally in Fig. 8, we find again the influence of the temperature of the foundation on the temperature field of the body and on the final deformed configurations.

3 Dynamic Contact Problems with Free Boundary Condition

We present here a class of dynamic thermal subdifferential contact problems with friction for long memory visco-elastic materials and without the clamped condition. The boundary Γ of the body Ω is partitioned into three disjoint measurable parts Γ_1, Γ_2, and Γ_3, with $\mathrm{meas}(\Gamma_1) = 0$.

The model leads to a system defined by a second order evolution inequality, coupled with a first order evolution equation. We establish an existence and uniqueness result. Finally a fully discrete scheme for numerical approximations is provided and corresponding various numerical computations in dimension two will be given for the cases where Γ_1 is reduced to one point or is an empty set (see [2, 10]).

The dynamic mechanical problem for long memory thermo-viscoelastic materials subjected to subdifferential contact condition and to non-clamped condition is then formulated as follows.

Problem Q **:** Find a displacement field $u : \Omega \times [0,T] \longrightarrow \mathbb{R}^d$ and a stress field $\sigma : \Omega \times [0,T] \longrightarrow S_d$ and a temperature field $\theta : \Omega \times [0,T] \longrightarrow \mathbb{R}_+$ such that for a.e. $t \in (0,T)$:

$$\sigma(t) = \mathcal{A}\varepsilon(\dot{u}(t)) + \mathcal{G}\varepsilon(u(t)) + \int_0^t \mathcal{B}(t-s)\,\varepsilon(u(s))\,ds - \theta(t)\,C_e \quad \text{in} \quad \Omega, \qquad (41)$$

$$\ddot{u}(t) = \mathrm{Div}\,\sigma(t) + f_0(t) \quad \text{in} \quad \Omega, \tag{42}$$

$$\sigma(t)\nu = f_2(t) \quad \text{on} \quad \Gamma_2, \tag{43}$$

$$u(t) \in U, \quad \varphi(w) - \varphi(\dot{u}(t)) \geq -\sigma(t)\nu \cdot (w - \dot{u}(t)) \quad \forall w \in U \quad \text{on} \quad \Gamma_3, \tag{44}$$

$$\dot{\theta}(t) - \mathrm{div}(K_c \nabla\theta(t)) = -c_{ij}\frac{\partial \dot{u}_i}{\partial x_j}(t) + q(t) \quad \text{on} \quad \Omega, \tag{45}$$

$$-k_{ij}\frac{\partial \theta}{\partial x_j}(t)\, n_i = k_e\,(\theta(t) - \theta_R) \quad \text{on} \quad \Gamma_3, \tag{46}$$

$$\theta(t) = 0 \quad \text{on} \quad \Gamma_2, \tag{47}$$

$$\theta(0) = \theta_0 \quad \text{in} \quad \Omega, \tag{48}$$

$$u(0) = u_0, \quad \dot{u}(0) = v_0 \quad \text{in} \quad \Omega. \tag{49}$$

It is worth to notice that the new feature here is due to the absence of the usual claimed condition. However, there is coerciveness with regard to the temperature by (46).

To derive the variational formulation of the mechanical problems (41)–(49), let us introduce the spaces V and E defined by

$$\mathcal{D}(\Omega)^d \subset V = H_1 \cap U,$$

$$E = \{\eta \in H^1(\Omega),\ \eta = 0 \quad \text{on} \quad \Gamma_2\}, \quad F = L^2(\Omega).$$

On V, we consider the inner product given by

$$(u, v)_V = (\varepsilon(u), \varepsilon(v))_{\mathcal{H}} + (u, v)_H \qquad \forall u, v \in V,$$

and the associated norm

$$\|v\|_V^2 = \|\varepsilon(v)\|_{\mathcal{H}}^2 + \|v\|_H^2 \qquad \forall v \in V.$$

It follows that $\|\cdot\|_{H_1}$ and $\|\cdot\|_V$ are equivalent norms on V and therefore $(V, \|\cdot\|_V)$ is a real Hilbert space.

In the study of the mechanical problem (41)–(49), we put again the analogous assumptions as in Sect. 2 on the different operators and data.

The viscosity operator $\mathcal{A} : \Omega \times S_d \longrightarrow S_d$, $(x, \tau) \longmapsto (a_{ijkh}(x)\,\tau_{kh})$ is linear on the second variable and satisfies the usual properties of ellipticity and symmetry, i.e.,

$$\begin{cases} \text{(i) } a_{ijkh} \in W^{1,\infty}(\Omega), \\ \text{(ii) } \mathcal{A}\sigma \cdot \tau = \sigma \cdot \mathcal{A}\tau \ \ \forall \sigma, \tau \in S_d, \text{ a.e. in } \Omega, \\ \text{(iii) there exists } m_{\mathcal{A}} > 0 \text{ such that} \\ \qquad \mathcal{A}\tau \cdot \tau \geq m_{\mathcal{A}}\,|\tau|^2 \quad \forall \tau \in S_d, \text{ a.e. in } \Omega. \end{cases} \tag{50}$$

The elasticity operator $\mathcal{G} : \Omega \times S_d \longrightarrow S_d$ satisfies:

$$\begin{cases} \text{(i)} & \text{there exists } L_{\mathcal{G}} > 0 \text{ such that} \\ & |\mathcal{G}(x,\varepsilon_1) - \mathcal{G}(x,\varepsilon_2)| \leq L_{\mathcal{G}}|\varepsilon_1 - \varepsilon_2| \\ & \forall \varepsilon_1, \varepsilon_2 \in S_d, \text{ a.e. } x \in \Omega, \\ \text{(ii)} & x \longmapsto \mathcal{G}(x,\varepsilon) \text{ is Lebesgue measurable on } \Omega, \forall \varepsilon \in S_d, \\ \text{(iii)} & \text{the mapping } x \longmapsto \mathcal{G}(x,0) \in \mathcal{H}. \end{cases} \tag{51}$$

The relaxation tensor $\mathcal{B} : [0,T] \times \Omega \times S_d \longrightarrow S_d,\ (t,x,\tau) \longmapsto (B_{ijkh}(t,x)\,\tau_{kh})$ satisfies

$$\begin{cases} \text{(i) } B_{ijkh} \in W^{1,\infty}(0,T;L^\infty(\Omega)), \\ \text{(ii) } \mathcal{B}(t)\sigma \cdot \tau = \sigma \cdot \mathcal{B}(t)\tau \\ \qquad \forall \sigma, \tau \in S_d, \text{ a.e. } t \in (0,T), \text{ a.e. in } \Omega. \end{cases} \tag{52}$$

We suppose the body forces and surface tractions satisfy

$$f_0 \in W^{1,2}(0,T;H), \qquad f_2 \in W^{1,2}(0,T;L^2(\Gamma_F)^d). \tag{53}$$

For the thermal tensors and the heat sources density, we suppose that

$$C_e = (c_{ij}), \quad c_{ij} = c_{ji} \in L^\infty(\Omega), \quad q \in W^{1,2}(0,T;L^2(\Omega)). \tag{54}$$

The boundary thermal data satisfy

$$k_e \in L^\infty(\Omega;\mathbb{R}^+), \quad \theta_R \in W^{1,2}(0,T;L^2(\Gamma_3)). \tag{55}$$

The thermal conductivity tensor verifies the usual symmetry end ellipticity: for some $c_k > 0$ and for all $(\xi_i) \in \mathbb{R}^d$,

$$K_c = (k_{ij}), \quad k_{ij} = k_{ji} \in L^\infty(\Omega), \quad k_{ij}\,\xi_i\xi_j \geq c_k\,\xi_i\xi_i. \tag{56}$$

Finally we have to put technical assumptions on the initial data and the sub-differential condition on the contact surface as to use classical results on first order set valued evolution equations. Here we use a general theorem taken in [20, p. 46], in a simplified case, which is enough for our proposal and applications.

We assume that the initial data satisfy the conditions

$$u_0 \in V, \quad v_0 \in V \cap H_0^2(\Omega)^d, \quad \theta_0 \in E \cap H_0^2(\Omega). \tag{57}$$

On the contact surface, the following frictional contact function

$$\psi(w) := \int_{\Gamma_3} \varphi(w)\,da$$

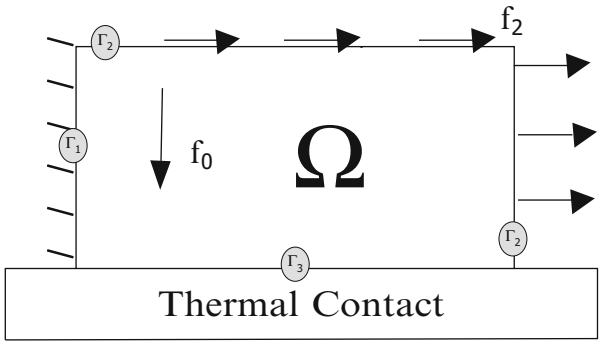


Fig. 9 Initial configuration

verifies

$$\begin{cases} \text{(i)}\ \ \psi : V \longrightarrow \mathbb{R} \quad \text{is well defined, continuous, and convex,} \\ \text{(ii)}\ \ \text{there exists a sequence of differentiable convex functions} \\ \qquad (\psi_n) : V \longrightarrow \mathbb{R} \quad \text{such that} \quad \forall w \in L^2(0,T;V), \\ \qquad \displaystyle\int_0^T \psi_n(w(t))\,dt \longrightarrow \int_0^T \psi(w(t))\,dt,\ n \longrightarrow +\infty, \\ \text{(iii) for all sequence } (w_n) \text{ and } w \text{ in } W^{1,2}(0,T;V) \text{ such that} \\ \qquad w_n \rightharpoonup w,\ w_n' \rightharpoonup w' \ \text{weakly in } L^2(0,T;V), \\ \qquad \text{then } \liminf_{n\longrightarrow+\infty} \displaystyle\int_0^T \psi_n(w_n(t))\,dt \geq \int_0^T \psi(w(t))\,dt, \\ \text{(iv) if } w \in V \text{ and } w = 0 \text{ on } \Gamma_3, \text{ then } \forall n \in \mathbb{N},\ \psi_n'(w) = 0_{V'}. \end{cases} \tag{58}$$

The weak formulation of the mechanical problem Q is then formulated as follows.

Problem QV **:** Find $u : [0,T] \to V$, $\theta : [0,T] \to E$ satisfying a.e. $t \in (0,T)$:

$$\begin{cases} \langle \ddot{u}(t) + A\,\dot{u}(t) + B\,u(t) + C\,\theta(t), w - \dot{u}(t)\rangle_{V'\times V} \\ + \left(\displaystyle\int_0^t B(t-s)\,\varepsilon(u(s))\,ds, \varepsilon(w) - \varepsilon(\dot{u}(t)) \right)_{\mathcal{H}} + \psi(w) - \psi(\dot{u}(t)) \\ \qquad \geq \langle f(t), w - \dot{u}(t)\rangle_{V'\times V} \quad \forall w \in V, \\ \dot{\theta}(t) + K\,\theta(t) = R\dot{u}(t) + Q(t) \qquad \text{in} \quad E', \\ u(0) = u_0, \quad \dot{u}(0) = v_0, \quad \theta(0) = \theta_0. \end{cases}$$

The different operators are here defined as in Sect. 2. Then we obtain our main existence and uniqueness result stated as below (see for details [2]):

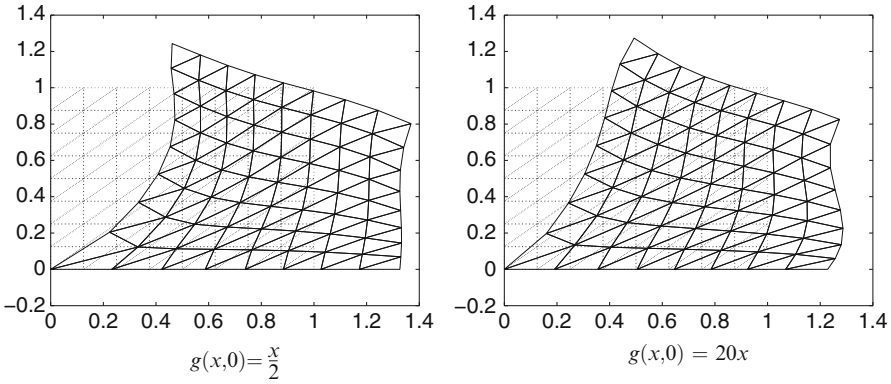


Fig. 10 Deformed configurations at final time, $\theta_R(t) = 0,\ 0 \le t \le 1$

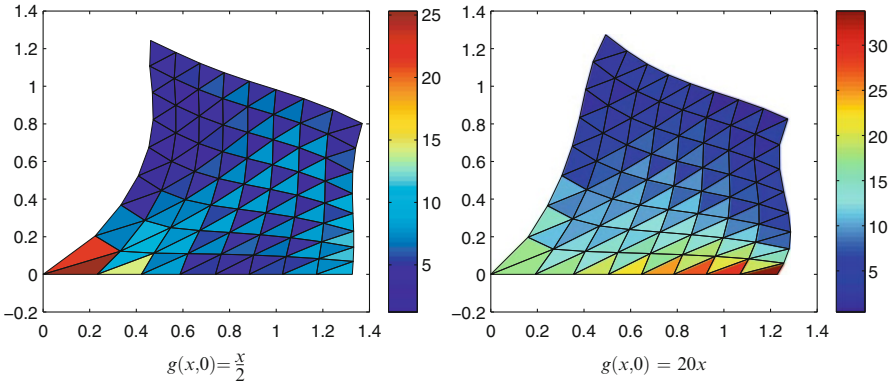


Fig. 11 Von Mises' norm in deformed configurations, $\theta_R(t) = 0,\ 0 \le t \le 1$

Theorem 4. *Assume that (50)–(58) hold, then there exists an unique solution* $\{u, \theta\}$ *to problem QV with the regularity:*

$$\begin{cases} u \in W^{2,2}(0,T;V) \cap W^{2,\infty}(0,T;H), \\ \theta \in W^{1,2}(0,T;E) \cap W^{1,\infty}(0,T;F). \end{cases} \tag{59}$$

3.1 Numerical Simulations A

Here, Γ_1 is reduced to one point (see [10]). We take again the two typical examples and the analogous data as in Sect. 2, except the followings:

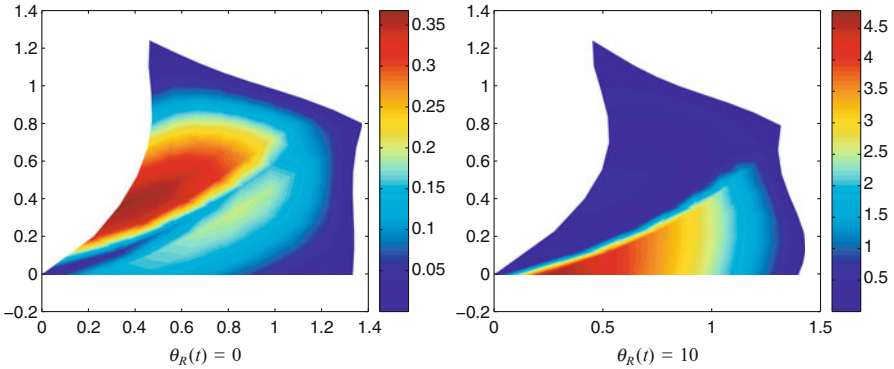


Fig. 12 Temperature field at final time, $g(x,0) = \frac{x}{2},\ 0 \leq x \leq 1$

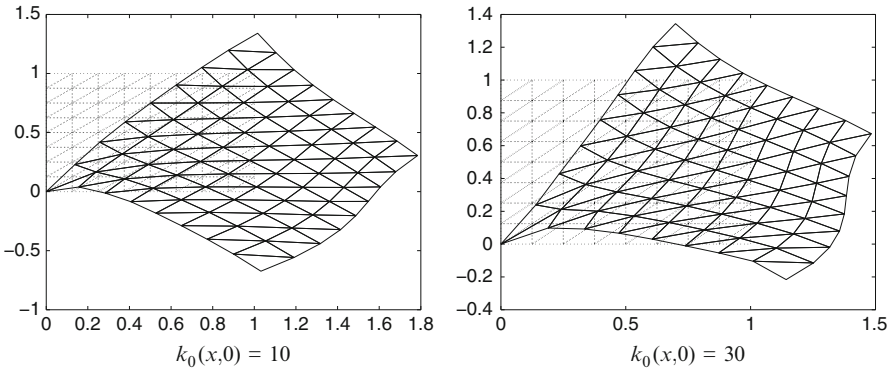


Fig. 13 Deformed configurations at final time, $\theta_R(t) = 0,\ 0 \leq t \leq 1$

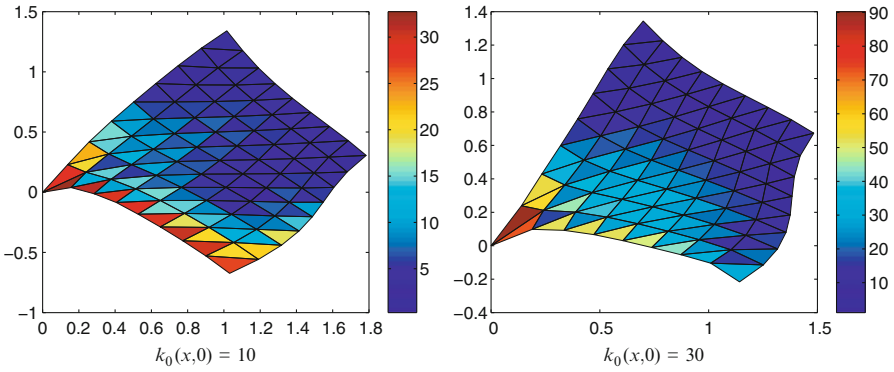


Fig. 14 Von Mises' norm in deformed configurations, $\theta_R(t) = 0,\ 0 \leq t \leq 1$

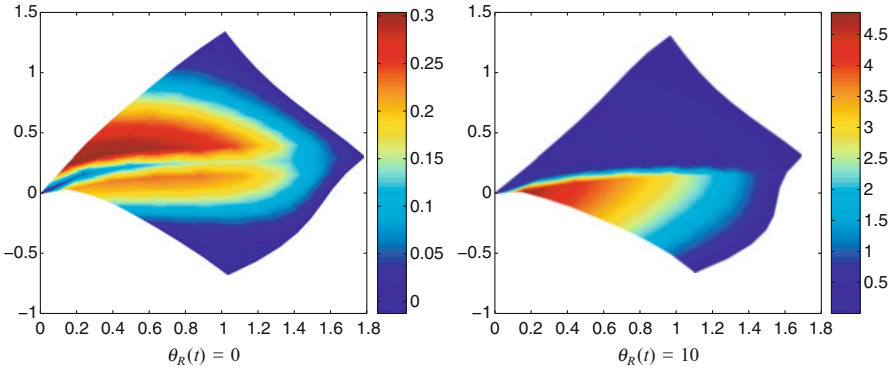


Fig. 15 Temperature field at final time, $k_0(x,0) = 10,\ 0 \le x \le 1$

$$\Omega = (0,L_1) \times (0,L_2),$$
$$\Gamma_1 = \{(0,0)\}; \quad \Gamma_2 = (\{0\}\times]0,L_2]) \cup (]0,L_1] \times \{L_2\}) \cup (\{L_1\} \times [0,L_2]),$$
$$\Gamma_3 =]0,L_1[\times\{0\},$$
$$f_2(x,t) = (0,0), \quad \forall x \in (\{0\}\times]0,L_2]), \quad \forall t \in [0,T],$$
$$f_2(x,t) = (1,0), \quad \forall x \in (]0,L_1] \times \{L_2\}) \cup (\{L_1\} \times [0,L_2]), \quad \forall t \in [0,T].$$

Similar conclusions as in Sect. 2 can be stated. See also Figs. 9, 10, 11, 12, 13, 14, and 15.

3.2 Numerical Simulations B

Here $\Gamma_1 = \emptyset$, $\Gamma_F = \Gamma_2$, $f_F = f_2$, $\Gamma_u = \Gamma_3$ (see [2]). We take again the two typical examples as in Sect. 2 with the following data:

$$\Omega = (0,L_1) \times (0,L_2),$$
$$\Gamma_F = (\{0\} \times [0,L_2]) \cup ([0,L_1] \times \{L_2\}) \cup (\{L_1\} \times [0,L_2]), \quad \Gamma_u =]0,L_1[\times\{0\},$$
$$f_2(x,t) = (0,0), \quad \forall x \in \{0\} \times [0,L_2], \quad \forall t \in [0,T],$$
$$f_2(x,t) = (1,0), \quad \forall x \in (]0,L_1] \times \{L_2\}) \cup (\{L_1\} \times [0,L_2]), \quad \forall t \in [0,T].$$

Analogous conclusions as in Sect. 2 can be stated. See also Figs. 16, 17, 18, 19, 20, 21, and 22.

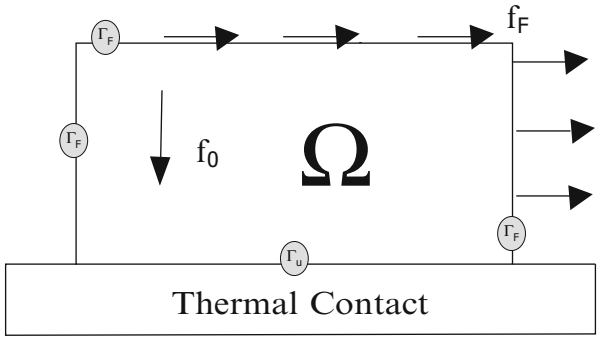


Fig. 16 Initial configuration

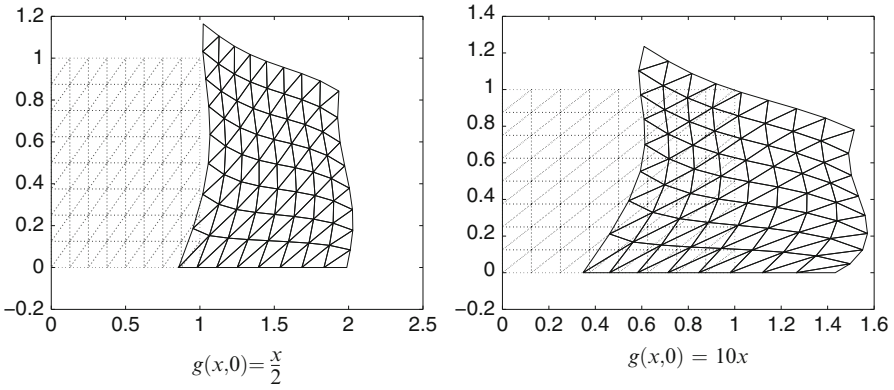


Fig. 17 Deformed configurations at final time, $\theta_R(t) = 10,\ 0 \le t \le 1$

4 A Duality Numerical Method

The fully discrete scheme (36) is equivalent to the variational inequality

$$\langle \mathcal{A} v_n^{hk}, w^h - v_n^{hk} \rangle + \psi(w^h) - \psi(v_n^{hk}) \ge \langle \mathcal{L}, w^h - v_n^{hk} \rangle_{V' \times V}, \forall w^h \in V^h \tag{60}$$

where operator $\mathcal{A} : V \to V'$ is defined by

$$\langle \mathcal{A} v, w \rangle = \left(\frac{v}{k}, w \right)_H + \langle A\, v, w \rangle_{V' \times V}, \quad \forall w \in V^h, \tag{61}$$

and $\mathcal{L} : V \to \mathbb{R}$ is defined by

$$\begin{aligned} \langle \mathcal{L}, w \rangle = & \langle \frac{v_{n-1}^{hk}}{k} - B\, u_{n-1}^{hk} - C\, \theta_{n-1}^{hk} + f(t_n), w \rangle_{V' \times V} \\ & - \left(k \sum_{m=0}^{n-1} \mathcal{B}(t_n - t_m)\, \varepsilon(u_m^{hk}), \varepsilon(w) \right)_{\mathcal{H}}, \quad \forall w \in V^h. \end{aligned} \tag{62}$$

For clearness we drop the indexes and consider in the sequel the problem:

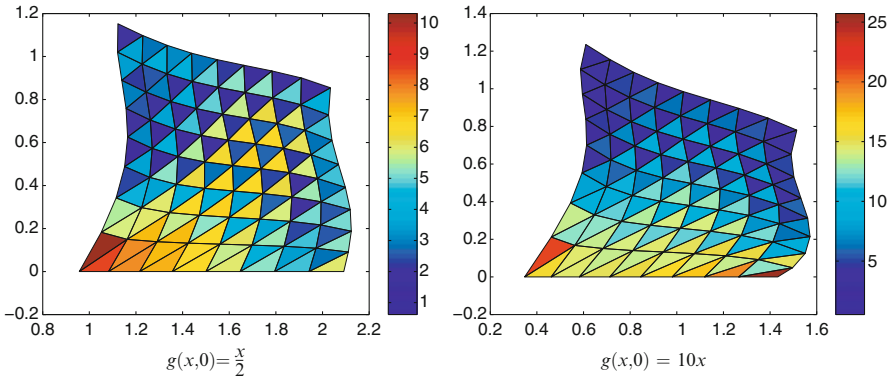


Fig. 18 Von Mises' norm in deformed configurations, $\theta_R(t) = 10,\ 0 \le t \le 1$

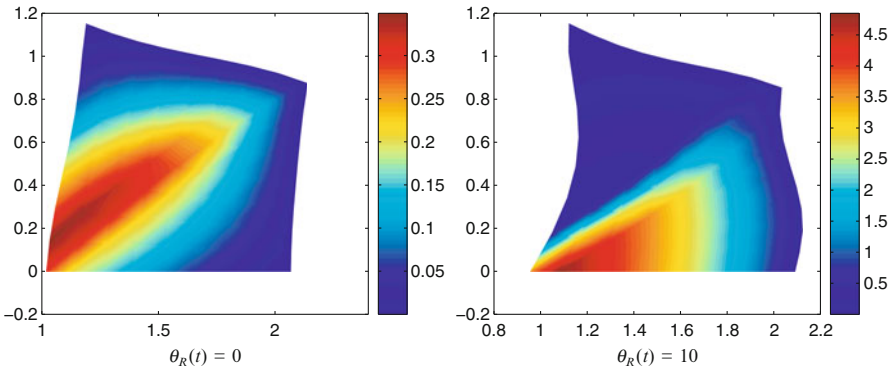


Fig. 19 Temperature field at final time, $g(x,0) = \frac{x}{2},\ 0 \le x \le 1$

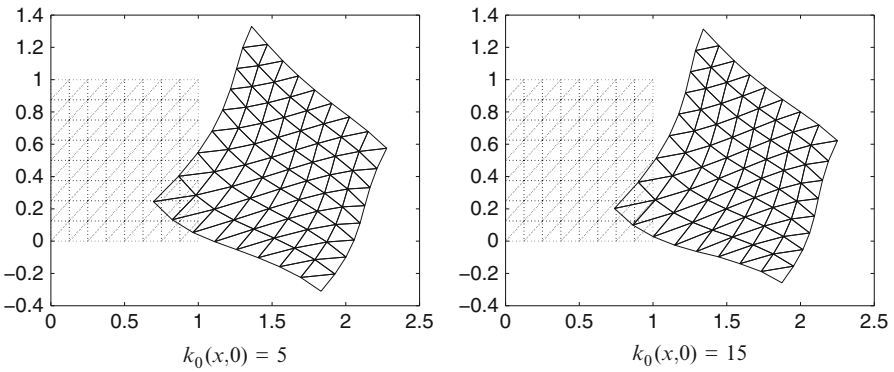


Fig. 20 Deformed configurations at final time, $\theta_R(t) = 10,\ 0 \le t \le 1$

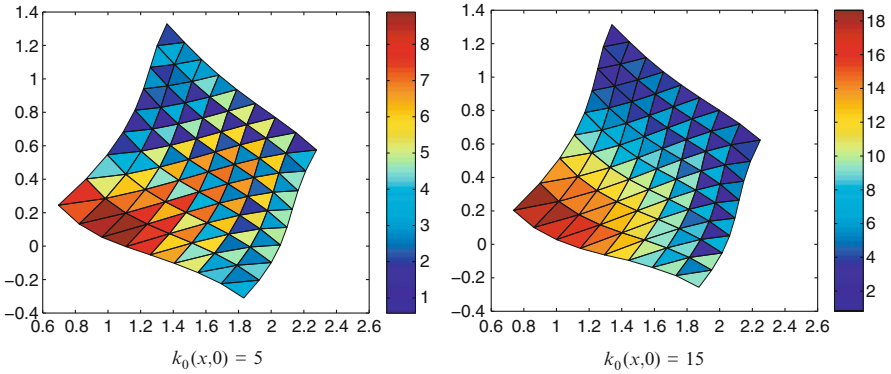


Fig. 21 Von Mises' norm in deformed configurations, $\theta_R(t) = 10,\ 0 \le t \le 1$

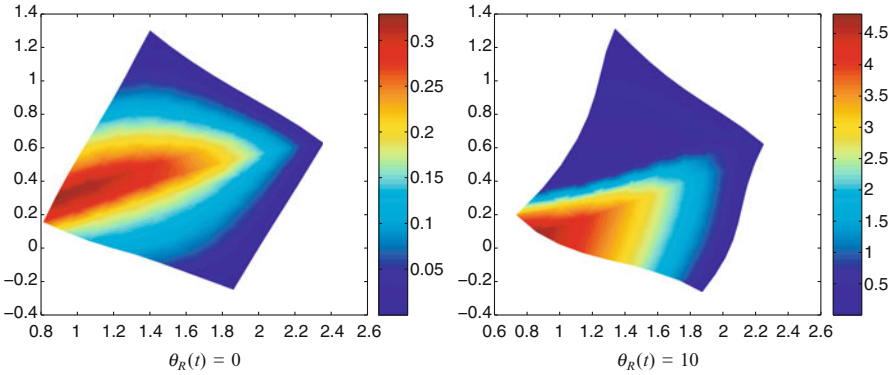


Fig. 22 Temperature field at final time, $k_0(x,0) = 15,\ 0 \le x \le 1$

Find $v \in V^h$ such that

$$\langle \mathcal{A}v, w - v\rangle + \psi(w) - \psi(v) \ge \langle \mathcal{L}, w - v\rangle_{V' \times V}, \quad \forall w \in V^h. \tag{63}$$

Numerical approach of (63) can be placed in the frame of duality methods for variational inequalities [4]. These methods are based on some classical results for monotone maps given in [6, 24] for instance. For convenience, we give first a brief introduction to the monotonous maximal operator theory.

Let G be a maximal monotone multivalued map on a Hilbert space H, and let λ be a nonnegative parameter. It can be proved that for all $f \in H$ there exists a unique $y \in H$ such that $f \in (I + \lambda G)(y)$. The single-valued map $J_\lambda^G = (I + \lambda G)^{-1}$ is a well defined and contraction map on H, and its called the resolvent operator of G (see [6]).

The map

$$G_\lambda = \frac{I - J_\lambda^G}{\lambda}$$

is called the Moreau–Yosida approximation of G. It is a maximal monotone, single-valued, and $\frac{1}{\lambda}$-Lipschitz continuous map. Moreover, G_λ satisfies the following important property on which is based our method.

Lemma 1. *Let be G a maximal monotone map on a Hilbert space H and G_λ, with $\lambda > 0$, its Yosida approximation. Then for all y and u in H, we have*

$$u \in G(y) \iff u = G_\lambda(y + \lambda u). \tag{64}$$

Proof. Let be $u = G_\lambda(x)$. Then

$$\begin{aligned} u = \frac{I - J_\lambda^G}{\lambda} x &\iff \lambda u = x - J_\lambda x \\ &\iff J_\lambda x = x - \lambda u \\ &\iff x \in (I + \lambda G)(x - \lambda u) = x - \lambda u + \lambda G(x - \lambda u) \\ &\iff \lambda u \in \lambda G(x - \lambda u) \\ &\iff u \in G(x - \lambda u) \end{aligned}$$

and by taking $x - \lambda u = y$ we get: $u \in G(y) \iff u = G_\lambda(y + \lambda u)$ □

Now returning to problem (63), the frictional contact function ψ is continuous and convex , and therefore its subdifferential $\partial\psi$ is a maximal monotone operator in V_h, and (63) can be written using the subdifferential operator:

Find $v \in V^h$ such that

$$\begin{cases} \langle \mathcal{A}v, w\rangle + (\gamma, v)_{L^2(\Gamma_3)} = \langle \mathcal{L}, w\rangle_{V' \times V}, \ \forall w \in V^h, \\ \gamma \in \partial\psi(v). \end{cases} \tag{65}$$

Using relation (64), we obtain equivalently:

Find $v \in V^h$ such that

$$\begin{cases} \langle \mathcal{A}v, w\rangle + (\gamma, w)_{L^2(\Gamma_3)} = \langle \mathcal{L}, w\rangle_{V' \times V}, \ \forall w \in V^h, \\ \gamma = (\partial\psi)_\lambda(v + \lambda\gamma), \end{cases} \tag{66}$$

where $\lambda > 0$ and $(\partial\psi)_\lambda$ is the Yosida approximation of $\partial\psi$.

Thereby, we apply the following algorithm to solve (66).

(0) Start with some arbitrary value of the multiplier γ^0.

(1) For γ^j known, compute v_j solution to

$$\langle \mathcal{A}v^j, w\rangle + (\gamma^j, w)_{L^2(\Gamma_3)} = \langle \mathcal{L}, w\rangle_{V' \times V}, \ \forall w \in V^h. \tag{67}$$

(2) Update multiplier γ^j as

$$\gamma^{j+1} = (\partial\psi)_\lambda(v^j + \lambda\gamma^j). \tag{68}$$

(3) Go to (1) until stop criterion is reached.

Theorem 5. *If $\mathcal{A}$ is an elliptic operator and $\lambda > \frac{1}{2\alpha}$, where α is the elliptic constant of $\mathcal{A}$, we have*

$$\lim_{j\to\infty} \|v^j - v\| = 0.$$

Proof. The mapping $(\partial\psi)_\lambda$ is $\dfrac{1}{\lambda}$-Lipschitz and thus

$$\begin{aligned}
\|\gamma - \gamma^{j+1}\|^2 &= \|(\partial\psi)_\lambda(v + \lambda\gamma) - (\partial\psi)_\lambda(v^j + \lambda\gamma^j)\|^2 \\
&\leq \frac{1}{\lambda^2}\|(v + \lambda\gamma) - (v^j + \lambda\gamma^j)\|^2 \\
&= \frac{1}{\lambda^2}\|(v - v^j) + \lambda(\gamma - \gamma^j)\|^2 \\
&= \frac{1}{\lambda^2}\|v - v^j\|^2 + \frac{2}{\lambda}((v - v^j), (\gamma - \gamma^j)) + \|\gamma - \gamma^j\|^2.
\end{aligned}$$

Therefore

$$\|\gamma - \gamma^j\|^2 - \|\gamma - \gamma^{j+1}\|^2 \geq -\frac{1}{\lambda^2}\|v - v^j\|^2 - \frac{2}{\lambda}((v - v^j), (\gamma - \gamma^j)) \tag{69}$$

Using now (66) and (67), we obtain

$$\langle \mathcal{A}(v - v^j), w\rangle + (\gamma - \gamma^j, w) = 0,\ \forall w \in V^h.$$

Thus

$$\begin{aligned}
\alpha\|v - v^j\|^2 &\leq \langle \mathcal{A}(v - v^j), (v - v^j)\rangle \\
&= -(\gamma - \gamma^j, v - v^j).
\end{aligned}$$

Substituting in (69) we get

$$\begin{aligned}
\|\gamma - \gamma^j\|^2 - \|\gamma - \gamma^{j+1}\|^2 &\geq -\frac{1}{\lambda^2}\|v - v^j\|^2 + \frac{2\alpha}{\lambda}\|v - v^j\|^2 \\
&= \frac{1}{\lambda}(2\alpha - \frac{1}{\lambda})\|v - v^j\|^2.
\end{aligned}$$

Recalling that $\lambda > \dfrac{1}{2\alpha}$, we obtain

$$\|\gamma - \gamma^j\|^2 - \|\gamma - \gamma^{j+1}\|^2 \geq \|v - v^j\|^2 \geq 0.$$

The sequence $(\|\gamma-\gamma^j\|^2)_{j\geq 0}$ is decreasing and positive, therefore

$$\lim_{j\to\infty}\|\gamma^j-\gamma\|^2=0$$

and finally

$$\lim_{j\to\infty}\|v^j-v\|^2=0.$$

□

Remark 2. Under symmetric property of operator $\mathcal{A}$ and if ψ is differentiable this algorithm is the Uzawa one to reach the sadlle-point of the Lagrangien:

$$L(v,q)=\frac{1}{2}\langle\mathcal{A}v,v\rangle-\langle\mathcal{L},w\rangle+(q,\psi(v))$$

$$L(u,q)\leq L(u,p)\leq L(v,p).$$

Now we turn out to determine the Yosida approximation $(\partial\psi)_\lambda$. Note first that

$$(\partial\psi)_\lambda=((\partial\psi)^{-1}+\lambda I)^{-1}. \tag{70}$$

Indeed, $\forall u,x\in V$ we have from (64):

$$\begin{aligned}u=(\partial\psi)_\lambda(x)&\Longleftrightarrow u\in(\partial\psi)(x-\lambda u)\\&\Longleftrightarrow(x-\lambda u)\in(\partial\psi)^{-1}(u)\\&\Longleftrightarrow x\in(\lambda I+(\partial\psi)^{-1})(u)\\&\Longleftrightarrow u=(\lambda I+(\partial\psi)^{-1})^{-1}(x).\end{aligned}$$

And from (70) we note that the Yosida approximation of $(\partial\psi)$ and the resolvent of $(\partial\psi)^{-1}$ are linked by

$$(\partial\psi)_\lambda(x)=J_{1/\lambda}^{\partial\psi^{-1}}\left(\frac{x}{\lambda}\right). \tag{71}$$

Let be $u,x\in V$, we have

$$\begin{aligned}u=(\partial\psi)_\lambda(x)&\Longleftrightarrow u=((\partial\psi)^{-1}+\lambda I)^{-1}(x)\\&\Longleftrightarrow x\in((\partial\psi)^{-1}+\lambda I)(u)\\&\Longleftrightarrow\frac{x}{\lambda}\in\left(\frac{1}{\lambda}(\partial\psi)^{-1}+I\right)(u)\\&\Longleftrightarrow u=\left(\frac{1}{\lambda}(\partial\psi)^{-1}+I\right)^{-1}(x/\lambda)=J_{1/\lambda}^{\partial\psi^{-1}}(x/\lambda).\end{aligned}$$

Definition 1. The map

$$\begin{aligned}\psi^*:V&\longrightarrow\mathbb{R}\\x&\longmapsto\sup_{y\in V}\{(x,y)_{L^2(\Gamma_3)}-\psi(y)\}\end{aligned}$$

is the Fenchel conjugate of ψ.

Theorem 6. *The Fenchel conjugate ψ^* is well defined, continuous and convex. Its subdifferential is the inverse subdifferential $(\partial\psi)^{-1}$ of ψ, and we have*

$$y \in (\partial\psi)(x) \iff x \in (\partial\psi^*)(y). \tag{72}$$

From (71) and (72), we get the equality linking the resolvent of $\partial\psi$ and the Yosida approximation of $(\partial\psi^*)$,

$$J_\lambda^{\partial\psi}(x) = (\partial\psi^*)_{\frac{1}{\lambda}}(\lambda x). \tag{73}$$

Let us set:

$$K = \{f \in L^2(\Gamma_3) : (f,w) - \psi(w) \le 0,\ \forall w \in L^2(\Gamma_3)\}.$$

Theorem 7. *The Fenchel conjugate ψ^* satisfies*

$$\psi^* = I_K, \quad \text{on} \quad L^2(\Gamma_3), \tag{74}$$

where I_K is the indicator function of K.

Proof. Let $f \in L^2(\Gamma_3)$ be given. There are two possibilities. If there exists $w \in L^2(\Gamma_3)$ such that

$$(f,w) - \psi(w) > 0,$$

then for $r > 0$:

$$(f,rw) - \psi(rw) = r\big((f,w) - \psi(w)\big),$$

and

$$\psi^*(f) = \sup_{w \in L^2(\Gamma_3)} \{(f,y) - \psi(w)\} = +\infty.$$

If such a w does not exist then

$$(f,w) - \psi(w) \le 0,\ \forall w \in L^2(\Gamma_3),$$

but this quantity vanishes for $w = 0$, so that:

$$\psi^*(f) = 0$$

□

Consequently, we can compute $(\partial\psi^*)$ as (∂I_K) and we obtain

$$(\partial\psi^*)(y) = (\partial I_K)(y) = N_K(y),\ \forall y \in K, \tag{75}$$

where

$$N_K(y) = \{f \in L^2(\Gamma_3) : (f, w-y) \le 0,\ \forall w \in K\}.$$

is the normal cone of K in y.

On the other hand, we can easily prove that the Yosida approximation of (∂I_K) is

$$(\partial I_K)_\lambda = \frac{I - P_K}{\lambda}, \tag{76}$$

where P_K is the projection operator on K.

Now, taking into account (73), (75), and (76) we have for all $x \in V$

$$\begin{aligned} J_\lambda^{\partial\psi}(x) &= (\partial\psi^*)_{\frac{1}{\lambda}}(\lambda x) \\ &= (\partial I_K)_{\frac{1}{\lambda}}(\lambda x) \\ &= \lambda(I - P_K)(\lambda x) \\ &= \lambda^2 x - \lambda P_K(\lambda x). \end{aligned}$$

We get finally

$$\begin{aligned} (\partial\psi)_\lambda(x) &= \frac{I - J_\lambda^{\partial\psi}}{\lambda}(x) \\ &= \frac{x - \lambda^2 x + \lambda P_K(\lambda x)}{\lambda} \\ &= \frac{1 - \lambda^2}{\lambda} x + P_K(\lambda x). \end{aligned}$$

Therefore the multiplier γ^j in (68) is updated by the formula

$$\gamma^{j+1} = \frac{1 - \lambda^2}{\lambda}(v^j + \lambda\gamma^j) + P_K(\lambda v^j + \lambda^2\gamma^j). \tag{77}$$

References

1. Addi, K., Chau, O., Goeleven, D.: On some frictional contact problems with velocity condition for elastic and visco-elastic materials. Discrete Continuous Dyn. Syst. **31**(4), 1039–1051 (2011)
2. Adly, S., Chau, O.: On some dynamical thermal non clamped contact problems, Mathematical Programming, serie B, (2013) doi: 10.1007/s10107-013-0657-9
3. Adly, S., Chau, O., Rochdi, M.: Solvability of a class of thermal dynamical contact problems with subdifferential conditions. Numer. Algebra Control Optim. **2**(1), 89–101 (2012)
4. Bermudez, A., Moreno, C.: Duality methods for solving variational inequalities. Comput. Math. Appl. **7**, 43–58 (1981)
5. Brézis, H.: Problèmes unilatéraux. J. Math. Pures Appl. **51**, 1–168 (1972)
6. Brezis, H.: Operateurs Maximaux Monotones et Semigroups de Contractions dans les Espaces de Hilbert. North-Holland, Amsterdam (1973)
7. Brézis, H.: Analyse fonctionnelle, Théorie et Application. Masson, Paris (1987)
8. Chau, O., Awbi, B.: Quasistatic thermovisoelastic frictional contact problem with damped response. Appl. Anal. **83**(6), 635–648 (2004)
9. Chau, O., Motreanu, D., Sofonea, M.: Quasistatic frictional problems for elastic and viscoelastic materials. Appl. Math. **47**(4), 341–360 (2002)

10. Chau, O., Goeleven, D., Oujja, R.: A numerical treatment of a class variational inequalities 567 arising in the study of viscoelastic materials. Int. J. Appl. Math. Mech. (to appear)
11. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North Holland, Amsterdam (1978)
12. Ciarlet, P.G.: Mathematical Elasticity, Vol. I : Three-Dimensional Elasticity. North-Holland, Amsterdam (1988)
13. Duvaut, G., Lions, J.L.: Les Inéquations en Mécanique et en Physique. Dunod, Paris (1972)
14. Ekeland, I., Teman, R.: Analyse Convexe et Problèmes Variationelles. Gautier-Villars, Paris (1984)
15. Glowinski, R.: Numerical Methods for Nonlinear Variational Problems. Springer, New York (1984)
16. Goeleven, D., Motreanu, D., Dumont, Y., Rochdi, M.: Variational and Hemivariational Inequalities, Theory, Methods and Applications, Volume I: Unilateral Analysis and Unilateral Mechanics. Kluwer Academic, Boston (2003)
17. Han, W., Reddy, B.D.: Plasticity: Mathematical Theory and Numerical Analysis. Springer, New York (1999)
18. Jeantet, R., Croguennec, T., Schuck, P., Brul, G.: Science des Aliments. Lavoisier, Paris (2006)
19. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity : A Study of Variational Inequalities and Finite Element Methods. SIAM, Philadelphia (1988)
20. Lions, J.L.: Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires. Dunod et Gauthier-Villars, Paris (1969)
21. Nečas, J., Hlavaček, I.: Mathematical Theory of Elastic and Elastoplastic Bodies: An Introduction. Elsevier, Amsterdam (1981)
22. Panagiotopoulos, P.D.: Inequality Problems in Mechanical and Applications. Birkhauser, Basel (1985)
23. Panagiotopoulos, P.D.: Hemivariational Inequalities, Applications in Mechanics and Engineering. Springer, Berlin (1993)
24. Pazy, A.: Semigroups of non-linear contractions in Hilbert spaces. In: Problems in Nonlinear Analysis. C.I.M.E. Ed. Cremonese, Roma (1971)
25. Raous, M., Jean, M., Moreau, J.J. (eds.): Contact Mechanics. Plenum Press, New York (1995)
26. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. Springer, New York (1997)

Neighboring Local-Optimal Solutions and Its Applications

Hsiao-Dong Chiang and Tao Wang

1 Introduction

Local-optimal solutions are of fundamental importance to the study of nonlinear optimization, which also closely resemble some concepts in biochemistry and electrical power engineering. The present work on local-optimal solutions is closely related to the various studies of complexity. Indeed, an effective tool for our study is provided by Sperner's lemma [1–4], which is stated as follows.

Theorem 1 (Sperner's Lemma). *Every Sperner labeling of a triangulation of an n-dimensional simplex contains a fully labeled cell that is labeled with a complete set of the labels.*

This lemma yields the Brouwer fixed point theorem [5] and plays an important role in the proof of Monsky's theorem [6] that a square cannot be cut into an odd number of equal-area triangles. On the finite covering by non-polyhedral closed sets, there are also many interesting corollaries [7, 8] derived from Sperner's lemma, and a proposition (see [7, Lemma 2–26]) is rephrased below.

Theorem 2. *Consider an n-dimensional simplex* Ω^*, *and* $(n+1)$ *closed sets* $V_k \subseteq \mathbb{R}^n, 0 \leq k \leq n$. *Let* $\{v_k; 0 \leq k \leq n\}$ *be the set of vertices of* Ω^*, *and* Q_k *be the* $(n-1)$*-dimensional face of* Ω^* *opposite to the vertex* v_k. *Suppose that* $\Omega^* \subseteq \bigcup_{i=0}^{n} V_n$, *and* $(Q_k \cap V_k) = \emptyset$ *with* $v_k \in V_k$ *for all* $0 \leq k \leq n$. *Then, the intersection* $(\bigcap_{k=0}^{n} V_k) \neq \emptyset$.

On a proof of Theorem 2, the key ingredient (see Fig. 1) is that every point in V_k is labeled by L_k, and under the specified conditions the Sperner's lemma yields the existence of a fully labeled cell in an arbitrary finite triangulation of S. By making the

H-.D. Chiang • T. Wang (✉)
School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA
e-mail: chiang@ece.cornell.edu; tw355@cornell.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_3,

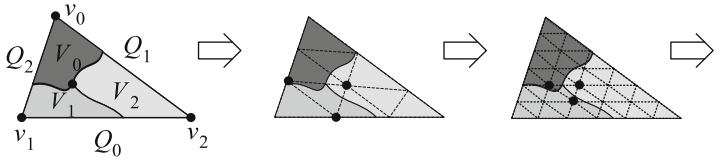


Fig. 1 Theorem 2 can be proved by Sperner's lemma, and the fully labeled cells are indicated by *bold dots* in the triangulations

triangulation smaller and smaller, one can easily show that the collection of fully labeled cells contains a convergent subsequence whose limit is a common point shared by V_k's. Indeed, Sperner's lemma and Theorem 2 can be alternatively interpreted in engineering design and nonlinear optimization, and motivate the study of the algebraic structure of the collection of local-optimal solutions.

The present work is devoted to estimating the number of neighboring local-optimal solutions, which provides an important index for evaluating the complexity of nonlinear systems in biochemistry and electrical engineering, and the computational complexity of solution methods for nonlinear optimization. First of all, we show that there are at least $2n$ local-optimal solutions neighboring to the given solution, if the corresponding gradient system of the optimization problem is spatially periodic in $\mathbb{R}^n$. Here the gradient is called spatially periodic, if it repeats the values in regular intervals or periods along n linearly independent directions. On the lower bound, it is expected that an improved estimation $n(n+1)$ can be obtained by investigating the local-independence of a collection of neighboring local-optimal solutions. The local-independence has been proved for $n = 2$, which is followed by an example for validating the derived bounds. Moreover, some engineering applications are elaborated at last, for interpreting Sperner's lemma and the present study.

2 Mathematical Preliminaries

2.1 Nonlinear Optimization and Local-Optimal Solution

We consider an optimization (minimization) problem of the form

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where the function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable over $\mathbb{R}^n$. By convention, a point x^* is called a *local-optimal solution* of (1) if there is a neighborhood $U \subseteq \mathbb{R}^n$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in U$. It should be apparent that any unconstrained maximization problem can be directly converted to the form (1), by negating the objective function. In addition, the gradient $\nabla f(x) = 0$, at a local-optimal solution x_*. When the determinant of Hessian matrix $\det(\nabla^2 f) \neq 0$ at a point x_*, one has that

the point x^* is a local-optimal solution for (1), if and only if the point x^* is a stable equilibrium point of the gradient system $\dot{x} = -\nabla f(x(t))$.

Therefore, we can use the term "local-optimal solution" and "stable equilibrium point" interchangeably, without causing any confusion. Moreover, two local-optimal solutions are called neighboring to each other, if their stability regions of the gradient system intersect on the boundary. For the constrained problems, we shall examine the projected gradient system [9–11].

2.2 Gradient System and Equilibrium Point

To define the neighboring solutions, we must introduce the gradient system of (1), say

$$\dot{x}(t) = F(x(t)) = -\nabla f(x(t)) \in \mathbb{R}^n, \tag{2}$$

where the state vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, and $F(x) = (F_1(x), \ldots, F_n(x))$. Here, F_i is a scalar function, for all $1 \leq i \leq n$. The solution of (2) starting from $x_0 \in \mathbb{R}^n$ at $t = 0$ is called a *trajectory* and denoted by $\phi(\cdot, x_0) : \mathbb{R} \mapsto \mathbb{R}^n$.

A state vector $x^* \in \mathbb{R}^n$ is called an *equilibrium point* of (2), if $F(x^*) = 0$. In addition, an equilibrium point $x^* \in \mathbb{R}^n$ is *hyperbolic*, if the Jacobian matrix of $F(\cdot)$ at x^* has no eigenvalues with zero real part, which implies $\det(\nabla^2 f(x^*)) \neq 0$. Furthermore, a *type-k* equilibrium point refers to a hyperbolic equilibrium point at which the Jacobian has exactly k eigenvalues with positive real part. In particular, a hyperbolic equilibrium point is called a (*asymptotically*) *stable equilibrium point* if at the point each eigenvalue of the Jacobian has negative real part, while it is called an *unstable equilibrium point* if all the eigenvalues have a positive real part, which are an equilibrium point of type-0 and of type-n respectively.

Given a type-k equilibrium point x^*, its stable manifold $W^s(x^*)$ and unstable manifold $W^u(x^*)$ are defined as,

$$W^s(x^*) \doteq \{x \in \mathbb{R}^n : \lim_{t \to \infty} \phi(t, x) = x^*\},$$
$$W^u(x^*) \doteq \{x \in \mathbb{R}^n : \lim_{t \to -\infty} \phi(t, x) = x^*\}$$

where the dimension of $W^u(x^*)$ and $W^s(x^*)$ are k and $(n-k)$, respectively. The *stability region* (or *region of attraction*) of stable equilibrium point x_s is

$$A(x_s) \doteq \{x \in \mathbb{R}^n : \lim_{t \to \infty} \phi(t, x) = x_s\}.$$

As mentioned earlier, there is a one-to-one correspondence between the stable equilibrium points of (2) and the local-optimal solutions of (1) under the hyperbolic assumption, and then the neighboring solutions are well defined as follows. Consider two local-optimal solutions x_s' and x_s of the problem (1), we say that the point x_s' is a local-optimal solution *neighboring* to x_s, if the closure of stability region $A(x_s)$ intersects that of $A(x_s')$, i.e., the set $(\overline{A(x_s')} \cap \overline{A(x_s)}) \neq \emptyset$. Accordingly, such $A(x_s')$ is called a stability region neighboring to $A(x_s)$. Here $\overline{A}$ denotes the closure of A. Apparently, a stability region is uniquely determined by a stable equilibrium

point. We thus can just estimate the number of neighboring stability regions when necessary.

Nevertheless, the structure of stability boundary $\partial A(x_s)$ for the nonlinear system (2) is complex in general, and the quasi-stability boundary is commonly studied instead. Indeed, the *quasi-stability boundary* $\partial A_p(x_s)$ of a stable equilibrium point x_s is defined by $\partial\overline{A(x_s)}$, and the *quasi-stability region* $A_p(x_s)$ is the open set $\text{int}(\overline{A(x_s)})$, where $\text{int}(\cdot)$ refers to the interior. It is known that the quasi-stability region $A_p(x_s) \subseteq \overline{A_p(x_s)} \subseteq \overline{A(x_s)}$, and the quasi-stability boundary $\partial A_p(x_s) \subseteq \partial A(x_s)$. We shall show that the neighboring local-optimal solutions can be equivalently defined by quasi-stability boundaries, which relies on a general proposition [12].

Proposition 1 (*Intersection of Quasi-Stability Boundary*). *Let $x_s, x_s' \in \mathbb{R}^n$ be two distinct stable equilibrium points of (2). On the quasi-stability boundary, one has $(\partial A_p(x_s) \cap \partial A_p(x_s')) = (\overline{A(x_s)} \cap \overline{A(x_s')})$.*

Proof: See Appendix.

Proposition 1 suggests an equivalent definition of the neighboring solution.

Remark 1. Given two distinct local-optimal solutions x_s, x_s' of (1), the solution x_s' is neighboring to x_s, if and only if $(\partial A_p(x_s) \cap \partial A_p(x_s')) \neq \emptyset$.

Hence, a stability boundary will always refer to the quasi-stability boundary ∂A_p, without causing any confusion.

2.3 Characterization of Stability Boundary

Prior to introducing the characterization of stability boundary, we recall that a set $K \subseteq \mathbb{R}^n$ is *invariant* regarding the dynamics at (2), if every trajectory of (2) starting in K stays in K for all $t \in \mathbb{R}$. By the definition, the stable manifold is always invariant. Moreover, for two submanifolds M_1 and M_2 of a manifold M, they meet the *transversality condition*, if either (1) $(M_1 \cap M_2) = \emptyset$, or (2) at every point $y \in (M_1 \cap M_2)$, the tangent spaces of M_1 and M_2 span the tangent spaces of M at y.

On the stability boundary, we make the following assumptions.

(A1) *All the equilibrium points are hyperbolic and are finite in number on a stability boundary.*

(A2) *The stable and unstable manifolds of equilibrium points on the stability boundary satisfy the transversality condition.*

(A3) *Every trajectory approaches an equilibrium point as $t \to +\infty$.*

Here (A1) and (A2) are generic properties [13] for nonlinear dynamical systems. Moreover, (A3) is not generic; however, it is satisfied by a large class of nonlinear dynamical systems, as the electric power system. To study the neighboring solutions, we need the characterization theorems.

Theorem 3. *(Theorem 4.2 [14]: Complete Characterization of Quasi-Stability Boundary) Consider a stable equilibrium point x_s of the nonlinear dynamical system (2) satisfying the assumptions (A1)–(A3). Let $x_e^i, i \in \mathbb{N}$ be the equilibrium points on the quasi-stability boundary $\partial A_p(x_s)$. Then, the quasi-stability boundary*

$$\partial A_p(x_s) = \bigcup_{x_e^i \in \partial A_p(x_s)} W^s(x_e^i).$$

This implies that the intersection of stability boundaries is also the union of stable manifolds of the equilibrium points in the intersection.

2.4 Spatially Periodic Dynamical Systems

A function $F = (F_1, F_2, \ldots, F_m) : \mathbb{R}^n \mapsto \mathbb{R}^m$ is called spatially periodic [15–19], if there exist n constants $p_i > 0$ for $1 \le i \le n$, such that $F_j(x) = F_j(x + p_i e_i)$ for all $x \in \mathbb{R}^n$ and $1 \le j \le m$. It is worthwhile noting that F_j is a scalar function, and e_i denotes the vector in $\mathbb{R}^n$ with 1 in the ith coordinate and 0's elsewhere. In addition, given a spatially periodic function $F(x)$, an n-tuple $p_* = (p_1^*, p_2^*, \ldots, p_n^*)$ is called the spatial periods, if each $p_i^* > 0$ is the minimum positive number p_i such that $F_j(x) = F_j(x + p_i e_i)$ for all $x \in \mathbb{R}^n, 1 \le j \le m$. In literature, there have been many reports on the applications of spatially periodic dynamics and systems [15–19] in physics, chemistry, and electrical engineering, etc. Moreover, the dynamical system (2) is spatially periodic if the gradient ∇f is spatially periodic.

Indeed, a spatially periodic function with $p_i^* \neq 2\pi$ can be transformed into a function having $p_i^* \equiv 2\pi$ for all $1 \le i \le n$. More precisely, given a spatially periodic function $f_1(x)$ with spatial periods p_*, it is easy to check that the function $f_2(x) = f_1(x \otimes p_*/2\pi) \doteq f_1(x_1 p_1^*/2\pi, \ldots, x_n p_n^*/2\pi)$ is also spatially periodic, with the spatial period $= 2\pi$ for all x_i's. This suggests, without loss of generality we can assume $p_i^* = 2\pi$ for all $1 \le i \le n$, if the system is spatially periodic.

Additional hypotheses are imposed on the system (2).

(A4) *The system (2) is spatially periodic with the spatial period $p_i^* = 2\pi$ for $1 \le i \le n$. Moreover, there is at most one stable equilibrium point in each region of the form $\Pi_{i=1}^n [x_i, x_i + 2\pi) \subseteq \mathbb{R}^n$, for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$.*

In other words, if x_s is a stable equilibrium point of (2), then any stable equilibrium point $\tilde{x}_s \in \mathbb{R}^n$ can be represented by $\tilde{x}_s = x_s + \zeta$, for some $\zeta \in \mathcal{P}$. By the condition (A4), we can write the set $\mathcal{P} \doteq 2\pi\mathbb{Z}^n$, where $\mathbb{Z}^n$ is the n-dimensional integer lattice. Moreover, a vector $\zeta \in \mathcal{P}$ is called a *(spatial-) period vector*, and the set $\mathcal{P}$ is the collection of period vectors.

(A5) *Every stability region $A(x_s)$ is bounded.*

The boundedness assumption in (A5) ensures that all the stability regions are uniformly bounded for the given spatially periodic gradient system.

3 Symmetry and Number of Neighboring Local-Optimal Solutions

In this section, we derive a lower bound on the number of neighboring local-optimal solutions [12]. The key propositions are presented below, and their proof and other intermediate results are contained in the appendix. It should be noted that the following analysis and propositions are presented under the hypotheses (A1)–(A5), until specified otherwise.

We introduce the translation operator $T_\zeta(x) \doteq x + \zeta$, for $x, \zeta \in \mathbb{R}^n$ and denote by $\mathcal{S}$ the set of all local-optimal solutions of (1). It is straightforward to see that the inverse $T_\zeta^{-1} = T_{-\zeta}$, and the set $\mathcal{S}$ coincides with the collection of all stable equilibrium points of (2). To begin with, the spatial-periodicity of (2) manifestly leads to the following proposition, and the proof is omitted.

Proposition 2 (*Spatial-Periodicity of Equilibrium Points*). *Let $\zeta \in \mathcal{P}$ be a spatial-period vector, and x_e be an equilibrium point of (2). Then, $T_\zeta(x_e)$ is also an equilibrium point of (2), and $T_\zeta(W^s(x_e))$ is the stable manifold of $T_\zeta(x_e)$. Moreover, the set $(T_\zeta(W^s(x_e)) \cap W^s(x_e)) = \emptyset$, and the closure $\overline{T_\zeta(W^s(x_e))} = T_\zeta(\overline{W^s(x_e)})$.*

By Proposition 2, the spatial-periodicity of the gradient system (2) yields the periodicity of local-optimal solutions in $\mathcal{S}$ and also that of the corresponding stability regions of (2). With the properties of extreme points on a convex hull [20, Corollaries 18.3.1 and 18.5.3], one can easily derive the proposition below.

Proposition 3 (*Existence of Extreme Point*). *For an arbitrary finite point set $\mathcal{E} = \{x^i; 1 \le i \le q\} \subseteq \mathbb{R}^n$, there always is a point $x^* \in \mathcal{E}$ (Fig. 2), such that*

$$\{T_\zeta(x^*), T_\zeta^{-1}(x^*)\} \setminus \mathcal{E} \neq \emptyset,$$

for all nonzero $\zeta \in \mathbb{R}^n$. Here, such x^* is called an extreme element of the set $\mathcal{E}$.

As an important application, the extreme element in Proposition 3 will serve as the center of symmetry in the derivation of the proposed lower bound.

By the definition, there is a one-to-one correspondence between the (neighboring) local-optimal solutions of (1) and the (neighboring) stability regions of (2). To estimate the number of neighboring solutions, we here estimate the number of

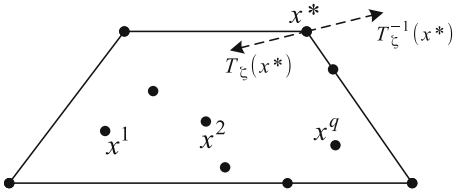


Fig. 2 An illustration for Proposition 3, where the *dots* are the points in $\mathcal{E}$

neighboring stability regions, as presented in Proposition 4. This proposition follows from the Sperner's lemma and Theorem 2, by constructing the closed sets V_k, $0 \leq k \leq n$.

Indeed, by Proposition 2 the collection $\mathscr{A}$ of all distinct stability regions of (2) must be countable and can be described as

$$\mathscr{A} = \{A_i; i \in \mathbb{N}\} = \{T_\zeta(A_s); \zeta \in \mathcal{P}\},$$

where the set A_s is any given stability region of (2). Moreover, we denote by $x_s^i \in \mathcal{S}$ the unique stable equilibrium point satisfying $A_i = A(x_s^i)$.

Proposition 4 (Existence of Neighboring Stability Regions). *There exist* $(n+1)$ *distinct stability regions* $\{A_{i_k}; 0 \leq k \leq n\} \subseteq \mathscr{A}$, *such that the intersection* $(\bigcap_{k=0}^{n} \overline{A}_{i_k}) \neq \emptyset$.

Proof: See Appendix.

This proposition shows that any given stability region must have at least n neighboring stability regions. Based on this intermediate result, we will further prove at Theorem 4 that $2n$ gives a general lower bound on the number of neighboring solutions. Without loss of generality, we assume that $i_k = k$ for $0 \leq k \leq n$ at Proposition 4, and x_s^0 is an extreme point in $\{x_s^k; 0 \leq k \leq n\}$ satisfying the property at Proposition 3. Clearly, A_0 is an extreme element in the collection

$$\mathscr{A}_n \doteq \{A_i; 0 \leq i \leq n\}. \tag{3}$$

Besides, we denote by $\zeta_i \in \mathcal{P}$ the vector such that $T_{\zeta_i}(A_i) = A_0$, or to say $T_{\zeta_i}^{-1}(A_0) = A_i$ for $1 \leq i \leq n$, with $\zeta_0 = 0$.

Proposition 5. *The elements in the augmented collection*

$$\mathscr{A}_n^* \doteq \{T_{\zeta_i}^{-1}(A_0); 1 \leq i \leq n\} \cup \{T_{\zeta_i}(A_0); 1 \leq i \leq n\} \tag{4}$$

are pairwise different, and $(\overline{A} \cap \overline{A}_0) \neq \emptyset$ *for all* $A \in \mathscr{A}_n^*$.

Proof: See Appendix.

The assertion on the augmented collection (4) directly shows that, for any local-optimal solution x_s of the problem (1), the stability region $A(x_s)$ has at least $2n$ neighboring stability regions. Thus, we are ready to state a theorem on the lower bound for the number of neighboring local-optimal solutions [24].

Theorem 4 (*Estimation Obtained by Symmetry*). *Consider an optimization problem* $\min_{x \in \mathbb{R}^n} f(x)$ *at (1), such that the objective* f *is twice-differentiable and the dynamical system* $\dot{x} = -\nabla f(x)$ *at (2) satisfies the conditions (A1)–(A5). Then, any local-optimal solution* x_s *of (1) has no less than* $2n$ *neighboring local-optimal solutions.*

Proof: Apparently, the number of neighboring stability regions gives a lower bound on the neighboring local-optimal solutions. In the remainder of the proof, we estimate the number of neighboring stability regions.

In light of Propositions 4 and 5, there are $(n+1)$ stability regions, say $\mathscr{A}_n = \{A_i; 0 \le i \le n\}$ as defined at (3), such that $\bigcap_{i=0}^{n} \overline{A}_i \neq \emptyset$, and A_0 is an extreme element in $\mathscr{A}_n$. By taking A_0 as the center of symmetry, we obtain an augmented collection of stability regions $\mathscr{A}_n^*$ at (4). As showed in Proposition 5, there are $2n$ distinct elements in $\mathscr{A}_n^*$, and $(\overline{A} \cap \overline{A}_0) \neq \emptyset$ for all $A \in \mathscr{A}_n^*$. This implies that A_0 has at least $2n$ neighboring stability regions. In other words, the solution x_s^0 has at least $2n$ neighboring local-optimal solutions $\{T_{\zeta_i}^{-1}(x_s^0); 1 \le i \le n\} \cup \{T_{\zeta_i}(x_s^0); 1 \le i \le n\}$, where $T_{\zeta_i}(x_s^i) = x_s^0$ for all $0 \le i \le n$. Proposition 2 suggests that any local-optimal solution x_s and the solution x_s^0 must have exactly the same number of neighboring local-optimal solutions. The proof is completed. □

4 Local-Independence and Proof for Planar Case

By Theorem 4, there are at least $4 = 2n$ neighboring local-optimal solutions for the optimization problems in $\mathbb{R}^2$. In fact, for the planar problems, we can derive an improved bound on the number of neighboring local-optimal solutions, by investigating the local-independence of the collection $\Theta \doteq \{\zeta_i; 0 \le i \le n\}$, where $\zeta_0 = 0$ and ζ_i's are defined at (4) for $i \ge 1$. Here the collection Θ is called locally independent, if the vector difference of any two distinct vectors in Θ is unique. Note that the collection of neighboring solutions obtained by symmetry at (4) and Theorem 4 forms a subset of the collection of neighboring solutions obtained by vector differences (of the vectors in Θ) at Theorem 5. The main result is stated below.

Theorem 5 (Estimation Obtained by Local-Independence). *Consider a planar optimization problem* $\min_{x \in \mathbb{R}^2} f(x)$ *at (1), such that f is twice-differentiable and the dynamical system* $\dot{x} = -\nabla f(x)$ *at (2) satisfies the conditions (A1)–(A5). Then, the collection Θ is locally independent for $n = 2$, and any local-optimal solution x_s of (1) has at least six neighboring local-optimal solutions.*

To show the local-independence of Θ, we need an auxiliary proposition on the collinear stability regions.

Proposition 6 (*Separation of Collinear Elements*). *The set* $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) = \emptyset$, *for all* $\zeta \in \Theta \setminus \{\zeta_0\}$, $\alpha > 1$ *satisfying* $\alpha\zeta \in \mathcal{P}$ *(Fig. 3).*

Proof: See Appendix.

Proposition 6 shows that, for any three collinear stability regions, the middle region must separate the other two. It also implies that the vectors in $\Theta \setminus \{\zeta_0\}$ are linearly independent, which yields the local-independence of Θ. Now we are ready to give a complete proof of the theorem on the improved bound.

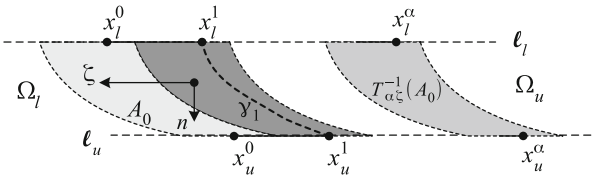


Fig. 3 An illustration of Proposition 6, where the point $x_l^0 = x_l$ and $x_u^0 = x_u$

Proof of Theorem 5: In light of Proposition 4, there are three distinct stability regions $\{A_{i_0}, A_{i_1}, A_{i_2}\}$, such that $(\bigcap_{q=0}^{2} \overline{A}_{i_q}) \neq \emptyset$. The collection Θ is defined by $\Theta = \{\zeta_q; 0 \leq q \leq 2\}$, where the vector $\zeta_q \in \mathcal{P}$ is uniquely determined by $T_{\zeta_q}(A_{i_q}) = A_0$. Now we consider the collection of vector differences

$$\Theta^* \doteq \{\zeta_{j_1} - \zeta_{j_2};\ \zeta_{j_1}, \zeta_{j_2} \in \Theta,\ j_1 \neq j_2\}.$$

To complete the proof, we will show that $T_\zeta^{-1}(\overline{A}_0) \cap \overline{A}_0 \neq \emptyset$ for all $\zeta \in \Theta^*$, and the collection Θ^* consists of six distinct nonzero vectors.

(i) We begin by showing $T_\zeta^{-1}(\overline{A}_0) \cap \overline{A}_0 \neq \emptyset$ for all $\zeta \in \Theta^*$. To fix the ideas, we consider a vector $\zeta \in \Theta^*$, and by the construction there must be distinct vectors $\zeta', \zeta'' \in \Theta$ with $\zeta' \neq \zeta''$, such that $\zeta = (\zeta' - \zeta'')$. From the choice of Θ, one has

$$\overline{A}_0 \cap T_{\zeta'}^{-1}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0) \neq \emptyset, \tag{5}$$

no matter whether the zero vector $0 \in \{\zeta', \zeta''\}$. Then, the set

$$\begin{aligned} T_\zeta^{-1}(\overline{A}_0) \cap \overline{A}_0 &= T_{\zeta'-\zeta''}^{-1}(\overline{A}_0) \cap \overline{A}_0 = T_{\zeta''}\big(T_{\zeta'}^{-1}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0)\big) \\ &\supseteq T_{\zeta''}\big(\overline{A}_0 \cap T_{\zeta'}^{-1}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0)\big) \neq \emptyset \end{aligned}$$

owing to (5) and the fact that the translation $T_{\zeta''}(\cdot)$ preserves the set cardinality. The claim that $T_\zeta^{-1}(\overline{A}_0) \cap \overline{A}_0 \neq \emptyset$ for all $\zeta \in \Theta^*$ has been justified.

(ii) It remains to prove that Θ^* contains six distinct vectors, or to say, the vector difference of any two distinct vectors in Θ is unique. First, one trivially has $-\zeta' \neq -\zeta''$ and $(\zeta' - \zeta'') \neq (\zeta'' - \zeta')$, if $\zeta' \neq \zeta'' \in \Theta$. Two more claims need to be clarified.

- First, we claim $\zeta' \neq -\zeta''$ if $\zeta' \neq \zeta''$. On the contrary, if $\zeta' = -\zeta''$, then

$$\begin{aligned} \overline{A}_0 \cap T_{\zeta'}^{-1}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0) &= \overline{A}_0 \cap T_{\zeta''}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0) \\ &= T_{\zeta''}\big(T_{\zeta''}^{-1}(\overline{A}_0) \cap \overline{A}_0 \cap T_{2\zeta''}^{-1}(\overline{A}_0)\big) \subseteq T_{\zeta''}\big(\overline{A}_0 \cap T_{2\zeta''}^{-1}(\overline{A}_0)\big) = \emptyset \end{aligned} \tag{6}$$

in light of Proposition 6. However, this contradicts the property (5). We have ruled out the case that $\zeta' = -\zeta''$. Consequently, $\zeta' \neq -\zeta''$ if $\zeta' \neq \zeta'' \in \Theta$.

- Next, we claim $(\zeta_1' - \zeta_1'') \neq (\zeta_2' - \zeta_2'')$, if $\zeta_1', \zeta_1'', \zeta_2', \zeta_2'' \in \Theta$ with $\zeta_1' \notin \{\zeta_1'', \zeta_2'\}$ and $\zeta_1'' \notin \{\zeta_1', \zeta_2''\}$. Clearly, there must be a nonzero vector in $\{\zeta_1', \zeta_1'', \zeta_2', \zeta_2''\}$, in view of $\zeta_1' \notin \{\zeta_1'', \zeta_2'\}$ and $\zeta_1'' \notin \{\zeta_1', \zeta_2''\}$. One can suppose without loss of generality that $\zeta' = \zeta_1' \neq 0$. Since Θ only contains two nonzero vectors at $n = 2$, we thus can denote by ζ'' the unique nonzero vector in $\Theta \setminus \{\zeta_1'\}$. Consequently, $\{\zeta_1'', \zeta_2'\} \subseteq \{0, \zeta''\}$.

  To prove the assertion by contradiction, we assume on the contrary that the difference $(\zeta_1' - \zeta_1'') = (\zeta_2' - \zeta_2'')$ for some such $\zeta_1', \zeta_1'', \zeta_2', \zeta_2'' \in \Theta$. Then $\zeta_1' = \zeta_1'' + (\zeta_2' - \zeta_2'') = (\zeta_1'' + \zeta_2') - \zeta_2''$. By recalling that $\{\zeta_1'', \zeta_2'\} \subseteq \{0, \zeta''\}$, one must have the vector $(\zeta_1'' + \zeta_2') = \alpha\zeta''$ for some $\alpha \in [0, 2]$. Now we examine the vector ζ_2''.
- If $\zeta_2'' \in \{0, \zeta''\}$, then $\zeta' = \zeta_1' = \alpha'\zeta''$ for some $\alpha' \in [-1, 2]$. We observe that $\alpha' \notin \{-1, 0, 1\}$; otherwise, a contradiction can be derived similar to (6). When $\alpha' \in (-1, 0)$, an application of Proposition 6 leads to

$$T_{\zeta'}^{-1}(\overline{A}_0) \cap T_{\zeta''}^{-1}(\overline{A}_0) = T_{\zeta'}^{-1}\left(\overline{A}_0 \cap T_{(1-\alpha')\zeta''}^{-1}(\overline{A}_0)\right) = \emptyset \tag{7}$$

  due to $(1 - \alpha') > 1$. When $\alpha' \in (0, 1)$, the vector $\zeta'' = \zeta'/\alpha'$, and by analogy we have

$$\overline{A}_0 \cap T_{\zeta''}^{-1}(\overline{A}_0) = \overline{A}_0 \cap T_{(1/\alpha')\zeta'}^{-1}(\overline{A}_0) = \emptyset \tag{8}$$

  in view of $1/\alpha' > 1$. When $\alpha' \in (1, 2)$, it also turns out that

$$\overline{A}_0 \cap T_{\zeta'}^{-1}(\overline{A}_0) = \overline{A}_0 \cap T_{\alpha'\zeta''}^{-1}(\overline{A}_0) = \emptyset. \tag{9}$$

  Obviously, (7)–(9) all contradict the property (5) satisfied by the vectors in Θ.

- Otherwise, if $\zeta_2'' = \zeta_1'$, it leads to that $\zeta' = \zeta_1' = \alpha/2 \cdot \zeta''$, where $\alpha/2 \in [0, 1]$. A contradiction can be derived, by an argument analogous to (8).

A contradiction always arises if $(\zeta_1' - \zeta_1'') = (\zeta_2' - \zeta_2'')$. Hence, the claim $(\zeta_1' - \zeta_1'') \neq (\zeta_2' - \zeta_2'')$ must be true, for all $\zeta_1', \zeta_1'', \zeta_2', \zeta_2'' \in \Theta$ with $\zeta_1' \notin \{\zeta_1'', \zeta_2'\}$ and $\zeta_1'' \notin \{\zeta_1', \zeta_2''\}$. This claim shows that the vector difference of any two distinct vectors in Θ is unique. That is, we have justified the local-independence of Θ.

To sum up, Θ is locally independent for $n = 2$. Moreover, the collection Θ contains three distinct vectors, and then there are $6 = 3 \cdot 2$ nonzero difference vectors, as collected in Θ^*. The verified claims suggest that the collection Θ^* includes six distinct vectors, and $T_\zeta^{-1}(\overline{A}_0) \cap \overline{A}_0 \neq \emptyset$ for all $\zeta \in \Theta^*$. In other words, the stability region A_0 has at least six neighboring stability regions, so does any stability region $A(x_s)$, in light of the spatial-periodicity by Proposition 2. As a consequence, the solution x_s must have at least six neighboring local-optimal solutions. The proof is completed. □

Example: To validate the derived bound, we consider a nonlinear optimization problem $\min_{x \in \mathbb{R}^2} f(x)$, say

$$f(x) = -3\cos(x_1) - \cos(x_2) - \cos(x_1 - x_2) - 0.04x_1 - 0.06x_2. \tag{10}$$

A simple computation shows that the gradient $-\nabla f(x) = (F_1(x), F_2(x))$ with

$$F_1(x) = -3\sin x_1 - \sin(x_1 - x_2) + 0.04, \quad F_2(x) = -\sin x_2 - \sin(x_2 - x_1) + 0.06.$$

One can easily check that the gradient $\nabla f(x)$ is spatially periodic, but not the objective function f. The point $x_s^0 = (0.0200, 0.0400)$ is a local-optimal solution to the problem (10). With reference to Fig. 4 left, the points x_s^i for $1 \le i \le 6$ are the neighboring local-optimal solutions of x_s^0, by seeing that there is a point x_e^{2i-1} shared by the stability boundaries $\partial A_p(x_s^i)$ and $\partial A_p(x_s^0)$ for $1 \le i \le 6$. They have been summarized in Fig. 4 right. Hence, the solution x_s^0 has exactly six neighboring local-optimal solutions.

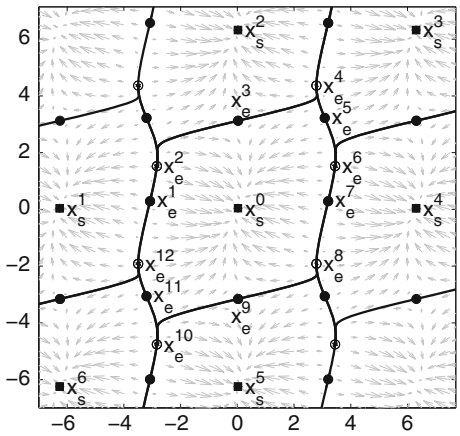
As proved by Theorem 5, there should be at least $6 = n(n+1)$ neighboring local-optimal solutions for x_s^0 at $n = 2$. This equals to the actual number of neighboring local-optimal solutions. Thus, the number $6 = n(n+1)$ provides the optimal lower bound on the number of neighboring local-optimal solutions, and this (lower) bound cannot be improved anymore for the planar optimization problems.

5 Engineering Interpretations

There are a number of engineering applications and interpretations related to the Sperner's lemma and the present study of local-optimal solutions.

First of all, Sperner's lemma has found interesting applications in software engineering [21, 22] and robust machines [23]. Clearly, software systems are of critical importance in the modern society, and their safety and quality have direct and immediate effects on our daily lives. In the manufacture and quality assurance process, an important element is the testing of software and hardware systems, to prevent the catastrophic consequences caused by software failure. In the industry, an affordable approach is to use the test suites generated from combinatorial designs, which involves identifying parameters that define the space of possible test scenarios, then selecting test scenarios to cover all the pairwise interactions between these parameters and their values. This process is called the *construction of efficient combinatorial covering suites*, and lower bounds on the size of covering suites [21, 22] have been derived by using the Sperner's lemma.

Furthermore, it should be noted that some concepts in power engineering [24], chemical engineering [25, 26], and molecular biology [27] resemble the local-optimal solutions in nonlinear optimization. Take the protein folding [27] as an example, the proteins are chains of amino acids and must self-assemble into well-defined conformations before fulfilling their biological functions, which can be achieved through a myriad of conformational changes (see Fig. 5). By convention, a conformation refers to a possible structure of the protein, and a conformational change is a transition between conformations. The resulting structure of folded protein is called the native state, determined by the sequence of amino acids, which can be interpreted as the state attaining the global minimum of the Gibbs free energy.



Local-Optimal Solutions Neighboring to x_s^0	
(1) x_s^1	(2) x_s^2
−6.2632	0.0200
0.0400	6.3232
(3) x_s^3	(4) x_s^4
6.3032	6.3032
6.3232	0.0400
(5) x_s^5	(6) x_s^6
0.0200	−6.2632
−6.2432	−6.2432

Fig. 4 *Left*: The stability region $A_p(x_s^0)$ is the area enclosed by the *bold curve* that the points x_e^k, $1 \le k \le 12$ lie on. Moreover, the local-optimal solution x_s^i is neighboring to the solution x_s^0, for $1 \le i \le 6$. *Right*: A summary of the neighboring local-optimal solutions, e.g., $x_s^1 = (-6.2632, 0.0400) \in \mathbb{R}^2$

The protein folding has a multi-state nature, and there can be many meta-stable states that can trap the folding and hinder the progress toward the native state. In this spontaneous optimization process, the meta-stable states play the role of local-optimal solutions, which may pertain to severe mammalian diseases.

Besides, the present work sheds light on the study of feasible components of the optimal power flow problem. In a typical power flow model [24], the power balance equations for the real and reactive power at node $k \in \{1, 2, \ldots, N\}$ are described by

$$\begin{aligned} H_{2k-1} &\doteq (P_k^G - P_k^L) - \textstyle\sum_{i=1}^N V_k V_i (G_{ki} \cos(\theta_k - \theta_i) + B_{ki} \sin(\theta_k - \theta_i)) = 0, \\ H_{2k} &\doteq (Q_k^G - Q_k^L) - \textstyle\sum_{i=1}^N V_k V_i (G_{ki} \sin(\theta_k - \theta_i) - B_{ki} \cos(\theta_k - \theta_i)) = 0. \end{aligned}$$

A solution $\theta = (\theta_1, \theta_2, \ldots, \theta_N) \in \mathbb{R}^N$ to the above equations (fixing P_k^G, P_k^L, Q_k^G, Q_k^L, V_k) must be a local-optimal solution of the minimization problem: $\min_{\theta \in \mathbb{R}^N} \frac{1}{2}\|H\|^2$, where the vector function $H \doteq (H_1, H_2, \ldots, H_{2N}) \in \mathbb{R}^{2n}$, and an associated gradient system is given by $\dot{\theta} = -\nabla_\theta H \cdot H$. The task of finding the local-optimal solutions of the minimization problem $\min_{\theta \in \mathbb{R}^N} \frac{1}{2}\|H\|^2$ thus is transformed to seek the stable equilibrium manifolds of the gradient system, where the stable manifolds of the stable equilibrium manifold are defined by [24]. Hence, the number of neighboring stable equilibrium manifolds can be estimated similar to Theorem 4, if the associated gradient system satisfies the recast conditions in terms of stable equilibrium manifolds, corresponding to (A1)–(A5).

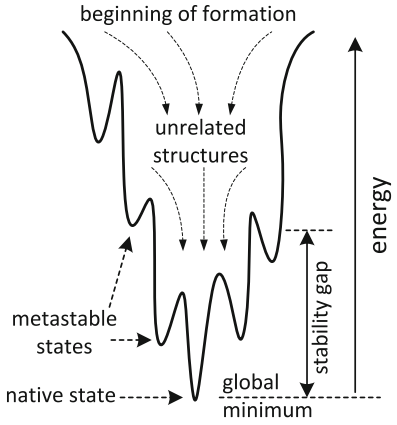


Fig. 5 The energy landscape of protein folding

6 Concluding Remarks

We have developed lower bounds for the number of neighboring local-optimal solutions for a class of nonlinear optimization problems. By the symmetry of the neighboring solutions, it is shown that there are at least $2n$ local-optimal solutions neighboring to a given one, where n is the dimensional of the state space. Moreover, for the planar problems, we can obtain an improved lower bound $6 = n(n+1) > 4 = 2n$ at $n = 2$. This is derived from the local-independence of the $(n+1)$ neighboring elements at Proposition 4. Nevertheless, it remains unclear whether $n(n+1)$ also provides an optimal lower bound on the number of neighboring local-optimal solutions for the optimization problems in $\mathbb{R}^n$ with $n \geq 3$.

Acknowledgements The presented work was partially supported by the CERT through the National Energy Technology Laboratory Cooperative Agreement No. DE-FC26-09NT43321, and partially supported by the National Science Foundation, USA, under Award #1225682.

Appendix

A Proof of Proposition 1

Proof: It should be apparent that $(A(x_s) \cap A(x'_s)) = \emptyset$, owing to $x_s \neq x'_s$. By the definition $\partial A_p(x_s) \doteq \overline{\partial A(x_s)}$, the task can be equivalently converted to show $(\overline{A(x_s)} \cap \overline{A(x'_s)}) = (\partial\overline{A(x_s)} \cap \partial\overline{A(x'_s)})$. The remaining analysis is given by examining the two possibilities of the intersection of closures.

- If $(\overline{A(x_s)} \cap \overline{A(x'_s)}) = \emptyset$, it is straightforward to see that $(\partial\overline{A(x_s)} \cap \partial\overline{A(x'_s)}) = \emptyset$. The conclusion is true.
- Otherwise, $(\overline{A(x_s)} \cap \overline{A(x'_s)}) \neq \emptyset$. Due to $(A(x_s) \cap A(x'_s)) = \emptyset$, one has $(\overline{A(x_s)} \cap \overline{A(x'_s)}) \subseteq (\partial A(x_s) \cap \overline{A(x'_s)}) \cup (\partial A(x'_s) \cap \overline{A(x_s)}) \subseteq (\partial A(x_s) \cup \partial A(x'_s))$. Recalling that $A(x_s)$ is open, we thus have the set $(\text{int}(\overline{A(x_s)}) \setminus \partial A(x_s)) \subseteq (\overline{A(x_s)} \setminus \partial A(x_s)) = A(x_s)$, so is at the point x'_s.

To prove by contradiction, we assume on the contrary that there is a point $y_\sharp \in (\overline{A(x_s)} \cap \overline{A(x'_s)})$, with either $y_\sharp \notin \partial\overline{A(x_s)}$ or $y_\sharp \notin \partial\overline{A(x'_s)}$. To fix the ideas, we suppose $y_\sharp \notin \partial\overline{A(x_s)}$. This implies the point $y_\sharp \in \text{int}(\overline{A(x_s)})$. Then, an open ball $B_\varepsilon(y_\sharp)$ exists in $\mathbb{R}^n$, with the center at $y_\sharp$ and the radius $\varepsilon > 0$, such that $B_\varepsilon(y_\sharp) \subseteq \text{int}(\overline{A(x_s)})$.

Since $y_\sharp \in (\overline{A(x_s)} \cap \overline{A(x'_s)}) \subseteq \overline{A(x'_s)}$, there must be a convergent sequence of points $y_k \in A(x'_s)$, $k \geq 1$ with $y_\sharp = \lim_{k\to\infty} y_k$. Recalling that $A(x'_s)$ is an open set, one thus can choose $\varepsilon_k > 0$ for $k \geq 1$, such that $B_{\varepsilon_k}(y_k) \subseteq A(x'_s)$ and $\lim_{k\to\infty} \varepsilon_k = 0$. The choice of y_k's ensures $B_{\varepsilon_k}(y_k) \subseteq B_\varepsilon(y_\sharp)$ for all k sufficiently large. From the construction of $B_\varepsilon(y_\sharp)$, it yields $B_{\varepsilon_k}(y_k) \subseteq \text{int}(\overline{A(x_s)}) \cap A(x'_s)$, for all large ks.

Let $\dim(\cdot)$ be the dimension [28] of a set in a Euclidean space. One can easily check that $\dim(B_{\varepsilon_k}(y_k)) = n$ for all $k \geq 1$. Meanwhile, $\dim(\partial A(x_s)) = (n-1)$ with $\dim(A(x_s)) = \dim(\overline{A(x_s)}) = n$, so are the sets for x'_s. This leads to $(B_{\varepsilon_k}(y_k) \setminus \partial A(x_s)) \neq \emptyset$, owing to $\dim(B_{\varepsilon_k}(y_k)) > \dim(\partial A(x_s))$. Then, for $k \geq 1$ sufficiently large,

$$\begin{aligned} A(x_s) \cap A(x'_s) &\supseteq (\text{int}(\overline{A(x_s)}) \setminus \partial A(x_s)) \cap A(x'_s) \\ &\supseteq (B_{\varepsilon_k}(y_k) \setminus \partial A(x_s)) \cap B_{\varepsilon_k}(y_k) = (B_{\varepsilon_k}(y_k) \setminus \partial A(x_s)) \neq \emptyset. \end{aligned}$$

As a consequence, $(A(x_s) \cap A(x'_s)) \neq \emptyset$. However, it violates the fact that the stability regions are disjoint for distinct stable equilibrium points. So the point $y_\sharp$ must belong to $\partial\overline{A(x_s)}$ and also belong to $\partial\overline{A(x'_s)}$ by analogy. This is valid for every point $y_\sharp \in (\overline{A(x_s)} \cap \overline{A(x'_s)})$. We thus can conclude that $(\overline{A(x_s)} \cap \overline{A(x'_s)}) = (\partial\overline{A(x_s)} \cap \partial\overline{A(x'_s)})$. □

B Proof of Proposition 4

Toward the proof of Proposition 4, first of all we need the existence of local-optimal solution (i.e., stable equilibrium point of (2)), which is stated below.

Proposition B.1 (*Existence of Local-Optimal Solution*). *Let $\mathscr{X} \doteq \{x_e; x_e \in \mathbb{R}^n\}$ be the set of all equilibrium points of (2). Then, there exists at least one stable equilibrium point of (2) in $\mathscr{X}$.*

Proof: From the condition (A1), the equilibrium points are all hyperbolic, which yields $\det(\nabla F) \neq 0$ for all equilibrium points of (2). Then, the equilibrium points are isolated. It follows that the set of all equilibrium points of (2) in $\mathbb{R}^n$ is countable. In the sequel, we can represent the set of equilibrium points as $\mathscr{X} = \{x_e^q; q \in \mathbb{N}\}$.

To show the conclusion, we assume on the contrary that no point in $\mathscr{X}$ is stable. Clearly, the stable manifold $W^s(x_e^q)$ is of dimension $\dim(W^s(x_e^q)) \leq (n-1) < n$, for all $q \in \mathbb{N}$. Besides, the stable manifold $W^s(x_e^q)$ is locally diffeomorphic to a Euclidean space. Thus, $W^s(x_e^q)$ is the union of countably many closed discs B_q^j with $\dim(B_q^j) = \dim(W^s(x_e^q)) \leq (n-1)$, $j \in \mathbb{N}$. In light of the condition (A3)–(A5), any point in $\mathbb{R}^n$ belongs to the stable manifold of an equilibrium point in $\mathscr{X}$, which shows the space $\mathbb{R}^n = \bigcup_{q\in\mathbb{N}} \bigcup_{j\in\mathbb{N}} B_q^j$. It follows from Sum Theorem [28, Theorem 1.5.3] that

$$\dim(\mathbb{R}^n) \leq \max\{\dim(B_q^j); q, j \in \mathbb{N}\} \leq (n-1).$$

However, this contradicts the fact $\dim(\mathbb{R}^n) = n$. Hence, the contrary proposition must be false. In a word, there must be one stable equilibrium point in $\mathscr{X} \subseteq \mathbb{R}^n$. □

Remark B.1. By (A4) and Proposition B.1, there is exactly one stable equilibrium point x_s^* in the subset $[0, 2\pi)^n$ of the state space $\mathbb{R}^n$, which includes exactly a single (spatial-) period for each x_i, $1 \leq i \leq n$. Proposition 2 also implies that $T_\zeta(x_s^*)$ is the only stable equilibrium point in the region $T_\zeta([0,2\pi)^n)$, for all $\zeta \in \mathcal{P}$. Moreover, let x_s be a stable equilibrium point in $\mathcal{S}$, and A_s be the stability region of x_s. Together with Proposition B.1, Proposition 2, and the assumption (A4), one has $\mathcal{S} = \{T_\zeta(x_s); \zeta \in \mathcal{P}\} = T_\zeta(\mathcal{S})$, for all $\zeta \in \mathcal{P}$, which is countable and consists of infinitely many points. Besides, the hypotheses (A3)–(A5) guarantee that the entire space $\mathbb{R}^n$ is the union of the closure of the stability regions of the points in $\mathcal{S}$. Or to say, $\mathbb{R}^n$ is the union of closures of the stability regions in $\mathscr{A} \doteq \{T_\zeta(A_s);\ \zeta \in \mathcal{P}\}$.

When applying the Sperner's lemma to prove Proposition 4, we need that the intersection of any compact set with the closures of stability regions in $\mathscr{A}$ is a union of finitely many closed sets, which yields that the union is a close set as well. To this end, an auxiliary proposition is summarized.

Proposition B.2 (*Finite Intersection with Compact Set*). *For any compact set $\Psi \subseteq \mathbb{R}^n$, there are only finitely many stability regions in $\mathscr{A}$ whose closures intersect Ψ.*

Proof: By the condition (A5), each stability region is bounded. Let ℓ be the diameter of a stability region $\overline{A}_s$, and $\|\cdot\|$ be the usual Euclidean norm of a vector. By the triangle inequality, given an arbitrary $\ell_\delta > 0$, if $\|\zeta\| > \ell_\delta + \ell$, then $\|x-y\| \geq \|y - T_\zeta^{-1}(y)\| - \|x - T_\zeta^{-1}(y)\| \geq \|\zeta\| - \ell > \ell_\delta$, for all points $x \in \overline{A}_s$ and $y \in T_\zeta(\overline{A}_s)$. Here, the inverse $T_\zeta^{-1} \doteq T_{-\zeta}$.

Let $\rho > 0$ be the diameter of Ψ, and $A_s \in \mathscr{A}$ be a stability region such that $(\Psi \cap \overline{A}_s) \neq \emptyset$. By setting $\ell_\delta = 2\rho$, we thus have $\|y-x\| > \ell_\delta = 2\rho > \rho$, for all $x \in (\overline{A}_s \cap \Psi), y \in T_\zeta(\overline{A}_s)$, $\zeta \in \mathcal{P}$ with $\|\zeta\| > \ell_\delta + \ell$. Then, $(T_\zeta(\overline{A}_s) \cap \Psi) = \emptyset$ for all $\zeta \in \mathcal{P}$ with $\|\zeta\| > \ell + 2\rho$, due to $(\Psi \cap \overline{A}_s) \neq \emptyset$.

In other words, $(T_\zeta(\overline{A}_s) \cap \Psi) \neq \emptyset$, only if $\zeta \in \mathcal{P}$ with $\|\zeta\| \leq \ell + 2\rho$. Observe that there are only finite number of vectors $\zeta \in \mathcal{P}$ satisfying $\|\zeta\| \leq \ell + 2\rho$. The proof is completed. □

Remark B.2. In view of (A1) and Lemma B.2, there are only finite number of equilibrium points in the region $[0, 2\pi)^n \subseteq \mathbb{R}^n$.

Proof of Proposition 4: Consider a closed set Ω^* in $\mathbb{R}^n$ which is the closure of a non-degenerated simplex, with $\{Q_k; 0 \le k \le n\}$ being the $(n-1)$-dimensional faces of Ω^*. Besides, $\{V_k; 0 \le k \le n\}$ are $(n+1)$ closed sets in $\mathbb{R}^n$ such that $\Omega^* = (\bigcup_{k=0}^n V_k)$, and $(V_k \cap Q_k) = \emptyset$ for all $0 \le k \le n$, with the vertex opposite to Q_k being contained in V_k. By Sperner's lemma and Theorem 2, the set $(\bigcap_{k=0}^n V_k) \neq \emptyset$.

To this end, we shall construct such closed sets V_k's, by using the closures of stability regions. First of all, we arbitrarily choose a simplex $\Omega^* \in \mathbb{R}^n$ with $\dim(\Omega^*) = n$, where $\{Q_k; 0 \le k \le n\}$ are the $(n-1)$-dimensional face set of Ω^*, and the point q_k is the vertex of Ω^* opposite to Q_k. Moreover, the simplex can be selected sufficiently large, such that Ω^* contains an open ball $B_\rho(x^*)$, where $x^* \in \Omega^*$ and the radius $\rho > 0$ is the diameter of a stability region. Then, every stability region $\overline{A}_i$ doesn't intersect all the faces of Ω^*, with either $(\overline{A}_i \cap Q_k) = \emptyset$ or $q_k \notin \overline{A}_i$ for each $0 \le k \le n$. In light of Proposition B.2, there are only finitely many A_i's $\in \mathscr{A}$ such that $(\overline{A}_i \cap \Omega^*) \neq \emptyset$.

The sets V_k's are obtained by induction as follows. Let $\mathscr{A}_0$ be the set of all $A_i \in \mathscr{A}$ such that $(\overline{A}_i \cap Q_0) = \emptyset$ with $(\overline{A}_i \cap \Omega^*) \neq \emptyset$, and $V_0 \doteq \bigcup \{(\overline{A}_i \cap \Omega^*); A_i \in \mathscr{A}_0\}$. Suppose that the collection $\mathscr{A}_j$ and the closed set V_j have been obtained, for all $0 \le j \le k$. We denote by $\mathscr{A}_{k+1}$ the collection of $A_i \in \mathscr{A}$ such that

$$A_i \notin \bigcup_{j=0}^k \mathscr{A}_j, \quad (\overline{A}_i \cap Q_{k+1}) = \emptyset \quad \text{and} \quad (\overline{A}_i \cap \Omega^*) \neq \emptyset.$$

The closed set $V_{k+1} \doteq \bigcup \{(\overline{A}_i \cap \Omega^*); A_i \in \mathscr{A}_{k+1}\}$. This process is terminated, once V_n is obtained. Clearly, $\bigcup_{k=0}^n V_k = \Omega^*$, and $(V_k \cap Q_k) = \emptyset$ for all $0 \le k \le n$.

Next, we verify that each V_k is not empty. Clearly, $V_0 \neq \emptyset$, and we suppose $V_j \neq \emptyset$ for all $j = 0, 1, \ldots, k$. It remains to show $V_{k+1} \neq \emptyset$. Apparently, the vertices $\{q_j;\ k < j \le n\} \subseteq \bigcap_{j=0}^k Q_j$, which implies $q_j \notin \overline{A}_i$, for all $A_i \in \bigcup_{j=0}^k \mathscr{A}_j$ and $k < j \le n$. Since $\Omega^* \subseteq \mathbb{R}^n = \bigcup_{A_i \in \mathscr{A}} \overline{A}_i$, there must exist one stability region $\overline{A}_{i^*}$ that contains the vertex q_{k+1}. It is straightforward to see that $A_{i^*} \notin \bigcup_{j=0}^k \mathscr{A}_j$. Moreover, by the choice of Ω^*, the set $\overline{A}_{i^*}$ doesn't intersect the face Q_{k+1} opposite to q_{k+1}. Hence, $V_{k+1} \supseteq (\overline{A}_{i^*} \cap \Omega^*) \neq \emptyset$. By this inductive argument, we conclude that $V_k \neq \emptyset$, for all $0 \le k \le n$.

Then, it follows from Theorem 2 that $(\bigcap_{k=0}^n V_k) \neq \emptyset$. Let y_* be a point in this nonempty intersection. Since $y_* \in V_k$, there must be a set $A_{i_k} \in \mathscr{A}_k$ such that $y_* \in (\overline{A}_{i_k} \cap \Omega^*) \subseteq V_k$. We thus obtain a finite subset $\{A_{i_k}; 0 \le k \le n\} \subseteq \mathscr{A}$, which satisfies $(\bigcap_{k=0}^n \overline{A}_{i_k}) \supseteq \bigcap_{k=0}^n (\overline{A}_{i_k} \cap \Omega^*) \supseteq \{y_*\} \neq \emptyset$. The first assertion is proved. □

C Proof of Proposition 5

Proof: Above all, the vector $\zeta_i \neq 0$, due to $x_s^0 \neq x_s^i$, $A_0 \neq A_i$ and $T_{\zeta_i}(A_i) = A_0$ for $1 \leq i \leq n$. Meanwhile, $\zeta_i \neq \zeta_j$ for all $i \neq j$, in light of $A_i \neq A_j$ and $T_{\zeta_i}(A_i) = A_0 = T_{\zeta_j}(A_j)$. In addition, $T_{\zeta_i}(x_s^0) = x_s^i$ for all $1 \leq i \leq n$, where the stability region $A_i = A(x_s^i)$.

To show that the elements in $\mathscr{A}_n^*$ at (4) are pairwise different, it suffices to only clarify that $T_{\zeta_i}(A_0) \neq A_j$ for all $1 \leq i, j \leq n$, and $T_{\zeta_i}(A_0) \neq T_{\zeta_j}(A_0)$ for all $i \neq j$.

- To justify $T_{\zeta_i}(A_0) \neq A_j$, we assume on the contrary that $T_{\zeta_i}(A_0) = A_j$ for some $1 \leq i, j \leq n$. In other words, the corresponding stable equilibrium points satisfy $T_{\zeta_i}(x_s^0) = x_s^j$. Observe that $T_{\zeta_i}(x_s^i) = x_s^0$ or $T_{\zeta_i}^{-1}(x_s^0) = x_s^i$, by the definition of ζ_i. Then, $x_s^i/2 + x_s^j/2 = 1/2 \cdot T_{\zeta_i}^{-1}(x_s^0) + 1/2 \cdot T_{\zeta_i}(x_s^0) = x_s^0$. That is, x_s^0 can be represented by a convex combination of two elements in $\{x_s^1, \ldots, x_s^n\}$. However, this contradicts the choice of x_s^0, which is an extreme point in the point set. A contradiction arises. We thus can conclude that $T_{\zeta_i}(A_0) \neq A_j$ for all $1 \leq i, j \leq n$.
- Next we show $T_{\zeta_i}(A_0) \neq T_{\zeta_j}(A_0)$, for all $i \neq j$, $1 \leq i, j \leq n$. To prove by contradiction, we assume on the contrary that $T_{\zeta_i}(A_0) = T_{\zeta_j}(A_0)$ for some $i \neq j$. Trivially, it yields $A_0 = T_{\zeta_j - \zeta_i}(A_0)$, and then $\zeta_i = \zeta_j$, which violates the choice that $\zeta_i \neq \zeta_j$ for all $i \neq j$. A contradiction arises. Thus, the contrary proposition must be false. We have completed the proof of the assertion that $T_{\zeta_i}(A_0) \neq T_{\zeta_j}(A_0)$ for all $i \neq j$ with $1 \leq i, j \leq n$.

  In a word, the elements in $\mathscr{A}_n^*$ are pairwise different. Moreover, $\big(T_{\zeta_i}(\overline{A}_0) \cap \overline{A}_0\big) = \big(T_{\zeta_i}(\overline{A}_0) \cap T_{\zeta_i}(\overline{A}_i)\big) = T_{\zeta_i}(\overline{A}_0 \cap \overline{A}_i) \neq \emptyset$, due to $(\overline{A}_0 \cap \overline{A}_i) \supseteq \bigcap_{q=0}^{n} \overline{A}_{i_q} \neq \emptyset$. Therefore, the set $(\overline{A} \cap \overline{A}_0) \neq \emptyset$, for all $A \in \mathscr{A}_n^*$. The proposition is proved. □

D Proof of Proposition 6

Proposition 6 will be proved by contradiction, which relies on the following result on the plane geometry and simple curves.

Proposition D.1 (Intersection of Simple Curves). *Let v be a unit vector in the plane, and the line ℓ be defined by $\ell \doteq \{\lambda v; \lambda \in \mathbb{R}\}$, with S being an open segment $\subseteq \ell$. The set H refers to a connected component of the set $(\mathbb{R}^2 \setminus \ell)$, which is a half-plane.*

Let $\gamma \subseteq H$ be a simple curve satisfying that the set $\Gamma \doteq (\overline{\gamma} \cup S)$ forms a Jordan curve, and the length $m_1(\Gamma) < \infty$. Then, the set

$$\gamma \cap T_{\alpha v}(\gamma) \neq \emptyset, \tag{11}$$

if the set $(S \cap T_{\alpha v}(S)) \neq \emptyset$ for some $\alpha \in \mathbb{R}$.

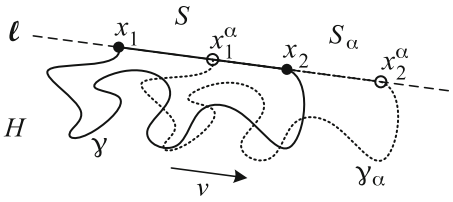


Fig. 6 In Proposition D.1, the *curve* γ_α must intersect γ, if $(S \cap S_\alpha) \neq \emptyset$

Proof: Let x_1, x_2 be the endpoints of the segment $\overline{S}$, and x_1^α, x_2^α be that of the segment $\overline{S}_\alpha$, where $S_\alpha = T_{\alpha v}(S)$ and $\gamma_\alpha \doteq T_{\alpha v}(\gamma)$ (Fig. 6). Without loss of generality we fix the point $x_2 = T_{\alpha_s v}(x_1)$, where $\alpha_s = \|S\| > 0$ is the length of S. That is, the point x_2 lies downstream of x_1 on ℓ.

Observe that the set $(H \cap T_{\alpha v}(S)) \subseteq (H \cap \ell) = \emptyset$, for all $\alpha \in \mathbb{R}$. Clearly, $\Gamma_\alpha \doteq T_{\alpha v}(\overline{\gamma} \cup S) = T_{\alpha v}(\Gamma)$ is a Jordan curve. One can also easily check that, the set $I_{\Gamma_\alpha} \subseteq H$ for all $\alpha \in \mathbb{R}$, in view of $I_{\Gamma_\alpha} = T_{\alpha v}(I_\Gamma) \subseteq H$. Here I_Γ refers to the bounded connected component of the set $(\mathbb{R}^2 \setminus \Gamma)$, or to say I_Γ is the interior enclosed by the Jordan curve Γ. The conclusion is obviously true at $\alpha = 0$. It remains to examine the case for $\alpha \neq 0$. To fix the ideas, we consider the case that $\alpha > 0$ in the sequel.

Since $x_2 = T_{\alpha_s c}(x_1)$ with $\alpha_s = \|S\| > 0$, one must have that, the condition $(S \cap S_\alpha) \neq \emptyset$ implies the point $x_2 \in S_\alpha$ and $x_1^\alpha \in S$, if $\alpha > 0$. To show $(\gamma \cap \gamma_\alpha) \neq \emptyset$, we need to claim that $(\gamma \cap I_{\Gamma_\alpha}) \neq \emptyset$, and $(\gamma \setminus \bar{I}_{\Gamma_\alpha}) \neq \emptyset$.

(i) We start by justifying the claim that $(\gamma \cap I_{\Gamma_\alpha}) \neq \emptyset$.

First of all, we claim the set $(B_\varepsilon(x_2) \cap H) \subseteq I_{\Gamma_\alpha}$, for some $\varepsilon > 0$. Observe that the curve $\gamma_\alpha \subseteq H$ and the point $x_2 \in \ell \subseteq (\mathbb{R}^2 \setminus H)$. Then, the point $x_2 \notin \overline{\gamma}_\alpha$, and thereby $(B_\varepsilon(x_2) \cap \gamma_\alpha) = \emptyset$ for all $\varepsilon > 0$ sufficiently small. Moreover, the set $(B_\varepsilon(x_2) \cap H) \cap \Gamma_\alpha = \emptyset$ for all $\varepsilon > 0$ sufficiently small, in view of $(B_\varepsilon(x_2) \cap H) \cap \ell = \emptyset$. By the simple-connectivity of $(B_\varepsilon(x_2) \cap H)$, the set

$$(B_\varepsilon(x_2) \cap H) \subseteq I_{\Gamma_\alpha} \text{ or } (B_\varepsilon(x_2) \cap H) \subseteq (\mathbb{R}^2 \setminus \bar{I}_{\Gamma_\alpha}). \tag{12}$$

It should be apparent that the ball $B_\varepsilon(x_2)$ must intersect I_{Γ_α}, owing to the point $x_2 \in S_\alpha \subseteq \Gamma_\alpha$ and Γ_α is the boundary of I_{Γ_α}. Then,

$$(B_\varepsilon(x_2) \cap H) \cap I_{\Gamma_\alpha} \supseteq (B_\varepsilon(x_2) \cap I_{\Gamma_\alpha}) \cap I_{\Gamma_\alpha} = (B_\varepsilon(x_2) \cap I_{\Gamma_\alpha}) \neq \emptyset$$

in light of $\Gamma_\alpha \subseteq H$ and $I_{\Gamma_\alpha} \subseteq H$ by Conway [29, Corollary 13.1.11]. We can conclude that the set $(B_\varepsilon(x_2) \cap H) \subseteq I_{\Gamma_\alpha}$ for all $\varepsilon > 0$ sufficiently small, in view of (12). The auxiliary claim is proved.

One can easily check that $(\gamma \cap B_\varepsilon(x_2)) \neq \emptyset$ for all $\varepsilon > 0$, due to $x_2 \in \overline{\gamma}$. It turns out that the set

$$(\gamma \cap I_{\Gamma_\alpha}) \supseteq \gamma \cap (B_\varepsilon(x_2) \cap H) = (\gamma \cap B_\varepsilon(x_2)) \neq \emptyset$$

for all $\varepsilon > 0$ sufficiently small, owing to $\gamma \subseteq H$ and the verified auxiliary claim $(B_\varepsilon(x_2) \cap H) \subseteq I_{\Gamma_\alpha}$. We complete the proof for the claim (i).

(ii) Next we prove that $(\gamma \setminus \bar{I}_{\Gamma_\alpha}) \neq \emptyset$.

We begin by showing the point $x_1 \notin \bar{I}_{\Gamma_\alpha}$. Clearly, the point $x_1 \notin S_\alpha$. By the boundedness of the set $\bar{I}_{\Gamma_\alpha}$, one can easily check that $T_{\lambda v}(x_1)$ doesn't belong to $\bar{I}_{\Gamma_\alpha}$, for all $\lambda \in \mathbb{R}$ with $|\lambda|$ being sufficiently large. Let x_1^* be the point $T_{\lambda v}(x_1)$, for some $\lambda < 0$ with $|\lambda|$ being sufficiently large. It should be apparent that the segment

$$S^* \doteq \{T_{\lambda v}(x_1);\, \lambda < \lambda' < 0\}$$

doesn't intersect S and S_α, in view of $\alpha > 0$. By recalling that $(S^* \cap \gamma_\alpha) \subseteq (S^* \cap H) = \emptyset$, we thus obtain the set $(S^* \cap \Gamma_\alpha) = \emptyset$. That is, either the segment $S^* \subseteq I_{\Gamma_\alpha}$, or the set $(S^* \cap I_{\Gamma_\alpha}) = \emptyset$. Hence, the endpoints x_1^* and x_1 of the segment S^* must belong to a same connected component of the set $(\mathbb{R}^2 \setminus \Gamma_\alpha)$. It turns out that the point $x_1 \notin \bar{I}_{\Gamma_\alpha}$, owing to $x_1^* \notin \bar{I}_{\Gamma_\alpha}$. The claim is proved.

Since the set $\bar{I}_{\Gamma_\alpha}$ is closed, there must be an $\varepsilon > 0$ sufficiently small, such that $(B_\varepsilon(x_1) \cap \bar{I}_{\Gamma_\alpha}) = \emptyset$. Clearly, the set $(\gamma \cap B_\varepsilon(x_1)) \neq \emptyset$ for all $\varepsilon > 0$. Then, the set

$$(\gamma \setminus \bar{I}_{\Gamma_\alpha}) \supseteq (\gamma \cap B_\varepsilon(x_1)) \setminus \bar{I}_{\Gamma_\alpha} = (\gamma \cap B_\varepsilon(x_1)) \neq \emptyset$$

for all $\varepsilon > 0$ sufficiently small. That is, $(\gamma \setminus \bar{I}_{\Gamma_\alpha}) \neq \emptyset$. Claim (ii) is justified.

At last, the Jordan curve theorem yields $(\gamma \cap \Gamma_\alpha) \neq \emptyset$, in view of the verified claims $(\gamma \cap I_{\Gamma_\alpha}) \neq \emptyset$ and $(\gamma \setminus \bar{I}_{\Gamma_\alpha}) \neq \emptyset$. Evidently, the set $(\gamma \cap \bar{S}_\alpha) \subseteq (\gamma \cap \ell) = \emptyset$. Then, the set $(\gamma \cap T_{\alpha v}(\gamma)) = (\gamma \cap \gamma_\alpha) = (\gamma \cap \Gamma_\alpha) \neq \emptyset$, if $(S \cap T_{\alpha v}(S)) \neq \emptyset$ for some $\alpha > 0$.

Similarly, it can be shown that $(\gamma \cap T_{\alpha v}(\gamma)) \neq \emptyset$, if $(S \cap T_{\alpha v}(S)) \neq \emptyset$ with $\alpha < 0$. The proof of the proposition is completed. □

Proof of Proposition 6: To prove the conclusion, we will derive a contradiction for the contrary opposition by applying Proposition D.1. To this end, we construct the desirable lines and segments as follows (see Fig. 3).

(i) Let $n \in \mathbb{R}^2$ be a unit vector perpendicular to ζ. We consider an arbitrary point $x_* \in A_0$ and define a signed distance function by $d_n(y) \doteq \langle y - x_*, n \rangle$ for $y \in \mathbb{R}^2$. Recalling that the closure $\bar{A}_0$ is compact, we thus have, there are $x_l, x_u \in \bar{A}_0$ such that

$$d_n(x_l) = \inf\{d_n(y);\, y \in \bar{A}_0\}; \quad d_n(x_u) = \sup\{d_n(y);\, y \in \bar{A}_0\}.$$

The straight lines are defined by

$$\ell_l \doteq \{x_l + \lambda\zeta;\, \lambda \in \mathbb{R}\} = \{T_{\lambda\zeta}^{-1}(x_l);\, \lambda \in \mathbb{R}\}, \quad \ell_u \doteq \{x_u + \lambda\zeta;\, \lambda \in \mathbb{R}\}.$$

It is apparent that the region Ω confined between ℓ_l and ℓ_u, is a simply connected set, with $d_n(x_l) < d_n(y) < d_n(x_u)$ for all $y \in \Omega$.

(ii) By the assumption (A5), there is a simple curve $\gamma_0 \subseteq A_0$ satisfying that the curve $\bar{\gamma}_0$ connects the point x_l to x_u, with the length $m_1(\gamma_0) < \infty$. For convenience we use γ_α to refer to the curve $T_{\alpha\zeta}^{-1}(\gamma_0)$, where the endpoint $x_l^\alpha \doteq T_{\alpha\zeta}^{-1}(x_l)$ and $x_u^\alpha \doteq T_{\alpha\zeta}^{-1}(x_u)$. In addition, we fix the point $x_l^0 = x_l$ and $x_u^0 = x_u$.

On the set $(\Omega \setminus \gamma_1)$, there are only two connected components, say Ω_l and Ω_u. Apparently, the components $\Omega_l, \Omega_u \subseteq \mathbb{R}^2$ are simply connected, though they are not bounded. Without loss of generality we suppose that the ray $r_l \doteq \{T_{\lambda\zeta}^{-1}(x_l); \lambda < 1\} \subseteq \overline{\Omega}_l$, and $r_u \doteq \{T_{\lambda\zeta}^{-1}(x_l); \lambda > 1\} \subseteq \overline{\Omega}_u$. It can be easily checked that $(\overline{\Omega}_l \cap \overline{\Omega}_u) = \overline{\gamma}_1$.

(iii) We proceed by claiming that $A_0 \subseteq \Omega_l$, and $T_{\alpha\zeta}^{-1}(A_0) \subseteq \Omega_u$ for $\alpha > 1$.

Observe that the set $(A_0 \cap \gamma_1) \subseteq (A_0 \cap T_\zeta^{-1}(A_0)) = \emptyset$ and $A_0 \cap (\ell_l \cup \ell_u) = \emptyset$. Then, $(A_0 \cap \partial\Omega_l) = \emptyset$, where $\partial\Omega_l \subseteq \gamma_1 \cup (\ell_l \cup \ell_u)$. In other words,

$$A_0 \subseteq \Omega_l \quad \text{or} \quad (A_0 \cap \Omega_l) = \emptyset. \tag{13}$$

From the construction, we easily observe that the point $x_l^0 \in \Omega_l$, and $x_l^\alpha \in \Omega_u$ for all $\alpha > 1$. In view of $(B_{\alpha_\zeta}(x_l^0) \cap r_u) = \emptyset$ for $\alpha_\zeta \doteq \|\zeta\|$ and the point $x_l^0 \notin \ell_u$ with $x_l^0 \notin \overline{\gamma}_1$, one has $(B_\varepsilon(x_l^0) \cap \overline{\Omega}_u) = \emptyset$, for all $\varepsilon > 0$ sufficiently small. This further yields

$$(B_\varepsilon(x_l^0) \cap \Omega_l) = B_\varepsilon(x_l^0) \cap (\Omega_l \cup \gamma_1 \cup \Omega_u) = (B_\varepsilon(x_l^0) \cap \Omega) \neq \emptyset.$$

On the other hand, $(B_\varepsilon(x_l^0) \cap A_0) \neq \emptyset$ owing to $x_l^0 \in \overline{A}_0$. Together with $A_0 \subseteq \Omega$, it implies that

$$\begin{aligned}(A_0 \cap \Omega_l) &\supseteq (A_0 \cap B_\varepsilon(x_l^0)) \cap (B_\varepsilon(x_l^0) \cap \Omega_l)\\ &= (A_0 \cap B_\varepsilon(x_l^0)) \cap (B_\varepsilon(x_l^0) \cap \Omega)\\ &= (A_0 \cap \Omega) \cap B_\varepsilon(x_l^0) = (A_0 \cap B_\varepsilon(x_l^0)) \neq \emptyset.\end{aligned}$$

for all $\varepsilon > 0$ sufficiently small. Finally, the set $A_0 \subseteq \Omega_l$, in view of (13). An analogous argument shows that $T_{\alpha\zeta}^{-1}(A_0) \subseteq \Omega_u$ for $\alpha > 1$.

(iv) To prove the conclusion, we assume on the contrary that $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) \neq \emptyset$.

By the verified claims, one has $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) \subseteq (\overline{\Omega}_l \cap \overline{\Omega}_u) = \overline{\gamma}_1$. Proposition 1 implies that $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) = (\partial\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\partial\overline{A}_0))$, for all $\alpha \neq 0$ with $\alpha\zeta \in \mathcal{P}$, due to $(A_0 \cap T_{\alpha\zeta}^{-1}(A_0)) = \emptyset$. As a consequence, $(\gamma_1 \cap \overline{A}_0) = (\gamma_1 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) = \emptyset$ for $\alpha > 1$, in view of $\gamma_1 \subseteq T_\zeta^{-1}(A_0)$ and $(\gamma_1 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) \subseteq (T_\zeta^{-1}(A_0) \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) = \emptyset$. It turns out that the set $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) \subseteq (\overline{\gamma}_1 \setminus \gamma_1) = \{x_l^1, x_u^1\}$.

To fix the ideas, we suppose the point $x_l^1 \in (\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0))$. Then, the points

$$x_l^1, x_l^0, x_l^{-\alpha} \in \overline{A}_0, \ \text{ and } \ x_l^2, x_l^1, x_l^{1-\alpha} \in T_\zeta^{-1}(\overline{A}_0).$$

In light of the condition (A5), there is a simple curve $\gamma \subseteq A_0$ such that $\overline{\gamma}$ joins the points $x_l^{-\alpha}$ and x_l^1. Clearly, the curve $T_\zeta^{-1}(\gamma) \subseteq T_\zeta^{-1}(A_0)$, whose closure connects the point x_l^2 to $x_l^{1-\alpha}$. Let S be the segment joining the points x_l^1 and $x_l^{-\alpha}$, with the line $\ell = \ell_l$. An application of Proposition D.1 yields that $(\gamma \cap T_\zeta^{-1}(\gamma)) \neq \emptyset$, in view of $(S \cap T_\zeta^{-1}(S)) \supseteq \{x_l^0\} \neq \emptyset$. In other words, the set $(A_0 \cap T_\zeta^{-1}(A_0)) \supseteq (\gamma \cap T_\zeta^{-1}(\gamma)) \neq \emptyset$,

which, however, violates the fact $(A_0 \cap T_\zeta^{-1}(A_0)) = \emptyset$. The contrary opposition must be false. We thus can conclude that $(\overline{A}_0 \cap T_{\alpha\zeta}^{-1}(\overline{A}_0)) = \emptyset$ for $\alpha > 1$. The proof is completed. □

References

1. Sperner, E.: Neuer Beweis für die Invarianz der Dimensionszahl und des Gebietes. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg **6**(1), 265–272 (1928)
2. Bagemihl, F.: An extension of Sperner's lemma, with applications to closed-set coverings and fixed points. Fundam. Math. **40**(1), 3–12 (1953)
3. Kuhn, H.W.: Some combinatorial lemmas in topology. IBM J. Res. Dev. **4**(5), 518–524 (1960)
4. Cohen, D.I.: On the Sperner lemma. J. Comb. Theory **2**(4), 585–587 (1967)
5. de Longueville, M.: A Course in Topological Combinatorics, pp. 5–6. Springer, New York (2012)
6. Monsky, P.: On dividing a square into triangles. Am. Math. Mon. **77**(2), 161–164 (1970)
7. Naber, G.L.: Topological Methods in Euclidean Spaces. Courier Dover Publications, New York (2000)
8. Kuratowski, K.: Topology, vol. I. PWN-Polish Scientific Publishers/Academic, Warsaw/New York (1966)
9. Chiang, H.D., Chu, C.C.: A systematic search method for obtaining multiple local optimal solutions of nonlinear programming problems. IEEE Trans. Circuits Syst. **43**(2), 99–106 (1996)
10. Lee, J., Chiang, H.D.: A dynamical trajectory-based methodology for systematically computing multiple optimal solutions of nonlinear programming problems. IEEE Trans. Autom. Control **49**(6), 888–899 (2004)
11. Chiang, H.D., Lee, J.: Trust-tech paradigm for computing high-quality optimal solutions: method and theory. In: Lee, K.Y., El-Sharkawi, M.A. (eds.) Modern Heuristic Optimization Techniques: Theory and Applications to Power Systems, pp. 209–233. Wiley, Hoboken (2007)
12. Wang, T., Chiang, H.D., Huang L.X.: On the number of adjacent elements in normal and lattice tilings, Discrete & Computational Geometry, under review.
13. Hirsch, M.W.: Differential Topology. Springer, New York (1976)
14. Chiang, H.D., Fekih-Ahmed, L.: Quasi-stability regions of nonlinear dynamical systems: theory. IEEE Trans. Circuits Syst. **43**(8), 627–635 (1996)
15. Aiello, G., Alfonzetti, S., Borzì, G., Saleron, N.: Computing spatially-periodic electric fields by charge iteration. IEEE Trans. Magnetics **34**(5), 2501–2504 (1998)
16. Fardad, M., Jovanović, M.R., Bamieh, B.: Frequency analysis and norms of distributed spatially periodic systems. IEEE Trans. Autom. Control **53**(10), 2266–2279 (2008)
17. Rokhlenko, A., Lebowitz, J.L.: Modeling electron flow produced by a three-dimensional spatially periodic field emitter. J. Appl. Phys. **108**, 123301, 1–6 (2010)
18. Ordonez, C.A., Pacheco, J.L., Weathers, D.L.: Spatially periodic electromagnetic force field for plasma confinement and control. Open Plasma Phys. J. **5**, 1–10 (2012)
19. Kolokathis, P.D., Theodorou, D.N.: On solving the master equation in spatially periodic systems. J. Chem. Phys. **137**, 034112, 1–21 (2012)
20. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1997)
21. Hartman, A.: Software and hardware testing using combinatorial covering suites. In: Graph Theory, Combinatorics and Algorithms, pp. 237–266. Springer, New York (2005)
22. Greene, C.: Sperner families and partitions of a partially ordered set. In: Hall, M., Jr., van Lint, J. (eds.) Combinatorics, pp. 277–290. Dordrecht, Holland (1975)

23. Crescenzi, P., Silvestri, R.: Sperner's lemma and robust machines. In: Proceedings of the Eighth Annual, Structure in Complexity Theory Conference, pp. 194–199. IEEE, New York (1993)
24. Chiang, H.D., Wang, B., Jiang, Q.Y.: Applications of TRUST-TECH methodology in optimal power flow of power systems. In: Optimization in the Energy Industry, pp. 297–318. Springer, Berlin (2009)
25. Atkins, P., De Paula, J.: Physical Chemistry, 8th edn. Oxford University Press, New York (2006)
26. Tessier, S.R., Brennecke, J.F., Stadtherr, M.A.: Reliable phase stability analysis for excess Gibbs energy models. Chem. Eng. Sci. **55**(10), 1785–1796 (2000)
27. Onuchic, J.N., Luthey-Schulten, Z., Wolynes, P.G.: Theory of protein folding: the energy landscape perspective. Ann. Rev. Phys. Chem. **48**(1), 545–600 (1997)
28. Engelking, R.: Dimension Theory. PWN-Polish Scientific Publishers, Warszawa (1978)
29. Conway, J.B.: Functions of One Complex Variable II. Springer, New York (1995)

General Traffic Equilibrium Problem with Uncertainty and Random Variational Inequalities

Patrizia Daniele, Sofia Giuffrè, and Antonino Maugeri

1 Introduction

In the last decades some papers (see [3–5, 8]) have been devoted to the study of random variational inequalities or general random equilibrium problems. Particularly Gwinner and Raciti in [5] present a class of linear random variational inequalities on random sets and give measurability, existence, and uniqueness results in a Hilbert space setting. Moreover in a special case they provide an approximation procedure. In paper [6] the authors apply the theory of random variational inequalities to study a class of random equilibrium problems on networks in the linear case, whereas the application to nonlinear random traffic equilibrium problem is treated in [7]. In paper [9], which is devoted to the study of a general infinite dimensional complementarity problem, the authors consider a random traffic equilibrium problem in the framework of generalized complementarity problems.

The aim of this paper is to consider a general random traffic equilibrium problem, namely a traffic problem where the data are affected by a certain degree of uncertainty, to give a random generalized Wardrop equilibrium condition and to show that the equilibrium conditions are equivalent to a random variational inequality.

The need to develop a random model of the traffic network arises because the path flows as well as the travel demand are often variable over time in a non-regular and predictable manner. Such an uncertainty can be caused not only by several factors such as the particular hour of the day, the particular day of the week, the

P. Daniele (✉) • A. Maugeri
Department of Mathematics and Computer Science, University of Catania,
Viale A. Doria, 6, 95125 Catania, Italy
e-mail: daniele@dmi.unict.it; maugeri@dmi.unict.it

S. Giuffrè
D.I.I.E.S., "Mediterranea" University of Reggio Calabria, 89060 Reggio Calabria, Italy
e-mail: sofia.giuffre@unirc.it

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_4,

particular week of the year but also by a sudden accident or a maintenance work. Moreover, since the demand itself is dynamic and can change randomly, we propose a framework which is able to handle random constraints.

We choose for our model a Hilbert space setting, which allows us to obtain, under general assumptions, existence and uniqueness results. The paper is organized as follows. In Sect. 2 the detailed random traffic equilibrium model is presented, a random generalized Wardrop equilibrium condition is stated, and a variational characterization of the equilibrium is given. In Sect. 3 we provide the proof of the main result and in Sect. 4 some existence results are discussed. Finally Sect. 5 summarizes our results and future work.

2 The Model and Main Results

For the reader's utility we introduce in detail the model of random traffic equilibrium problem (see [2] for the deterministic case). A traffic network consists of a triple (N,A,W), where $N=(N_1,N_2,\dots,N_p)$ is the set of nodes, $A=(A_1,\dots,A_n)$ represents the set of the directed arcs connecting couples of nodes, and $W=\{w_1,\dots,w_l\}\subset N\times N$ is the set of the origin–destination (O/D) pairs. The flow on the arc A_i is denoted by f_i and the uncertainty which affects the knowledge of f_i is given by the dependence of f_i on ω, namely $f_i=f_i(\omega)$, where $\omega\in\Omega$ and $(\Omega,\mathcal{A},P)$ is a probability space. We will set $f(\omega)=(f_1(\omega),\dots,f_n(\omega))$. We call a set of consecutive arcs a path and assume that each O–D pair w_j is connected by $r_j\geq 1$ paths, whose set is denoted by $\mathcal{R}_j$, $j=1,\dots,l$. All the paths in the network are grouped into a vector $(R_1,\dots,R_m)$. We can describe the arc structure of the path by using the arc-path incidence matrix $\Delta=\{\delta_{ir}\}$, $i=1,\dots,n$, $r=1,\dots,m$, whose entries take the value 1 if $A_i\in R_r$ and 0 if $A_i\notin R_r$. To each path R_r there corresponds a flow $F_r(\omega)$, $\omega\in\Omega$, and the path flows are grouped into a vector $(F_1(\omega),\dots,F_m(\omega))$, which is called the path flow vector. The flow f_i on the arc A_i is equal to the sum of the flows on the paths which contain A_i so that $f(\omega)=\Delta F(\omega)$, $\omega\in\Omega$. Let us now introduce the unit cost of going through A_i as a function $c_i(f(\omega))\geq 0$ of the flows on the network, so that $c(f(\omega))=(c_1(f(\omega)),\dots,c_n(f(\omega)))$ denotes the arc cost on the network. Analogously $C(F(\omega))=(C_1(F(\omega)),\dots,C_m(F(\omega)))$ will denote the cost on the paths. Usually $C_r(F(\omega))$ is given by the sum of the costs on the arcs building the path: $C_r(F(\omega))=\sum_{i=1}^{n}\delta_{ir}c_i(f(\omega))$, $\omega\in\Omega$ or $C(F(\omega))=\Delta^T c(\Delta F(\omega))$. Instead of assuming that the paths have an infinite capacity, we suppose that there exist two random capacity vectors $\lambda(\omega)$, $\mu(\omega)$, $\lambda(\omega)<\mu(\omega)$, such that

$$0\leq\lambda(\omega)\leq F(\omega)\leq\mu(\omega)\ P\text{-a.s.}$$

For each pair w_j there is a given random traffic demand $D_j(\omega)\geq 0$ so that $(D_1(\omega),\dots,D_l(\omega))$ is the demand vector. We require that the so-called traffic conservation law is fulfilled, namely that the demand $D_j(\omega)$ verifies

$$\sum_{r=1}^{m} \varphi_{jr} F_r(\omega) = D_j(\omega) \quad j = 1,\dots,l \ \ P\text{-a.s.},$$

where $\Phi = \{\varphi_{jr}\}$, $j = 1,\dots,l$, $r = 1,\dots,m$, is the pair-incidence matrix whose elements φ_{jr} are equal to 1, if the path R_r connects the pair w_j, and equal 0 otherwise. In order to guarantee general existence results under minimal assumptions we set our problem in the framework of a Hilbert space and, precisely, we assume that $F(\omega) \in L^2(\Omega, P, \mathbb{R}^m)$, $D(\omega) \in L^2(\Omega, P, \mathbb{R}^l)$ and the random cost function $C(F(\omega)) : L^2(\Omega, P, \mathbb{R}^m) \to L^2(\Omega, P, \mathbb{R}^m)$. By $L^2(\Omega, P, \mathbb{R}^m)$ we denote the class of $\mathbb{R}^m$-valued functions defined in Ω, which are square integrable with respect to the probability measure P, while the symbol $\langle \cdot, \cdot \rangle$ will denote the standard scalar product in $\mathbb{R}^m$. Moreover we set

$$\langle\langle G, F \rangle\rangle = \int_{\Omega} \langle G(\omega), F(\omega) \rangle dP_{\omega} \quad \forall F, G \in L^2(\Omega, P, \mathbb{R}^m).$$

Then the set of random feasible flows is given by

$$\mathbb{K}_P = \{F(\omega) \in L^2(\Omega, P, \mathbb{R}^m) : \ \lambda(\omega) \le F(\omega) \le \mu(\omega), \Phi F(\omega) = D(\omega) P\text{-a.s.}\},$$

which is a closed, bounded, and convex subset of $L^2(\Omega, P, \mathbb{R}^m)$.

Setting $\forall \omega \in \Omega$

$$\mathbb{K}(\omega) = \{F(\omega) \in \mathbb{R}^m : \ \lambda(\omega) \le F(\omega) \le \mu(\omega), \Phi F(\omega) = D(\omega)\},$$

$\mathbb{K}_P$ may be rewritten as

$$\mathbb{K}_P = \{F(\omega) \in L^2(\Omega, P, \mathbb{R}^m) : F(\omega) \in \mathbb{K}(\omega) P\text{-a.s.}\}.$$

We can give the following definition of equilibrium.

Definition 1. A distribution $H \in \mathbb{K}_P$ is an equilibrium distribution from the user's point of view iff

$$\begin{gathered} \forall w_j \in W, \ \forall R_q, R_s \in \mathcal{R}_j \text{ and } P\text{-a.s. there holds} \\ C_q(H(\omega)) < C_s(H(\omega)) \Longrightarrow H_q(\omega) = \mu_q(\omega) \text{ or } H_s(\omega) = \lambda_s(\omega). \end{gathered} \tag{1}$$

Now we prove that an equilibrium distribution can be characterized by means of a variational inequality.

Theorem 1. *$H \in \mathbb{K}_P$ is an equilibrium flow according to Definition 1 iff it is a solution to the variational inequality:*

$$\langle\langle C(H), F - H \rangle\rangle = \int_{\Omega} \langle C(H(\omega)), F(\omega) - H(\omega) \rangle \, dP_{\omega} \ge 0, \quad \forall F \in \mathbb{K}_P, \tag{2}$$

or, in compact form, using the expectation E^p

$$E^p(\langle C(H), F - H \rangle) \ge 0,$$

where

$$E^p(\langle H_1, H_2\rangle) = \int_\Omega \langle H_1(\omega), H_2(\omega)\rangle \, dP_\omega \quad \forall H_1, H_2 \in L^2(\Omega, P, \mathbb{R}^m).$$

Remark 1. In Sect. 3, during the proof of Theorem 1, we implicitly prove that variational inequality (2): Find $H \in \mathbb{K}_P$ such that

$$\int_\Omega \langle C(H(\omega)), F(\omega) - H(\omega)\rangle \, dP_\omega \geq 0, \quad \forall F \in \mathbb{K}_P$$

is equivalent to the pointwise finite dimensional variational inequality (3) with random parameter:

Find $H \in \mathbb{K}_P$ such that P-a.s.

$$\sum_{j=1}^{l} \sum_{R_r \in \mathcal{R}_j} C_r(H(\omega))(F_r(\omega) - H_r(\omega)) \geq 0, \quad \forall F(\omega) \in \mathbb{K}(\omega). \tag{3}$$

In the paper [9] the authors study the problem: Find $H(\omega) \in \mathbb{K}(\omega)$ such that

$$\sum_{j=1}^{l} \sum_{R_r \in \mathcal{R}_j} C_r(H(\omega))(F_r(\omega) - H_r(\omega)) \geq 0, \quad \forall F(\omega) \in \mathbb{K}(\omega)$$

without a priori assuming that $H \in \mathbb{K}_P$, but looking for conditions which ensure that $H \in \mathbb{K}_P$. In Remark 1.1 in [9] they give a positive answer under suitable conditions.

Moreover the relationship between the two formulations is further specified in Proposition 1 in [7] (see also Remark 6.1).

Remark 2. Definition 1 is equivalent to the following condition: for every $w_j \in W$ there exists a random variable $C^j(\omega)$ such that for all $R_r \in \mathcal{R}_j$ and P-a.s.

$$\begin{array}{l} C_r(H(\omega)) < C^j(\omega) \Longrightarrow H_r(\omega) = \mu_r(\omega) \\ C_r(H(\omega)) > C^j(\omega) \Longrightarrow H_r(\omega) = \lambda_r(\omega) \end{array}.$$

If $\mu_r = +\infty$ for all r, then the above conditions can be rendered as follows: for every $w_j \in W$ and P-a.s. if $C^j(\omega) := \min_{R_r \in \mathcal{R}_j} C_r(H(\omega))$, then

$$(C_r(H(\omega)) - C^j(\omega))(H_r(\omega) - \lambda_r(\omega)) = 0 \quad \forall R_r \in \mathcal{R}_j \quad P\text{-a.s.}$$

3 Proof of Theorem 1

First, we prove that (1) implies (2). By assumption $H \in \mathbb{K}_P$ is an equilibrium distribution and it is enough to prove that P-a.s.

$$\sum_{j=1}^{l} \sum_{R_r \in \mathcal{R}_j} C_r(H(\omega))(F_r(\omega) - H_r(\omega)) \geq 0, \quad \forall F(\omega) \in \mathbb{K}(\omega), \tag{4}$$

because, by integrating in Ω, we get (2).

Let $w_j \in W$ be an arbitrary O/D pair. We denote by

$$A_j := \left\{R_q \in \mathcal{R}_j : H_q(\omega) < \mu_q(\omega)\right\} \quad \text{and} \quad B_j := \left\{R_s \in \mathcal{R}_j : H_s(\omega) > \lambda_s(\omega)\right\}.$$

From (1)

$$C_q(H(\omega)) \geq C_s(H(\omega)) \quad \text{for all } R_q \in A_j \text{ and } R_s \in B_j.$$

So, there exists $\gamma_j \in \mathbb{R}$ such that

$$\inf_{R_q \in A_j} C_q(H(\omega)) \geq \gamma_j \geq \sup_{R_s \in B_j} C_s(H(\omega)).$$

Let $F(\omega) \in \mathbb{K}_p$ be arbitrary. Then, for every $R_r \in \mathcal{R}_j$, $C_r(H(\omega)) < \gamma_j$ implies $R_r \notin A_j$, hence $H_r(\omega) = \mu_r(\omega)$, therefore $F_r(\omega) - H_r(\omega) \leq 0$ and

$$(C_r(H(\omega)) - \gamma_j)(F_r(\omega) - H_r(\omega)) \geq 0.$$

Likewise, $C_r(H(\omega)) > \gamma_j$ implies $(C_r(H(\omega)) - \gamma_j)(F_r(\omega) - H_r(\omega)) \geq 0$.

Thus,

$$\begin{aligned}\sum_{R_r \in \mathcal{R}_j} C_r(H(\omega))(F_r(\omega) - H_r(\omega)) &\geq \gamma_j \sum_{R_r \in \mathcal{R}_j} (F_r(\omega) - H_r(\omega)) \\ &= \gamma_j (D(\omega) - D(\omega)) = 0.\end{aligned}$$

Hence,

$$\ll C(H(\omega)), F(\omega) - H(\omega) \gg = \sum_{j=1}^{l} \sum_{R_r \in \mathcal{R}_j} C_r(H(\omega))(F_r(\omega) - H_r(\omega)) \geq 0$$
$$\forall F(\omega) \in \mathbb{K}_p,$$

and (2) holds true.

Now, we prove that (2) implies (1). Ad absurdum we assume that there exist $w_j \in W$, R_q, $R_s \in \mathcal{R}_j$ and a set $E \in \mathcal{A}$ with $P(E) > 0$ such that in E $C_q(H(\omega)) < C_s(H(\omega))$, but $H_q(\omega) < \mu_q(\omega)$ and $H_s(\omega) > \lambda_s(\omega)$.

We set: $\delta(\omega) = \min\left\{\mu_q(\omega) - H_q(\omega), H_s(\omega) - \lambda_s(\omega)\right\} > 0$ in E.
Consider now the flow $F^*(\omega)$ defined as:

$$F^*(\omega) = H(\omega) \text{ in } \Omega \setminus E$$

$$F_r^*(\omega) = \begin{cases} H_r(\omega) & \text{if } r \neq q, s \\ H_q(\omega) + \delta(\omega) & \text{if } r = q \\ H_s(\omega) - \delta(\omega) & \text{if } r = s \end{cases} \quad \text{in } E.$$

It is easy to verify that $F^*(\omega) \in \mathbb{K}_P$, then we can calculate the variational inequality (2) in $F^*(\omega)$:

$$\int_\Omega \langle C(H(\omega)), F^*(\omega) - H(\omega)) \, dP_\omega = \int_E [C_q(H(\omega) - C_s(H(\omega)] \, \delta(\omega) \, dP_\omega < 0$$

which is an absurdity.

4 Existence of Equilibria

There are two standard approaches to the existence of equilibria, namely with and without a monotonicity requirement (see [10]). We shall employ the following definitions.

Let E be a reflexive Banach space over the reals, $\mathbb{K} \subset E$ be a nonempty, closed, and convex set, $A : \mathbb{K} \to E^*$ be a map to the dual space E^* equipped with the *weak** topology.

Definition 2. A mapping A from $\mathbb{K}$ to X^* is called pseudomonotone in the sense of Brezis (B-pseudomonotone) iff

1. For each sequence u_n weakly converging to u (in short $u_n \rightharpoonup u$) in $\mathbb{K}$ and such that $\limsup_n \langle Au_n, u_n - u \rangle \leq 0$ it results that:

$$\liminf_n \langle Au_n, u_n - v \rangle \geq \langle Au, u - v \rangle, \quad \forall v \in \mathbb{K};$$

2. For each $v \in \mathbb{K}$ the function $u \mapsto \langle Au, u - v \rangle$ is lower bounded on the bounded subsets of $\mathbb{K}$.

The following Theorem holds (see [1, 10])

Theorem 2. *Let $\mathbb{K}$ be a nonempty convex and weakly compact subset of E and A a B-pseudomonotone mapping from $\mathbb{K}$ to E^*. Then variational inequality*

$$\langle Au, v - u \rangle \geq 0, \quad \forall v \in \mathbb{K}$$

admits solutions.

In our framework, since $\mathbb{K}_P$ is a nonempty convex and weakly compact subset of $L^2(\Omega, P, \mathbb{R}^m)$, Theorem 2 becomes

Theorem 3. *If $C : L^2(\Omega, P, \mathbb{R}^m) \to L^2(\Omega, P, \mathbb{R}^m)$ is B-pseudomonotone, namely*

1. *For each sequence H_n weakly converging to H (in short $H_n \rightharpoonup H$) in $\mathbb{K}_P$ and such that $\limsup_n \langle \langle C(H_n), H_n - H \rangle \rangle \leq 0$ it results that:*

$$\liminf_n \langle \langle C(H_n), H_n - H \rangle \rangle \geq \langle \langle C(H), H - v \rangle \rangle, \quad \forall v \in \mathbb{K}_P;$$

2. *For each $v \in \mathbb{K}_P$ the function $H \mapsto \langle\langle C(H), H - v\rangle\rangle$ is lower bounded on the bounded subsets of $\mathbb{K}_P$,*

then variational inequality (2) admits solutions.

In the case of monotone approach, we need the following definitions.

Definition 3. The map $A : \mathbb{K} \to E^*$ is said to be pseudomonotone in the sense of Karamardian (K-pseudomonotone) iff for all $u, v \in \mathbb{K}$

$$\langle Av, u - v\rangle \geq 0 \implies \langle Au, u - v\rangle \geq 0.$$

Definition 4. A mapping $A : \mathbb{K} \to E^*$ is lower hemicontinuous along line segments, iff the function

$$\xi \mapsto \langle A\xi, u - v\rangle$$

is lower semicontinuous for all $u, v \in \mathbb{K}$ on the line segments $[u, v]$.

The following theorem holds (see [10]).

Theorem 4. *If $\mathbb{K}$ is convex, closed, and bounded and A is a K-pseudomonotone and lower hemicontinuous along line segments mapping, then variational inequality*

$$\langle Au, v - u\rangle \geq 0, \quad \forall v \in \mathbb{K}$$

admits solutions.

In our framework, Theorem 4 becomes

Theorem 5. *If $C : L^2(\Omega, P, \mathbb{R}^m) \to L^2(\Omega, P, \mathbb{R}^m)$ is K-pseudomonotone, namely for all $H, v \in \mathbb{K}_P$*

$$\langle\langle C(v), H - v\rangle\rangle \geq 0 \implies \ll C(H), H - v \gg \geq 0$$

and lower hemicontinuous along line segments, namely the function

$$\xi \mapsto \langle\langle C(\xi), H - v\rangle\rangle$$

is lower semicontinuous for all $H, v \in \mathbb{K}_P$ on the line segments $[H, v]$, then variational inequality (2) admits solutions.

Let us remark that if we assume that C is continuous and verifies the condition

$$\exists c_1 > 0 : \|C(H(\omega))\| \leq c_1 \|H(\omega)\|, \ P\text{-a.s.},$$

then C results to be lower hemicontinuous along line segments.

5 Conclusions

In this paper we applied the random approach used in [10] to the traffic network problem (see also [6]) with capacity constraints on the path flows. Starting from the generalized Wardrop equilibrium condition governing the dynamic traffic networks in [2], we considered a model which includes uncertainty on the data, specifically on the path flows as well as on the travel demand. So we introduced a general random traffic equilibrium problem, and we gave a random generalized Wardrop equilibrium condition and showed that the equilibrium conditions are equivalent to a random variational inequality. Moreover we provided some existence theorems. Further work is to study in this framework the duality theory and to provide an approximation procedure, but also to extend the random approach to other situations such as the case of mergers/acquisitions.

References

1. Brezis, H.: Équations et inéquations non linéaires dans les espaces vectoriels en dualité. Annales de l'Institut Fourier **18**, 115–175 (1968)
2. Daniele, P., Maugeri, A., Oettli, W.: Time-dependent traffic equilibria. J. Optim. Theory Appl. **103**(3), 543–555 (1999)
3. Evstigneev, I.V., Taksar, M.I.: Equilibrium states of random economies with locally interacting agents and solutions to stochastic variational inequalities in $\langle L_1, L_\infty \rangle$. Ann. Oper. Res. **114**, 145–165(2002)
4. Ganguly, A., Wadhwa, K.: On random variational inequalities. J. Math. Anal. Appl. **206**, 315–321 (1997)
5. Gwinner, J., Raciti, F.: On a class of random variational inequalities on random sets. Numer. Funct. Anal. Optim. **27**(56), 619-636 (2006)
6. Gwinner, J., Raciti, F.: Random equilibrium problems on networks. Math. Comput. Model. **43**(7–8), 880–891 (2006)
7. Gwinner, J., Raciti, F.: Some equilibrium problems under uncertainty and random variational inequalities. Ann. Oper. Res. **200**(1), 299–319 (2012)
8. Gürkan, G., Özge, A.Y., Robinson, S.M.: Sample-path solution of stochastic variational inequalities. Math. Program. **84**, 313–333 (1999)
9. Maugeri, A., Raciti, F.: On general infinite dimensional complementarity problems. Optim. Lett. **2**, 71–90 (2008)
10. Maugeri, A., Raciti, F.: On existence theorems for monotone and nonmonotone variational inequalities. J. Convex Anal. **16**, 899–911 (2009)

Computational Complexities of Optimization Problems Related to Model-Based Clustering of Networks

Bhaskar DasGupta

1 Introduction

For complex systems of interaction in biology and social sciences, modeled as networks of pairwise interactions of components, many successful approaches to mathematical analysis of such networks rely upon viewing them as composed of subnetworks or modules whose behaviors are simpler and easier to understand. Coupled with appropriate interconnections, the goal is to deduce emergent properties of the complete network from the understanding of these simpler subnetworks. Such modular decomposition of networks appears quite often in the application domain. For example, in social networks it is a common practice to partition the nodes of a network into modules called communities such that nodes within each community are related more closely to each other than to nodes outside the community [14, 17, 21, 35–37, 42], and similarly in regulatory networks modular decomposition has been used in studying "monotone" parts of the dynamics of a biological system [12, 16] and more generally in studying a network in terms of interconnectivity of smaller parts with well-understood behaviors [22, 43]. These problems are also closely connected to many partitioning problems in graphs based on local densities studied in other computer science applications. Simplistic definitions of modules traditionally studied in the computer science literature, such as cliques, unfortunately do not apply well in the context of biological and social networks and therefore alternate methodologies are most often used [14, 17, 21, 35–37, 42]. As in virtually all works on network partitioning and community detection, we consider a *static* model of interaction in which the network connections do not evolve over time. In this chapter we focus on one approach of modular analysis of networks, namely the *model-based* approach.

B. DasGupta (✉)
Department of Computer Science, University of Illinois at Chicago, Chicago, IL 50507, USA
e-mail: bdasgup@uic.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_5,

2 Model-Based Decomposition

In the context of biological or social interaction networks, an important problem is to *partition* the nodes into a set of so-called communities or modules of statistically significant interactions. Such partitions facilitate studying interesting properties of these graphs in their applications, such as studying the behavioral patterns of a group of individuals in a society, and serve as important steps towards computational analysis of these networks. The *general* model-based decomposition approach can be described in the following manner:

- We have an appropriate "global null model" $\mathcal{G}$ of a background random graph providing, *implicitly* or *explicitly*, the probability $p_{u,v}$ of an edge between two nodes u and v.
- The general goal is to place nodes in the same module if their interaction patterns are significantly stronger than those inferred by $\mathcal{G}$ and in different modules if their interaction patterns are significantly weaker than those inferred by $\mathcal{G}$. No a priori assumptions are made about the number of modules as opposed to some other traditional graph clustering approaches.

As an example of applicability of the above framework of model-based clustering framework, consider the following maximization version of the standard $\{+,-\}$-correlation clustering that appears in the computer science literature extensively [5, 10, 46]:

Input: an undirected graph $G = (V,E)$ with each edge $\{u,v\} \in E$ having a label $\ell_{u,v} \in \{1,-1\}$.
Valid solution: a partition $V_1,\ldots,V_k$ of V.
Objective: *maximize* $\sum_{i=1}^{k} \sum_{u,v \in V_i} \ell_{u,v}$.

The above problem can be placed in the above model-based clustering framework in the following manner:

- Let H be the graph consisting of all edges labeled 1 in G.
- Let
$$p_{u,v} = \begin{cases} 0, \text{ if } \ell_{u,v} = 1 \\ 1, \text{ otherwise} \end{cases}$$
- Let the modularity of a partition V_i be
$$\mathsf{M}(V_i) = \sum_{u,v \in V_i} \Big(a_{u,v} - p_{u,v} \Big),$$
where
$$a_{u,v} = \begin{cases} 1, \text{ if } \{u,v\} \text{ is an edge of } H \\ 0, \text{ otherwise.} \end{cases}$$
- Let the total modularity of the partition $V_1,\ldots,V_k$ be defined as $\sum_{i=1}^{k} \mathsf{M}(V_i)$.

As is well known, *every graph decomposition procedure has both pros and cons, and there exists no universal decomposition procedure that works for every application*. Any decomposition method that relies on a global null model such as the one currently discussed suffers from the drawback that each node can get attached to any other node of the graph; for another possible criticism, see [18]. To design and analyze a model-based decomposition, one faces *at least* the following three choices, each being influenced by the appropriateness in the corresponding applications:

(C1) What should be an appropriate null model $\mathcal{G}$?
(C2) How should we precisely measure the statistical significance ("fitness") of an *individual* module of the given graph?
(C3) How should we *combine* the fitnesses of individual modules to get a total fitness value for the entire network?

In this chapter, we begin with a specific choice of (C1)–(C3) that leads us to the so-called *modularity clustering*, an extremely popular decomposition method in practice in the context of both social networks [1, 32, 37, 38] and biological networks [22, 43]. Subsequently, we discuss a few other choices for (C1)–(C3). An algorithm $\mathcal{A}$ for a maximization (resp., minimization) problem is said to have an approximation ratio of ε (or simply an ε-approximation) provided $\mathcal{A}$ runs in polynomial time in the size of the input and produces a solution with an objective value no smaller than $1/\varepsilon$ times (resp., no larger than ε times) the value of the optimum. We assume that the reader is familiar with standard concepts in algorithmic design and analysis such as found in textbooks [13, 19, 48].

3 Basic Modularity Clustering

To simplify discussion, suppose that our input is an undirected unweighted graph[1] $G = (V, E)$ of n nodes and m edges, let $A = [a_{u,v}]$ denote the *adjacency matrix* of G, i.e.,

$$a_{u,v} = \begin{cases} 1, \text{ if } (u,v) \in E \\ 0, \text{ otherwise,} \end{cases}$$

and let d_u denote the degree of node u.

3.1 Definitions

In the basic version of modularity clustering as proposed by Newman and others [21, 32, 35, 36, 38], the following options for (C1)–(C3) were selected.

[1] The definitions can be easily generalized for directed and weighted graphs; see Sect. 3.5.

Choice for ($\mathcal{C}1$): The null model $\mathcal{G}$ is dependent on the *degree-distribution* of the given graph G and is given by $p_{u,v} = \frac{d_u d_v}{m}$ *with* $u = v$ *being allowed*. Such a null model preserves the distribution of the degree of each node in the given graph *in expectation*, i.e., $\sum_{v \in V} p_{u,v} = d_u$.

Choice for ($\mathcal{C}2$): If nodes u and v belong to the same partition, then one would expect $a_{u,v}$ to be significantly higher than $p_{u,v}$. This is captured by adding the term $a_{u,v} - p_{u,v}$ to the objective value of the decomposition. Thus, for a subset of nodes $V' \subseteq V$, its fitness is given by $\mathsf{M}(V') = \sum_{u,v \in V'} (a_{u,v} - p_{u,v})$.

Choice for ($\mathcal{C}3$): A partition $\mathcal{S} = \{V_1, \ldots, V_k\}$ of nodes[2] has a total fitness ("modularity") of

$$\mathsf{M}(\mathcal{S}) = \frac{1}{2m} \sum_{i=1}^{k} \mathsf{M}(V_i) = \frac{1}{2m} \sum_{i=1}^{k} \left(\sum_{u,v \in V_i} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) \right) \tag{1}$$

and our goal is to *maximize* $\mathsf{M}(\mathcal{S})$ over all possible partitions $\mathcal{S}$ of V. The $\frac{1}{2m}$ factor is introduced only for a min–max normalization of the measure [23] so that $0 \leq \max_{\mathcal{S}} \{ \mathsf{M}(\mathcal{S}) \} < 1$.

Formally, the modularity clustering (MC) problem is defined as follows:

Problem name: modularity clustering (MC).
Input: an undirected graph $G = (V, E)$.
Valid solution: a partition $\mathcal{S} = \{V_1, \ldots, V_k\}$ of V.
Objective: *maximize* $\mathsf{M}(\mathcal{S}) = \frac{1}{2m} \sum_{i=1}^{k} \left(\sum_{u,v \in V_i} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) \right)$.

In the sequel, we use OPT to denote the maximum modularity value $\max_{\mathcal{S}} \{\mathsf{M}(\mathcal{S})\}$ of a given graph G. $\mathsf{M}(\mathcal{S})$ can be equivalently represented via simple algebraic manipulation [8, 15, 37, 38] as

$$\mathsf{M}(\mathcal{S}) = \sum_{i=1}^{k} \left[\frac{m_i}{m} - \left(\frac{D_i}{2m} \right)^2 \right] \tag{2}$$

where m_i is the number of weights of edges whose *both* endpoints are in the cluster V_i and $D_i = \sum_{v \in V_i} d_v$ is the sum of degrees of the nodes in V_i.

Yet another equivalent way to represent $\mathsf{M}(\mathcal{S})$, which was found to be quite useful in proving NP-completeness when inputs are restricted to graphs with the maximum degree of any node bounded by a constant, is the following. Let m_{ij} denote the number of edges one of whose endpoints is in V_i and the other in V_j and $D_i = \sum_{v \in V_i} d_v$ denote the sum of degrees of nodes in cluster V_i. Then,

$$\mathsf{M}(V_i) = \frac{1}{2m} \left(\sum_{u \in V_i, v \notin V_i} \left(\frac{d_u d_v}{2m} - a_{u,v} \right) \right)$$

[2] Each V_i is usually called a "cluster".

and this gives us the following third equation of modularity (note that now each pair of clusters contributes to the sum in Eq. (3) *exactly once*):

$$\mathsf{M}(\mathcal{S}) = \sum_{V_i, V_j \,:\, i<j} \left(\frac{D_i D_j}{2m^2} - \frac{m_{ij}}{m} \right) \tag{3}$$

An important special case of the MC problem arises [8, 15] if we restrict the maximum number of partitions of V to some pre-specified value κ. This special case, referred to as the *modularity* κ*-clustering* (κ-MC) problem, is thus formally defined as follows.

Problem name: modularity κ-clustering (κ-MC).
Input: an undirected graph $G = (V, E)$.
Valid solution: a partition $\mathcal{S} = \{V_1, \ldots, V_k\}$ of V with $k \leq \kappa$.
Objective: *maximize* $\mathsf{M}(\mathcal{S}) = \frac{1}{2m} \sum_{i=1}^{k} \left(\sum_{u,v \in V_i} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) \right)$.

In the sequel, we use OPT_κ to denote the maximum modularity value of the modularity κ-clustering problem for a given graph. The usefulness of the κ-MC problem in designing approximation algorithms for the MC problem is brought out by the following lemma.

Lemma 1 ([15]). *For any* $\kappa \geq 1$, $\mathsf{OPT}_\kappa \geq \left(1 - \frac{1}{k}\right) \mathsf{OPT}$.

Thus, in particular, $\mathsf{OPT}_2 \geq \mathsf{OPT}/2$ and, for large enough κ, OPT_κ approximates OPT very well.

3.2 *Absolute Bounds for* OPT *and* OPT_κ

Although it is difficult to specify accurately the range of values that OPT or OPT_κ may take for general graphs, it is possible to derive some bounds when the given graph G has some specific topologies. For example, bounds of the following kinds were demonstrated in [8, 15].

- If G is a complete graph, then $\mathsf{OPT} = 0$.
- If G is an union of k *disjoint* cliques each with n/k nodes, then $\mathsf{OPT} = 1 - \frac{1}{k}$.
- If G is a d-regular graph (i.e., a graph in which every node has a degree of *exactly* d), then

$$\mathsf{OPT} > \frac{0.26}{\sqrt{d}}, \text{ if } n > 40d^9$$
$$\mathsf{OPT} > \frac{0.86}{d} - \frac{4}{n}, \text{ otherwise}$$

- If G is a graph in which every node has a degree of *at most* d and $d < \frac{\sqrt[5]{n}}{16 \ln n}$, then $\mathsf{OPT} > \frac{1}{8d}$.
- For any graph G and any κ, $0 \leq \mathsf{OPT}_\kappa \leq 1 - \frac{1}{\kappa}$.

3.3 Computational Hardness Results

3.3.1 NP-Hardness Results

It was shown in [8] that computing OPT is NP-complete for sufficiently dense graphs (graphs in which nodes have degrees roughly $\Omega(\sqrt{n})$ for every node) and this NP-completeness result for dense graphs holds even if one wishes to compute just OPT_2. A basic idea behind many of these reductions is that large size cliques of the graph are properly contained within a community. The authors in [15] show that computing OPT_2 is NP-complete even if the given graph is G sparse and regular, namely even if G is a d-regular graph for any fixed $d \geq 9$. The NP-completeness proof in [15] for sparse graphs, motivated by the proof for this case in [8], is from the *graph bisection* problem for 4-regular graphs which is known to be NP-complete [28]. Intuitively, in this reduction an optimal solution for the modularity 2-clustering problem is constrained to have *exactly* the same number of nodes in each community.

3.3.2 Beyond NP-Hardness: APX-Hardness Results

A minimization problem is said to be APX-hard if it cannot be approximated within a factor of $1+\varepsilon$ for some constant $\varepsilon > 0$ under the assumption of $\mathsf{P} \neq \mathsf{NP}$. The authors in [15] showed that computing OPT_κ for any $\kappa > 1$ is APX-hard for dense regular graphs, namely for d-regular with $d = n-4$. This approximation gap is derived from the following approximation gap of the maximum independent set problem for 3-regular graphs [11]:

Problem name:	Maximum Independent Set for 3-regular graphs (3-MIS).
Input:	a graph $H = (V,E)$ that is 3-regular, i.e., every node has a degree of exactly 3.
Valid solution:	a subset $V' \subset V$ of nodes such that every pair of nodes u and v in V' is *independent*, i.e., $\{u,v\} \notin E$.
Objective:	*maximize* $\lvert V' \rvert$.
Approximation gap as derived in [11]:	NP-hard to decide if $\max_{V' \subseteq V} \left\{ \lvert V' \rvert \right\} \geq \frac{95}{194}\lvert V \rvert$ or if $\max_{V' \subseteq V} \left\{ \lvert V' \rvert \right\} \leq \frac{94}{194}\lvert V \rvert$.

The reduction is carried out by providing the edge-complement of the graph H as the input graph G to the MC problem, i.e., the input to MC is $G = (V,E)$ with $E = \{\{u,v\} \mid u,v \in V, \{u,v\} \notin F\}$. The reduction was completed in [15] by proving the following bounds for any κ:

- If $\max_{V' \subseteq V} \left\{ |V'| \right\} \geq \frac{95}{194}|V|$ then $\mathsf{OPT}_\kappa > \frac{0.9388}{|V|-4}$.
- If $\max_{V' \subseteq V} \left\{ |V'| \right\} \leq \frac{94}{194}|V|$ then $\mathsf{OPT}_\kappa < \frac{0.9382}{|V|-4}$.

This provides the desired inapproximability result with $\varepsilon = 1 - \frac{0.9388}{0.9382} \approx 0.0006$. The intuition behind a proof of the above bounds is that, for the type of sparse graphs H that is considered in the reduction, edge-complements of large-size independent set of nodes in H must be properly contained within a cluster of G and that $\mathsf{OPT}_\kappa \leq \mathsf{OPT}_2$ for any $\kappa > 2$.

3.4 Approximation Algorithms

In this section, we review several combinatorial and algebraic method for designing approximation algorithms for the MC and κ-MC problems.

3.4.1 Greedy Heuristics

As a first attempt at designing approximation algorithms for MC, one may be tempted to use a greedy approach of the following type that can easily be implemented to run in $\mathrm{O}\left(n^2 \log n\right)$ time [8]:

---

1. Start with each node being a separate cluster. Let $\mathcal{C}^0 = \left\{ \{v\} \,|\, v \in V \right\}$ be this initial clustering.

2. **for** $i = 1, 2, \ldots, n-1$ **do**

- Merge two clusters of $\mathcal{C}^{i-1}$ that yield a clustering with the largest increase or the smallest decrease in modularity.
- Let $\mathcal{C}^i$ be the new clustering obtained.

endfor

3. Return $\max\limits_i \left\{ \mathsf{M}\left(\mathcal{C}^i\right) \right\}$ as the solution.

Consider the graph $G = (V, E)$ consisting of the union of two *disjoint* cliques V_1 and V_2, each having $n/2$ nodes, along with $n/2$ additional edges corresponding to an arbitrary maximum bipartite matching $\left\{ \{u, v\} \,|\, u \in V_1, v \in V_2 \right\}$ among nodes in V_1 and V_2. Brandes et al. [8] observed that the above greedy approach has an unbounded approximation ratio on this graph by showing that the greedy algorithm obtains a modularity value of 0 even though OPT is very close to $1/2$. Thus, greedy approaches do not seem very promising in designing algorithms with bounded approximation ratios.

3.4.2 Linear Programming-Based Approach

It is possible to formulate the modularity clustering problem with arbitrarily many clusters as an integer linear program (ILP) in the following manner. For every two distinct nodes $u, v \in V$, let $x_{u,v}$ be a Boolean variable defined as:

$$x_{u,v} = \begin{cases} 0, & \text{if } u \text{ and } v \text{ belong to the same cluster} \\ 1, & \text{otherwise} \end{cases}$$

One constraint of partitioning the nodes into clusters is the so-called triangle inequality constraint:

> if u, v and v, z belong to the same cluster then u, z must also belongs to the same cluster.

This is easily described by the linear (inequality) constraint $x_{u,z} \leq x_{u,v} + x_{v,z}$. Noting that $1 - x_{u,v}$ is the contribution of a pair of distinct nodes u, v to the modularity value computed by Eq. (1), we arrive as the following equivalent ILP formulation of the MC problem [1, 8, 15]:

$$\begin{array}{l} \textit{maximize} \displaystyle\sum_{u,v \in V:\, u \neq v} \left(\frac{a_{u,v} - \frac{d_u d_v}{2m}}{2m} \right) (1 - x_{u,v}) - \sum_{v \in V} \frac{d_v^2}{2m} \\ \textit{subject to} \\ \quad \forall u \neq v \neq z : \; x_{u,z} \leq x_{u,v} + x_{v,z} \\ \quad \forall u \neq v : \; x_{u,v} \in \{0, 1\} \end{array}$$

However, solving an ILP exactly is in general an NP-hard problem. A natural approach is therefore to consider the linear programming (LP) relaxation of the ILP obtained by replacing the constraints "$\forall u \neq v : x_{u,v} \in \{0, 1\}$" by $\forall u \neq v : 0 \leq x_{u,v} \leq 1$, solving this LP in polynomial time [26], and then use some type of "rounding" scheme to convert fractional values of variables to Boolean values.[3] The authors in [1] used such an LP-relaxation with several rounding schemes for empirical evaluations.

Unfortunately, [15] showed that this LP-relaxation-based approach, irrespective of the rounding scheme used, may not be a very good choice for designing approximation algorithms with good guaranteed approximation ratio in the following manner. Let OPT_f denote the optimal objective value of the LP obtained from the ILP. Then, it was shown in [15] that, for every $d > 3$ and for all sufficiently large n, there exists a d-regular graph with n nodes such that the integrality gap $\mathsf{OPT}_f / \mathsf{OPT}$ is $\Omega(\sqrt{d})$, and thus an approximation ratio of $o(\sqrt{n})$ would be impossible to achieve irrespective of the rounding scheme used.

[3] See [48, part II] for further details of such an approach.

3.4.3 Spectral Partitioning Approach

Spectral partitioning methods for graph decomposition problems are well known [41, 45]. This approach was first suggested by Newman in [37] for the 2-MC problem but a theoretical analysis of the approximation ratio of this approach is *not* yet known. Consider the $n \times n$ symmetric matrix $W = [w_{u,v}]$ with $w_{u,v} = a_{u,v} - \frac{d_u d_v}{2m}$, and suppose that W has an eigenvector $\mathbf{u}_i$ with a corresponding eigenvalue b_i for $i = 1, 2, \ldots, n$. For every node $u \in V$, let x_u be a selection variable defined as:

$$x_u = \begin{cases} -1, & \text{if } u \text{ is assigned to cluster 1 } (V_1) \\ 1, & \text{if } u \text{ is assigned to cluster 2 } (V_2 = V \setminus V_1) \end{cases}$$

and let $X = [x_u]$ be the $1 \times n$ column vector of these selection variables such that $X = \sum_{i=1}^{n} a_i \mathbf{u}_i$ with $a_i = \mathbf{u}_i^{\mathrm{T}} X$. Then, it can be shown that

$$\mathsf{M}(S) = \frac{1}{4m} \sum_{i=1}^{n} (\mathbf{u}_i^{\mathrm{T}} X)^2 b_i.$$

Thus, one would like to select X proportional to the eigenvector with the largest eigenvalue to maximize $\mathsf{M}(S)$. However, such an eigenvector will in general have entries that are *not* ± 1 but real values. This would therefore require exploring some nontrivial "rounding scheme" for such an eigenvector to convert the real values of the components of the eigenvector to ± 1 such that the new value of objective does not decrease too much; currently, no such rounding scheme is known.

This approach can also be applied to the MC problem by using the same approach recursively to decompose the clusters V_1 and V_2 adjusting the objective function to reflect the fact that certain edges have been deselected by the partitioning, and continuing in this fashion until the modularity value cannot be improved further.

3.4.4 Quadratic Programming-Based Approach

Using the fact that $\mathsf{OPT}_2 \geq \mathsf{OPT}/2 \geq \mathsf{OPT}_\kappa/2$ for any $\kappa > 2$, it follows that an algorithm for 2-MC having an approximation ratio of ε also provides an algorithm for κ-MC having an approximation ratio of 2ε. The quadratic programming-based approach discussed in this section provides an approximation algorithm for 2-MC, thereby also providing an approximation algorithm for κ-MC for any $\kappa > 2$. As in the previous section, for every $u \in V$ let x_u be a selection variable defined as:

$$x_u = \begin{cases} -1, & \text{if } u \text{ is assigned to cluster 1 } (V_1) \\ 1, & \text{if } u \text{ is assigned to cluster 2 } (V_2 = V \setminus V_1) \end{cases}$$

Then, since $\sum_{u,v \in V} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) = 0$, Eq. (1) can be rewritten for the 2-MC problem as

$$\mathsf{M}(\mathcal{S}) = \frac{1}{4m} \left(\sum_{u,v \in V} w_{u,v} (1 + x_u x_v) \right) = \frac{1}{4m} \sum_{u,v \in V} w_{u,v} x_u x_v = \mathbf{x}^{\mathrm{T}} W \mathbf{x} \qquad (4)$$

where $w_{u,v} = \frac{a_{u,v} - \frac{d_u d_v}{2m}}{4m}$, $W = [w_{u,v}] \in \mathbb{R}^{n \times n}$ is the corresponding symmetric matrix of $w_{u,v}$'s and $\mathbf{x} \in \{-1, 1\}^n$ is a column vector of the indicator variables. Note that the $w_{u,v}$ values can be positive or negative, but $w_{u,u} = -\frac{d_u^2}{2m}$ is always negative.

Equation (4) describes a quadratic form with *arbitrary real* coefficients. As a first attempt, one might be tempted to use an existing semi-definite programming (SDP)-based approximation on quadratic forms to obtain an efficient algorithm. However, a direct application of many previously known results on SDP-based approximation is not possible. For example, the results in [9] cannot be directly applied since the diagonal entries $w_{u,u}$ are negative, the results in [40] cannot be directly applied since the coefficient matrix W is not necessarily positive-semidefinite, and even the elegant results on Grothendieck's inequality in [4] cannot be applied because we do not have a bipartition of the nodes.

However, the authors in [15] were able to adopt the techniques in [4, 9] in a nontrivial manner to provide a *randomized* approximation algorithm with an appoximation ratio of ρ, where

$$\mathbb{E}[\rho] = \begin{cases} 8.4 \ln d = \mathrm{O}(\log d), & \text{if } G \text{ is a } d\text{-regular graph with } d < \frac{n}{2 \ln n} \\ \mathrm{O}(\log d_{\max}), & \text{if } d_{\max}, \text{ the maximum degree over all nodes, is at most } \frac{\sqrt[5]{n}}{16 \ln n} \end{cases}$$

We briefly outline the proof for the $\mathrm{O}(\log d)$ bound when G is d-regular with $d < \frac{n}{2 \ln n}$. Consider the matrix $W' = [w'_{u,v}]$, where

$$w'_{u,v} = \begin{cases} 0, & \text{if } u = v \\ w_{u,v}, & \text{otherwise.} \end{cases}$$

First, it is shown that if $\mathsf{OPT}_2 = \max_{\mathbf{x} \in \{-1,1\}^n} \mathbf{x}^{\mathrm{T}} W \mathbf{x}$ and $\mathsf{OPT}'_2 = \max_{\mathbf{x} \in \{-1,1\}^n} \mathbf{x}^{\mathrm{T}} W' \mathbf{x}$, then $\mathsf{OPT}'_2 > \mathsf{OPT}_2 - \frac{1}{n}$. Then, the following lower bound on OPT_2 is derived:

$$\mathsf{OPT}_2 > \begin{cases} 0.13/\sqrt{d}, & \text{if } n > 40 d^9 \\ \frac{0.43}{d} - \frac{2}{n}, & \text{otherwise} \end{cases}$$

This shows that it suffices to approximate OPT'_2. Note that the diagonal entries of the matrix W' are now zeroes and $\mathsf{OPT}'_2 = \Omega(1/d)$. Next, we utilize the following algorithmic result on quadratic forms proven in [4, 9]. Consider the following randomized approximation algorithm:

Randomized approximation algorithm in [4, 9] for computing
$\mathsf{OPT}'_2 = \max_{\mathbf{x}\in\{-1,1\}^n} \mathbf{x}^T\mathbf{W}'\mathbf{x} = \max_{\forall \mathbf{u}:\, \mathbf{x_u}\in\{-1,1\}} \sum_{\mathbf{u},\mathbf{v}\in\mathbf{V}} \mathbf{w'_{u,v}x_u x_v}$

1. Solve the following maximization problem

$$\begin{aligned} &\textit{maximize} \sum_{\substack{u,v\in V \\ u\neq v}} w'_{u,v} X_u X_v \\ &\textit{subject to} \\ &\quad \forall u \in V:\ X_u \in \mathbb{R}^n \\ &\quad \forall u \in V:\ X_u \text{ is a symmetric positive semi-definite matrix} \end{aligned}$$

in polynomial time using the semidefinite programming approach.[4]
Let the solution vectors be X_u^* for $u \in V$.

2. Select a suitable real number $T > 1$.

3. Let $\mathbf{r}$ be a vector selected uniformly over the n-dimensional unit-norm hypersphere.

4. Set $x_u = \begin{cases} 1, & \text{if } \mathbf{Y}_u\mathbf{r} > T \\ -1, & \text{if } \mathbf{Y}_u\mathbf{r} < -T \end{cases}$
Otherwise, if $-T \le \mathbf{Y}_u\mathbf{r} \le T$, set $x_u = \begin{cases} 1 & \text{with probability } \frac{1}{2} + \frac{\mathbf{Y}_u\mathbf{r}}{2T} \\ -1 & \text{with probability } \frac{1}{2} - \frac{\mathbf{Y}_u\mathbf{r}}{2T} \end{cases}$

5. Return $\{x_u \,|\, u \in V\}$ as the solution.

The bounds in [4, 9] imply that the above algorithm returns a solution satisfying

$$\mathbb{E}\left[\sum_{u,v\in V} w'_{u,v} x_u x_v\right] \ge \frac{\mathsf{OPT}'_2}{T^2} - 8\,\mathrm{e}^{-T^2/2}\left(\sum_{u,v\in V} \left|w'_{u,v}\right|\right)$$

The proof can then be completed by showing that $\sum_{u,v\in V} \left|w'_{u,v}\right| < 2$ and selecting $T = \sqrt{4\ln d}$.

3.4.5 Other Heuristic Approaches

Other approaches for solving the MC problem include:

- simple heuristics without any guarantee of performance, and
- simulated-annealing type approaches that are exhaustive and slow [22] and therefore difficult to apply to large-scale networks with thousands of nodes.

[4] See [48, Chap. 26].

3.5 Extensions to Directed or Weighted Networks

An extension of the basic modularity clustering to a more general weighted directed network is easy and was done by Leicht and Newman [32] in the following manner. Suppose that our input is a directed weighted graph $G = (V, E, w)$ of n nodes where $w\colon E \mapsto \mathbb{R}^+$ denotes a function giving a positive weight to every edge in E, and let $A = [a_{u,v}]$ denote the weighted adjacency matrix of G, i.e.,

$$a_{u,v} = \begin{cases} w(u,v), & \text{if } (u,v) \in E \\ 0, & \text{otherwise.} \end{cases}$$

Let $d_u^{\text{in}} = \sum_{(v,u)\in E} w(v,u)$ and $d_u^{\text{out}} = \sum_{(u,v)\in E} w(u,v)$ denote the *weighted* in-degree and the *weighted* out-degree of node u, respectively, and let $m = \sum_{(u,v)\in E} w_{u,v}$ denote the sum of weights of all the edges. Then, Eq. (1) computing the modularity value of a cluster $C \subseteq V$ needs to be modified as

$$\mathsf{M}(C) = \frac{1}{m}\left(\sum_{u,v\in C}\left(a_{u,v} - \frac{d_u^{\text{out}} d_v^{\text{in}}}{m}\right)\right)$$

The authors in [15] showed that with some effort almost all our computational complexity results for modularity clustering on undirected networks can be extended to directed weighted networks.

4 Other Model-Based Graph Decomposition

In this section we discuss a few other choices for the (℃1)–(℃3) items for model-based graph decomposition.

4.1 Alternate Null Models (Alternate Choices for (℃1))

A natural objection to the basic modularity clustering is that the background degree-dependent null model may not be appropriate in all applications. We discuss a few other choices that have been explored in the literature.

4.1.1 Scale-Free Null Model

The choice of the linear preferential attachment model for the class of scale-free networks [6] may not be an appropriate choice since Karrer and Newman [27] showed that this may not provide a new null model. However, it is still an open question

as to whether other generative models for scale-free networks, such as the "copy" model by Kumar et al. [30] in which new nodes choose an existing node at random and copy a fraction of the links of this node, provide a new and useful null model.

4.1.2 Classical Erdös–Rényi Null Models

A theoretically appealing choice is the classical Erdös–Rényi random graph model, e.g., the random graph $G(n,p)$ in which each possible edge $\{u,v\}$ is selected uniformly and randomly with a probability of p. Although the Erdös–Rényi model has a rich and beautiful theory [7] with significant applications in other areas of computer science, it is by now agreed upon that such a model may be inadequate in many social and biological network applications. Nonetheless, a formal investigation of such a null model is of independent theoretical interest and may provide insight regarding the properties that an appropriate null model must satisfy. If p is selected such that the expected number of edges of the random graph is equal to the number of edges of the given graph, then maximizing modularity with this new null model is precisely the same as maximizing modularity in an appropriate regular graph [15]; otherwise, however, it is not clear what the complexity of computing this new modularity value is.

4.1.3 Application Specific Null Models

Sometimes null models motivated by specific applications in biology and social sciences are used by the researchers. Two such null models are described next.

Null Models for Transcriptional and Signaling Biological Networks

One of the most frequently reported topological characteristics of such networks is the distribution of in-degrees and out-degrees of nodes, which is close to a power law or a mixture of a power law and an exponential distribution [2, 20, 33]. Specifically, in biological applications, metabolic and protein interaction networks are heterogeneous in terms of node degrees and exhibit a degree distribution that is a mixture of a power law and an exponential distribution [2, 20, 24, 33, 34], whereas transcriptional regulatory networks exhibit a power law out-degree distribution and an exponential in-degree distribution [31, 44]. Based on these types of known topological characterizations, Albert et al. [3] suggested some degree distributions and network parameters for generating random transcriptional and signaling networks for the null model. Random networks with prescribed degree distributions can be generated in a variety of ways, e.g., by using the method suggested by Newman et al. in [39].

Markov-Chain Null Model

In this method, a random network for the null model is generated by starting with the given input network $G = (V, E)$ and repeatedly swapping randomly chosen pairs of connections in the following manner [25]:

repeat
- Select two edges, $\{a, b\}$ and $\{c, d\}$ randomly and uniformly among all edges in E.
- **If** $a = c$ or $b = d$ or $\{a, d\} \in E$ or $\{b, c\} \in E$
  then discard this pair of edges
  else add the edges $\{a, d\}$ and $\{b, c\}$ to E
  delete the edges $\{a, b\}$ and $\{c, d\}$ from E

until a specified percentage of edges of G has been replaced

4.2 Alternate Fitness Measures (Alternate Choices for (C2)–(C3))

Exact or approximate solutions to the modularity measure as described by (1) may tend to produce many trivial clusters of single nodes. For example, DasGupta and Desai in [15] showed that if the maximum node degree $d_{\max}$ of G satisfies $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$, then there is a clustering in which every cluster except one consists of a single node and the modularity value is at least 25 % of the optimal. One reason for such a consequence is due to the fact that the fitness measure for a modularity clustering is the *sum* of fitnesses of individual clusters (i.e., for a clustering $\mathcal{S} = \{V_1, V_2, \ldots, V_k\}$, $\mathsf{M}(\mathcal{S})$ is the summation of $\mathsf{M}(V_i)$'s), and one moderately large cluster sometimes over-compensates the negative effects of many small clusters.

Based on these observations, it is reasonable to investigate other suitable choices of the function that combines the individual fitness values into a global fitness measure without sacrificing the quality of the optimal decomposition. Some reasonable choices include the max-min objective, namely

$$\mathsf{M}^{\text{max-min}}(\mathcal{S}) = \min_{V_i \in \mathcal{S}} \mathsf{M}(V_i),$$

and the average objective, namely

$$\mathsf{M}^{\text{average}}(\mathcal{S}) = \frac{\sum_{i=1}^{k} \mathsf{M}(V_i)}{k}.$$

DasGupta and Desai investigated the max-min objective in [15] and showed that the max-min objective indeed avoids generating small-size trivial clusters and the optimal objective value for max-min objective is precisely scaled by a factor of 2 from that of the objective of the basic modularity clustering, thereby keeping the overall quantitative measure the same.

5 Conclusion and Further Research

There is still a large gap between the 1.0006 factor inapproximability result and logarithmic factor approximation algorithm known for modularity clustering problems. Designing better scalable algorithms for these problems would enable one to apply this method to much larger networks than that is currently done. A few interesting directions for future algorithmic research are as follows:

- Is it possible to do a nontrivial analysis of the spectral partitioning approach discussed in Sect. 3.4.3, perhaps by using the techniques presented in analysis of the spectral method for MAX-CUT such as in [47]?
- Is it possible to augment the ILP formulation for modularity clustering as discussed in Sect. 3.4.2 with additional redundant constraints using the cutting plane approach [29] to decrease the integrality gap substantially and perhaps thereby obtaining an improved approximation algorithm?

Acknowledgments The author was supported by NSF grant IIS-1160995.

References

1. Agarwal, G., Kempe, D.: Modularity-maximizing graph communities via mathematical programming. Eur. Phys. J. B **66**(3), 409–418 (2008)
2. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
3. Albert, R., DasGupta, B., Dondi, R., Kachalo, S., Sontag, E., Zelikovsky, A., Westbrooks, K.: A novel method for signal transduction network inference from indirect experimental evidence. J. Comput. Biol. **14**(7), 927–949 (2007)
4. Alon, N., Naor, A.: Approximating the cut-norm via Grothendieck's inequality. In: Proceedings of the 36th ACM Symposium on Theory of Computing, pp. 72–80. ACM, New York (2004)
5. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**(1–3), 89–113 (2004)
6. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
7. Bollobás, B.: Random Graphs, 2nd edn. Cambridge University Press, Cambridge (2001)
8. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Trans. Knowl. Data Eng. **20**(2), 172–188 (2007)
9. Charikar, M., Wirth, A.: Maximizing quadratic programs: extending Grothendieck's inequality. In: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, pp. 54–68 (2004)
10. Charikar, M., Guruswami, V., Wirth, A.: Clustering with qualitative information. In: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, Boston, pp. 524–533 (2003)
11. Chlebík, M., Chlebíková, J.: Complexity of approximating bounded variants of optimization problems. Theor. Comput. Sci. **354**(3), 320–338 (2006)
12. Coleman, T., Saunderson, J., Wirth, A.: Local-search 2-approximation for 2-correlation-clustering. In: Proceedings of the 16th Annual European Symposium on Algorithms. Lecture Notes in Computer Science, Springer Verlag, vol. 5193, pp. 308–319 (2008)

13. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT Press, Cambridge (2001)
14. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. P09008 **2005**(9)
15. DasGupta, B., Desai, D.: Complexity of Newman's community finding approach for social networks. J. Comput. Syst. Sci. **79**(1), 50–67 (2013)
16. DasGupta, B., Andres Enciso, G., Sontag, E., Zhang, Y.: Algorithmic and complexity results for decompositions of biological networks into monotone subsystems. Biosystems **90**(1), 161–178 (2007)
17. Flake, G.W., Lawrence, S.R., Giles, C.L., Coetzee, F.M.: Self-organization and identification of web communities. IEEE Comput. **35**, 66–71 (2002)
18. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proc. Natl. Acad. Sci. **104**(1), 36–41 (2007)
19. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman & Company, New York (1979)
20. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M.: A protein interaction map of *Drosophila melanogaster*. Science **302**(5651), 1727–1736 (2003)
21. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci **99**, 7821–7826 (2002)
22. Guimera, R., Sales-Pardo, M., Amaral, L.A.N.: Classes of complex networks defined by role-to-role connectivity profiles. Nat. Phys. **3**, 63–69 (2007)
23. Hann, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
24. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.-L.: The large-scale organization of metabolic networks. Nature **407**, 651–654 (2000)
25. Kannan, R., Tetali, P., Vempala, S.: Markov-chain algorithms for generating bipartite graphs and tournaments. Random Struct. Algorithms **14**, 293–308 (1999)
26. Karmarkar, N.: A new polynomial-time algorithm for linear programming. Combinatorica **4**, 373–395 (1984)
27. Karrer, B., Newman, M.E.J.: Random graph models for directed acyclic networks. Phys. Rev. E **80**, 046110 (2009)
28. Kefeng, D., Ping, Z., Huisha, Z.: Graph separation of 4-regular graphs is NP-complete. J. Math. Study **32**(2), 137–145 (1999)
29. Kelley, J.E., Jr.: The cutting-plane method for solving convex programs. J. Soc. Ind. Appl. Math. **8**(4), 703–712 (1960)
30. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: Stochastic models for the web graph. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science, Redondo Beach, pp. 57–65 (2000)
31. Lee, T.I., Rinaldi, M.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.-B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science **298**(5594), 799–804 (2002)
32. Leicht, E.A., Newman, M.E.J.: Community structure in directed networks. Phys. Rev. Lett. **100**, 118703 (2008)
33. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D.J., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J.,

Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., Vidal, M.: A map of the interactome network of the metazoan *C. elegans*. Science **303**, 540–543 (2004)
34. Maayan, A., Jenkins, S.L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N.J., Weng, G., Ram, P.T., Rice, J.J., Kershenbaum, A., Stolovitzky, G.A., Blitzer, R.D., Iyengar, R.: Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science **309**(5737), 1078–1083 (2005)
35. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256, (2003)
36. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B **38**, 321–330 (2004)
37. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**, 8577–8582 (2006)
38. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
39. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E **64**(2), 026118–026134 (2001)
40. Nesterov, Y.: Semidefinite relaxation and nonconvex quadratic optimization. Optim. Methods Softw. **9**, 141–160 (1998)
41. Pothen, A., Simon, D.H., Liou, K.P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. **11**, 430–452 (1990)
42. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**, 036106 (2007)
43. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L.: Hierarchical organization of modularity in metabolic networks. Science **297**(5586), 1551–1555 (2002)
44. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet. **31**, 64–68 (2002)
45. Simon, H.D., Teng, S.H.: How good is recursive bisection. SIAM J. Sci. Comput. **18**, 1436–1445 (1997)
46. Swamy, C.: Correlation clustering: maximizing agreements via semidefinite programming. In: Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, pp. 526–527 (2004)
47. Trevisan, L.: Max cut and the smallest eigenvalue. In: Proceedings of the 41st ACM Symposium on Theory of Computing, New York, pp. 263–272 (2009)
48. Vazirani, V.: Approximation Algorithms. Springer, Berlin (2001)

On Distributed-Lag Modeling Algorithms by *r*-Convexity and Piecewise Monotonicity

Ioannis C. Demetriou and Evangelos E. Vassiliou

1 Introduction

A linearly distributed lag model in time series data is used to predict current values of a dependent variable y based on both the current value of an independent variable x and lagged values of x. Specifically, the data are the pairs (x_t, y_t), $t = 1, 2, \ldots, n + m - 1$, where we assume that y_t is given approximately by a weighted sum of x_t and $m-1$ past values of x_t, where m is a prescribed positive number representing the lag length that is smaller than n. Thus, we have

$$y_t = \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \cdots + \beta_m x_{t-m+1} + \varepsilon_t, \tag{1}$$

where $\beta_1, \beta_2, \ldots, \beta_m$ are the unknown lag coefficients and ε_t is a random variable with zero mean and constant variance. Distributed-lag modeling refers to only the last n observations of y_t, $t = 1, 2, \ldots, m-1, m, \ldots, m+n-1$, because $m-1$ degrees of freedom are lost due to Eq. (1). With matrix notation, the unconstrained lag distribution problem is to determine a vector $\underline{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_m)$ that minimizes the objective function

$$F(\underline{\beta}) = (\underline{y} - X\underline{\beta})^T (\underline{y} - X\underline{\beta}), \tag{2}$$

where $\underline{y}^T = (y_m, y_{m+1}, \ldots, y_{m+n-1})$ and X is the $n \times m$ matrix of current and lagged values of x_t defined as

I.C. Demetriou (✉)
Division of Mathematics and Informatics, Department of Economics, University of Athens, 8 Pesmazoglou Street, Athens 10559, Greece
e-mail: demetri@econ.uoa.gr

E.E. Vassiliou
Department of Financial and Management Engineering, University of Aegean, 41 Kountourioti Street, Chios Island 82100, Greece
e-mail: e.vassiliou@aegean.gr

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_6,

$$X = \begin{pmatrix} x_m & x_{m-1} & \cdots & x_2 & x_1 \\ x_{m+1} & x_m & \cdots & x_3 & x_2 \\ x_{m+2} & x_{m+1} & \cdots & x_4 & x_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m+n-1} & x_{m+n-2} & \cdots & x_{n+1} & x_n \end{pmatrix}.$$

We assume that X has full rank. Then the unconstrained minimum of (2) is

$$\underline{\tilde{\beta}} = (X^T X)^{-1} X^T \underline{y}. \tag{3}$$

Such an estimation may give imprecise results, because of the multicollinearity that usually occurs among the lagged values of the independent variable. Nonetheless, if we avoid severe inaccuracies in the calculation of the lag coefficients due to (3), then there appear discernible patterns in the unconstrained estimate, which follow from the nature of the observations. Hence it may be necessary to assume some structure for the relation between the lag coefficients and so far there have been several suggestions in the literature.

The use of distributed lags in Economics and Electrical Engineering is very old. The paper of Levinson [28] had an important impact on the field directly and indirectly as Kailath [22] notes and several models are considered by [1, 4, 12, 13, 21, 26, 32] and [33], for instance. These models assume that the underlying function of the lag coefficients can be approximated closely by a form that depends on a few parameters. Over the years, literature on the subject agrees that some weak representation of the lag coefficients is a sensible requirement for a satisfactory model estimation (see, e.g., [18, 29] and references therein).

In this work we take the view that one knows *some properties* of an underlying relation, but one does not have sufficient information to put the relation into any simple parametric form. We assume that the rth derivative of the underlying relation allows a certain number of sign changes, which we call "prior knowledge." The prior knowledge is conveyed to the calculation through the requirement that the sequence $\{\Delta_j^r \beta : j = 1, 2, \ldots, m-r\}$ has a certain number of sign changes, where $\Delta_j^r \beta$ is the jth difference of order r of the lag coefficients β_i, $i = j, j+1, \ldots, j+r$, whose value is (see, e.g., [19])

$$\Delta_j^r \beta = (-1)^r \binom{r}{0} \beta_j + (-1)^{r-1} \binom{r}{1} \beta_{j+1} + \cdots +$$

$$\binom{r}{r-2} \beta_{j+r-2} - \binom{r}{r-1} \beta_{j+r-1} + \binom{r}{r} \beta_{j+r}. \tag{4}$$

Relation (4) is a linear combination of $\beta_i, i = j, j+1, \ldots, j+r$, where the coefficients of successive β_i are binomial coefficients prefixed by alternating signs. An immediate advantage of this approach to lag estimation is that it avoids the assumption that the relation has a form that depends on a few parameters, which occurs in many other techniques. Depending on the value of r and the number of

sign changes in (4), several particular methods may arise from this approach. In this paper we consider two methods that are both effective in lag modeling estimation and efficient in computation.

In Sect. 2 we give a brief description of a method for calculating lag coefficients that minimize (2) subject to nonnegative rth consecutive differences, where r is smaller than m [34]. It is a quadratic programming algorithm, which solves the problem very efficiently by taking advantage of certain submatrices of the Toeplitz matrices that occur during the calculation. An advantage of this method to lag coefficient estimation is that, due to the constraints on $\beta_1, \beta_2, \ldots, \beta_m$, it obtains particular properties that occur in a variety of underlying relations of the lag coefficients, such as monotonicity, convexity, concavity, and r-convexity. It is very useful that our condition on the lag coefficients allows mathematical descriptions for these nice properties.

In Sect. 3 we give brief descriptions of two procedures for estimating the lag coefficients $\beta_1, \beta_2, \ldots, \beta_m$ subject to the condition that the coefficients have at most k monotonic sections, where k is a prescribed positive number less than m. They are iterative algorithms that combine the steepest descent method and the conjugate gradient method with piecewise monotonicity on the components of $\underline{\beta}$ [11]. Starting from an initial estimate of $\underline{\beta}$, each iteration of these methods adjusts the lag coefficients by solving efficiently a combinatorial optimization calculation that imposes piecewise monotonicity constraints on the lag coefficients. An advantage of this idea to lag coefficient estimation is that piecewise monotonicity gives a property that occurs to a wide range of underlying relations of the lag coefficients.

In Sect. 4 we present an application of these methods to real quarterly macroeconomic data derived from the Federal Reserve Bank of St. Louis for the period 1959:Q2–2013:Q2. Dependent variable is the Annual Rate of Change of the GDP in United States and independent variable is the Annual Rate of Change of the Money Supply for United States. The values of m, r, and k were selected to provide a variety of models in the final lag coefficients. We present sufficient details of results intended for use as a guide to apply the methods. It is believed that the illustrative analysis of this section will be helpful for judging possible relations.

In Sect. 5 we present some concluding remarks. Numerical results that demonstrate the accuracy and the performance of these methods are presented by [10, 11] and [34].

2 Calculating the Lag Coefficients Subject to r-Convexity

In [34], we seek lag coefficients $\beta_1, \beta_2, \ldots, \beta_m$ that minimize the objective function (2) subject to the r-convexity constraints

$$\Delta_j^r \beta \geq 0, \quad j = 1, 2, \ldots, m - r. \tag{5}$$

Ideally, the fitted function of the lag coefficients is to have a nonnegative rth derivative. Functions like this are called r-convex (see [23] for a definition) and we

analogously call r-convex a vector whose components satisfy the constraints (5). Similarly the problem may well be defined for the case where the differences (4) are nonpositive,

$$\Delta_j^r \beta \leq 0, \quad j = 1,2,\ldots,m-r, \tag{6}$$

in which case we call the solution vector r-concave.

The cases $r = 1, 2$ in (5) allow very important applications. When $r = 1$, the constraints are

$$\beta_{i+1} - \beta_i \geq 0, \quad i = 1,2,\ldots,m-1, \tag{7}$$

which implies monotonically increasing coefficients (see, e.g., [31] for a general treatment of the subject of monotonic regression)

$$\beta_1 \leq \beta_2 \leq \cdots \leq \beta_m. \tag{8}$$

Analogously, the monotonically decreasing coefficients satisfy the inequalities

$$\beta_1 \geq \beta_2 \geq \cdots \geq \beta_m. \tag{9}$$

Inequalities (9) suggest that the lag coefficients become more significant as time proceeds. These constraints may be seen as a generalization of the method of [13], where the coefficients β_i are imposed to decline arithmetically.

In cases such as in production situations, the assumption $r = 2$ in (5) implies that the lag coefficients are subject to the increasing rates of change (see, [20] for a definition)

$$\beta_2 - \beta_1 \leq \beta_3 - \beta_2 \leq \cdots \leq \beta_m - \beta_{m-1}, \tag{10}$$

which is equivalent to assuming that $\{\beta_i : i = 1,2,\ldots,m\}$ satisfy the convexity conditions,

$$\beta_{i+2} - 2\beta_{i+1} + \beta_i \geq 0, \quad i = 1,2,\ldots,m-2. \tag{11}$$

By considering piecewise linear functions it can be proved that if $r = 1$ and $\underline{\beta}$ is optimal, then there exists a monotonic function that interpolates the points (i,β_i), $i = 1,2,\ldots,m$. Similarly, if $r = 2$, then there exists a convex function that interpolates (i,β_i), $i = 1,2,\ldots,m$. These statements do not generalize to larger values of r as it has been shown by Cullinan and Powell [3], which means that nonnegative differences of order $r \geq 3$ do not imply that there exists a function with a nonnegative rth derivative that interpolates the above points.

We can express the constraints (5) in the matrix form

$$D_r^T \underline{\beta} \geq 0, \tag{12}$$

where D_r is the $m \times (m-r)$ rectangular matrix, whose elements $(D_r)_{ij}$ are defined by the relation

$$(D_r)_{ij} = \begin{cases} (-1)^{r+j-i} \binom{r}{i-j}, & j \leq i \leq j+r, \quad j = 1,2,\ldots,m-r \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Now the nonzero components of the jth column of D_r are those that occur in the differences (4) giving a Toeplitz pattern that depends on the value of r.

Since the constraints on $\underline{\beta}$ are linear and consistent and since the second derivative matrix of (2) with respect to $\underline{\beta}$ is twice the positive definite matrix X^TX, the problem of minimizing (2) subject to (12) is a strictly convex quadratic programming problem. It has a unique solution, $\underline{\hat{\beta}}$ say, that is usually straightforward to calculate by standard quadratic programming methods (see, e.g., [14]). The solution depends on the Karush–Kuhn–Tucker optimality conditions (see, e.g., [27]), which state that $\underline{\beta}=\underline{\hat{\beta}}$ if and only if the constraints (12) are satisfied and there exist Lagrange multipliers $\lambda_i \geq 0, i \in \mathcal{A}$ such that the equation

$$2X^T(X\underline{\beta}-\underline{y})=\sum_{i\in\mathcal{A}}\lambda_i\underline{a}_i, \tag{14}$$

holds, where $\mathcal{A}$ is the subset $\{i:\Delta_i^r\beta=0\}$ of active constraint indices and $\underline{a}_i \in \mathrm{R}^m$ is the ith column of D_r. We define, $\lambda_i=0$, for $i\in[1,m-r]\backslash\mathcal{A}$ and denote the $(m-r)$-vector of Lagrange multipliers by $\underline{\lambda}$.

The method of [34] employs a version of the strictly convex quadratic programming calculation of [9]. It generates a sequence of subsets of the constraint indices $\{1,2,\ldots,m-r\}$, where for each subset $\mathcal{A}$ the equations

$$\underline{a}_i^T\underline{\beta}=0, \quad i\in\mathcal{A} \tag{15}$$

are satisfied and the vector $\underline{\beta}$ is obtained by minimizing the objective function (2) subject to (15). Moreover, unique Lagrange multipliers $\lambda_i, i\in\mathcal{A}$ are defined by the first order optimality condition (14). An outline of the quadratic programming method for minimizing (2) subject to (12) is as follows.

Step 0: Set $\mathcal{A}=\{1,2,\ldots,m-r\}$ and calculate the associated $\underline{\beta}$ and $\underline{\lambda}$. If any negative multipliers occur, then start removing the corresponding constraint indices from $\mathcal{A}$, one at a time, while recalculating each time $\underline{\beta}$ and $\underline{\lambda}$, until all the multipliers become nonnegative.

Step 1: If the constraints (12) are satisfied, then terminate. Otherwise record $\underline{\mu}=\underline{\lambda}$, find the most violated constraint, $\underline{a}_\kappa^T\underline{\beta}<0$ say, add κ to $\mathcal{A}$ and calculate $\underline{\beta}$ and $\underline{\lambda}$.

Step 2: If $\lambda_i\geq 0$, $i\in\mathcal{A}$, then branch to Step 1.

Step 3: Seek the greatest value of θ such that the numbers $(1-\theta)\mu_i+\theta\lambda_i$, $i\in\mathcal{A}$ are nonnegative, which implies $0\leq\theta<1$. If ρ is the value of i that gives $(1-\theta)\mu_i+\theta\lambda_i=0$, then remove ρ from $\mathcal{A}$, replace $\underline{\mu}$ by $(1-\theta)\underline{\mu}+\theta\underline{\lambda}$, calculate $\underline{\beta}$ and $\underline{\lambda}$ and branch to Step 2.

The implementation of this algorithm depends strongly on the Toeplitz structure of the constraint coefficient matrix (13) for deriving the solution of the equality constrained problem that minimizes (2) subject to (15) and the corresponding Lagrange multipliers that occur during the quadratic programming iterations. In particular, the following method is highly suitable, if, as it happens in the examples of [34], there

is a small number of inactive constraints. We express $\underline{\beta}$ as a linear combination of a basis of the subspace of vectors $\underline{\beta}$ that satisfy the equality constraints (15). Since there are no redundant equations in (15), we can find $m-p$ linearly independent m-vectors $\{\underline{u}_s : s \in S\}$ such that $\underline{a}_i^T \underline{u}_s = \underline{0}$, $s \in S$, for all $i \in \mathcal{A}$, where we let $p = |\mathcal{A}|$ be the number of elements of $\mathcal{A}$.

The following basis, proposed by Cullinan [2] and modified by Vassiliou and Demetriou [34], has the beautiful feature that the matrices that occur are banded and positive definite. We let the notation $\lfloor r/2 \rfloor^- = \lfloor r/2 \rfloor$ if r is odd and $\lfloor r/2 \rfloor^- = (\lfloor r/2 \rfloor - 1)$ if r is even, where $\lfloor q \rfloor$ denotes the largest integer that is smaller than q. Also, we let S be the index set

$$S = \{1, \ldots, \lfloor r/2 \rfloor\} \cup \{j + \lfloor r/2 \rfloor : \ j \notin \mathcal{A}\} \cup \left\{m - \lfloor r/2 \rfloor^-, \ldots, m\right\}$$

and the vectors $\underline{u}_s$, for $s \in S$ are defined by the equations

$$(\underline{u}_s)_t = \delta_{st}, \quad \text{for } s,t \in S \tag{16}$$

and

$$\underline{a}_i^T \underline{u}_s = 0, \quad \text{if } i \in \mathcal{A} \text{ and } s \in S, \tag{17}$$

where δ_{st} is the Kronecker's delta.

Each of the basis vectors is obtained by solving a $p \times p$ system of equations, whose coefficient matrix elements are derived by deleting a column of the $p \times m$ coefficient matrix (17) for each $s \in S$. Let M_r be the matrix so obtained. Depending on the sign pattern of D_r, M_r is positive definite if ($r \bmod 4 = 0, 3$) and negative definite if ($r \bmod 4 = 1, 2$) as it is proved by Demetriou and Lypitakis [7]. In addition, M_r inherits the bandwidth form of D_r. Hence, the unknown components $\{(\underline{u}_s)_{i+\lfloor r/2 \rfloor} : \ i \in \mathcal{A}\}$ can be calculated efficiently and stably by Cholesky factorization if r is even and by band LU factorization if r is odd. Since this process has to be repeated for each $s \in S$ in order to generate all the basis elements, we factorize M_r only once and subsequently use this factorization to derive the components of each basis vector.

Having defined this basis, we can work with reduced quantities throughout the calculation. We express any vector $\underline{\beta}$ that satisfies (15) in the form

$$\underline{\beta} = U \underline{\theta}, \tag{18}$$

where U is the $m \times (m-p)$ matrix whose columns are the vectors $\{\underline{u}_s : s \in S\}$ and $\underline{\theta}$ is an $m-p$ vector, whose components are to determined. Working with $\underline{\theta}$ instead of with the m-vector $\underline{\beta}$ provides two advantages. One is that there are much fewer variables, because $\mathcal{A}$ is usually kept large during the quadratic programming iterations and the other is that Eq. (15) are satisfied automatically. By substituting (18) into (2) we obtain the reduced quadratic function

$$\psi(\underline{\theta}) = \|XU\underline{\theta} - \underline{y}\|_2^2, \tag{19}$$

whose unique minimizer is calculated by applying Cholesky factorization to the first order condition

$$(XU)^T(XU)\underline{\theta} = (XU)^T\underline{y}. \tag{20}$$

Although the value of r is not a restriction to this calculation, the most popular choices restrict r to values smaller than 7 or 8. Hence, if p is close to $m-r$, as in the numerical results of [34], the amount of numerical work required to solve (20) for $\underline{\theta}$ is quite low.

Once $\underline{\beta}$ is available, the corresponding Lagrange multipliers $\{\lambda_i : i \in \mathcal{A}\}$ are defined by the first order conditions (14), which form an overdetermined system with $m-p$ redundant equations. So p equations may be chosen in order to specify the p unknowns multipliers and all possible choices will give the same solution, provided the chosen system is non-singular. In view of the magnitude of the elements of D_r, the central element (there are two such elements with opposite sign, if r is odd) of each column of D_r is also the largest in absolute value element of each column. Thus, by choosing the rows $(i+\lfloor r/2 \rfloor) \in \mathcal{A}$ of (14), we derive a system of equations whose coefficient matrix has diagonal dominance and is the transpose of the $p \times p$ matrix M_r. Since the factorization of M_r is already available from the calculation that provided the components of the basis vectors, it is remarkable that this choice, which has resulted to M_r, is also suitable to the calculation of the Lagrange multipliers.

3 A Conjugate Gradient Algorithm with Piecewise Monotonicity

In this section we consider the problem of estimating lag coefficients by minimizing (2) subject to the condition that the lag coefficients $\beta_1, \beta_2, \ldots, \beta_m$ have at most k monotonic sections, where k is a prescribed positive number that is less than m.

The problem when $k = 1$ may be solved by the structured quadratic programming calculation that is stated in Sect. 2. The problem when $k = 2$ concerns the minimization of (2) subject to the constraints

$$\left.\begin{array}{l} \beta_1 \le \beta_2 \le \cdots \le \beta_t \\ \beta_t \ge \beta_{t+1} \ge \cdots \ge \beta_m \end{array}\right\}, \tag{21}$$

where t is one of the variables of the optimization calculation. In order to identify the value of t that gives the least value of (2), one would solve $m-2$ separate quadratic programming problems in m variables, for $t = 2, 3, \ldots, m-1$.

When $k > 2$, we consider the problem of calculating a vector $\underline{\beta}$ that minimizes (2) subject to the piecewise monotonicity constraints

$$\left.\begin{array}{l} \beta_{t_{s-1}} \le \beta_{t_{s-1}+1} \le \cdots \le \beta_{t_s}, \quad \text{if } s \text{ is odd} \\ \beta_{t_{s-1}} \ge \beta_{t_{s-1}+1} \ge \cdots \ge \beta_{t_s}, \quad \text{if } s \text{ is even} \end{array}\right\}, \tag{22}$$

where the integers $\{t_s : s = 0, 1, \ldots, k\}$ satisfy the conditions

$$1 = t_0 \le t_1 \le \cdots \le t_k = m. \tag{23}$$

It is quite difficult to develop efficient optimization algorithms for calculating a solution to this problem, because the integers $\{t_s : s = 2,3,\ldots,k-1\}$ are not known in advance and they are variables in the optimization calculation. Indeed, the calculation of a global minimum of (2) would require about $O(m^k)$ separate quadratic programming calculations in m variables, which is not practicable.

Therefore, we consider an alternative form of the problem where an iterative algorithm attempts to minimize (2) by combining the conjugate gradient method with piecewise monotonicity constraints on the lag coefficients.

We begin the description with the process that involves the steepest descent method (details for the steepest descent method are given by [16], for instance) with piecewise monotonicity constraints on the lag coefficients, which we extend subsequently by the introduction of a term that requires little additional work and gives the conjugate gradient method. This process starts from an initial estimate $\underline{\beta}^{(0)}$ of $\underline{\beta}$ that satisfies the constraints (22) and generates a sequence of estimates $\{\underline{\beta}^{(j)} : j = 1,2,3,\ldots\}$ to $\underline{\beta}$ in two phases. In the first phase it takes a descent direction from the current estimate to a new estimate of $\underline{\beta}$. In the second phase it replaces the new estimate by its best piecewise monotonic approximation. In the first phase, the algorithm calculates a new estimate of the form

$$\underline{\beta}^{(j+1)} = \underline{\beta}^{(j)} + \alpha_j \underline{d}^{(j)}, \tag{24}$$

where α_j is a step-length and $\underline{d}^{(j)}$ is the search direction

$$\underline{d}^{(j)} = X^T(\underline{y} - X\underline{\beta}^{(j)}). \tag{25}$$

The step-length α_j with exact line search is determined by the minimization of the convex function of one variable $F(\underline{\beta}^{(j)} + \alpha \underline{d}^{(j)})$ which gives

$$\alpha_j = \frac{\underline{d}^{(j)T} X^T(\underline{y} - X\underline{\beta}^{(j)})}{\|X\underline{d}^{(j)}\|_2^2}. \tag{26}$$

Since (25) involves matrix X only by multiplication, ill-conditioning of X is irrelevant here. Having calculated $\underline{\beta}^{(j+1)}$, the algorithm proceeds to the second phase, which calculates a vector $\underline{\beta}$ that minimizes the sum of the squares of the residuals

$$\|\underline{\beta}^{(j+1)} - \underline{\beta}\|_2^2 = \sum_{i=1}^{m} (\beta_i^{(j+1)} - \beta_i)^2 \tag{27}$$

subject to (22), while the integers $\{t_s : s = 0,1,\ldots,k\}$ satisfy (23). This is a formidable combinatorial problem which has been solved efficiently by [8] in only $O(m^2 + km\log_2 m)$ computer operations. Some details of this calculation are given in this section, after the description of the conjugate gradient process below. The

algorithm finishes if the vector $\underline{\beta}$ found at the second phase satisfies the convergence condition

$$\|\underline{\beta} - \underline{\beta}^{(j+1)}\|_2 / \|\underline{\beta}\|_2 \leq \varepsilon, \tag{28}$$

where ε is a small positive tolerance. This test is applied at every estimate $\underline{\beta}^{(j+1)}$ including the first iteration as well. When the test (28) fails, then the algorithm replaces $\underline{\beta}^{(j+1)}$ by its best piecewise monotonic approximation vector $\underline{\beta}$, increases j by one and branches to the beginning of the first phase in order to calculate at least one new vector in the sequence $\{\underline{\beta}^{(j)} : j = 1,2,3,\ldots\}$. This gives the following algorithm.

Step 0: Set $j = 0$ and $\underline{\beta}^{(0)} = \underline{0}$.

Step 1: Calculate $\underline{d}^{(j)} = X^T(\underline{y} - X\underline{\beta}^{(j)})$.

Step 2: Calculate α_j and set $\underline{\beta}^{(j+1)} = \underline{\beta}^{(j)} + \alpha_j \underline{d}^{(j)}$.

Step 3: (Piecewise monotonic approximation) By employing Algorithm 2 of [5] calculate $\underline{\beta}$, namely a least squares approximation with k monotonic sections to $\underline{\beta}^{(j+1)}$.

Step 4: If criterion (28) is satisfied then quit, otherwise replace $\underline{\beta}^{(j+1)}$ by $\underline{\beta}$, increase j by one and branch to Step 1.

If we drop Step 3, which provides the best piecewise monotonic approximation to current $\underline{\beta}^{(j+1)}$, it can be proved that the algorithm terminates at the minimum of (2) (see, e.g., [17]). By incorporating the piecewise monotonicity constraints into this algorithm further restricts the solution because the monotonicity algorithm is norm reducing [31]. By invoking the strict convexity of (2) and the contraction mapping theorem, it can be proved (the convergence analysis of [25] is suitable to this case) that this algorithm meets the termination condition for some finite integer j. Thus it converges to a local minimum of (2) subject to the constraints (22). However, the numerical results of [10] show that this algorithm is very slow in practice, which makes it rather inefficient to be useful.

A vast improvement in efficiency is achieved by the method of [11] that combines the conjugate gradient method of Fletcher and Reeves with exact line searches as described by [14] with piecewise monotonicity constraints on the components of $\underline{\beta}$. The only change to the above steepest descent algorithm is that on most iterations the search direction is altered from (24) to the vector

$$\underline{d}^{(j)} = X^T(\underline{y} - X\underline{\beta}^{(j)}) + \gamma_j \underline{d}^{(j-1)}, \tag{29}$$

except that the last term is omitted if $j = 1$. The value of γ_j is determined by the Fletcher–Reeves [15] conjugacy condition

$$\gamma_j = \frac{\|X^T(\underline{y} - X\underline{\beta}^{(j)})\|_2^2}{\|X^T(\underline{y} - X\underline{\beta}^{(j-1)})\|_2^2}. \tag{30}$$

Then the algorithm proceeds as in the steepest descent case and is as follows.

Step 0: Set $j=0$, $\underline{\beta}^{(0)}=0$, and $\gamma_0=0$.

Step 1: Calculate $\underline{d}^{(j)}=X^T(\underline{y}-X\underline{\beta}^{(j)})+\gamma_j\underline{d}^{(j-1)}$.

Step 2: Calculate α_j and set $\underline{\beta}^{(j+1)}=\underline{\beta}^{(j)}+\alpha_j\underline{d}^{(j)}$.

Step 3: (Piecewise monotonic approximation) By employing Algorithm 2 of [5] calculate $\underline{\beta}$, namely a least squares approximation with k monotonic sections to $\underline{\beta}^{(j+1)}$.

Step 4: If criterion (28) is satisfied then quit, otherwise replace $\underline{\beta}^{(j+1)}$ by $\underline{\beta}$, calculate γ_j, increase j by one and go to Step 1.

The choice of γ_j is suitable, because the conjugate gradient method has the property that, if $F(.)$ is a convex quadratic function, if $\underline{d}^{(1)}=X^T(\underline{y}-X\underline{\beta}^{(1)})$, and if formula (29) is used for all $j\geq 2$, then (see [30]) in exact arithmetic the method terminates because $X^T(\underline{y}-X\underline{\beta}^{(j)})=\underline{0}$, for some $j\in[1,m+1]$. Hence, if we exclude Step 3 from the above algorithm, the remaining steps terminate at the unconstrained minimum of (2). In addition, the Fletcher–Reeves condition (30), in view of our quadratic function and the exact line searches we use, gives descent [14]. This is important, because the convergence of this algorithm can be established by arguments similar to those that follow the steepest descent algorithm above. Furthermore the numerical results of [11] show a considerably higher convergence speed than the method that uses the steepest descent.

In the rest of the section we discuss some properties of the piecewise monotonicity method that is employed by Step 3. Given a positive integer $k<m$, Step 3 seeks an m-vector $\underline{\beta}$ that is closest to $\underline{\beta}^{(j+1)}$ by minimizing (27) subject to the conditions that the components of $\underline{\beta}$ consist of at most k monotonic sections. Without loss of generality, we specify that the first monotonic section is increasing. The approximation process is a projection, because if $\underline{\beta}^{(j+1)}$ satisfies the constraints (22), then $\underline{\beta}=\underline{\beta}^{(j+1)}$. Therefore if $\underline{\beta}^{(j+1)}$ consists of more than k monotonic sections, then the piecewise monotonicity constraints prevent the equation $\underline{\beta}=\underline{\beta}^{(j+1)}$, which means that the integers $\{t_s: s=2,3,\ldots,k-1\}$ are all different.

The most important property of this calculation is that each monotonic section in a best piecewise monotonic fit is the optimal approximation to the corresponding data. Indeed, the components $\{\beta_i: i=t_{s-1},t_{s-1}+1,\ldots,t_s\}$ on $[t_{s-1},t_s]$ minimize the sum of the squares

$$\sum_{i=t_{s-1}}^{t_s}(\beta_i^{(j+1)}-\beta_i)^2 \tag{31}$$

subject to the constraints

$$\beta_i\leq\beta_{i+1}, i=t_{s-1},\ldots,t_s-1, \text{ if } s \text{ is odd} \tag{32}$$

and subject to the constraints

$$\beta_i\geq\beta_{i+1}, i=t_{s-1},\ldots,t_s-1, \text{ if } s \text{ is even.} \tag{33}$$

In the former case the sequence $\{\beta_i : i = t_{s-1}, t_{s-1}+1, \ldots, t_s\}$ is the best monotonic increasing fit to $\{\beta_i^{(j+1)} : i = t_{s-1}, t_{s-1}+1, \ldots, t_s\}$ and in the latter case the best monotonic decreasing one. Therefore, provided that $\{t_s : s = 2,3,\ldots,k-1\}$ are known, the components of $\underline{\beta}$ are generated by solving a separate monotonic problem on each section $[t_{s-1}, t_s]$ in only $O(t_s - t_{s-1})$ computer operations. We introduce the notation $\alpha(t_{s-1}, t_s)$ and $\beta(t_{s-1}, t_s)$ for the least value of (31) subject to the constraints (32) and (33), respectively. We denote by $\delta(k,n)$ the least value of (27) at the required minimum and, if k is odd, we obtain the expression

$$\delta(k,n) = \alpha(t_0,t_1) + \beta(t_1,t_2) + \alpha(t_2,t_3) + \cdots + \alpha(t_{k-1},t_k) \tag{34}$$

and analogously if k is even, where we replace the last term in this sum by $\beta(t_{k-1}, t_k)$.

Reversely now, we use separation and, an optimal $\underline{\beta}$ associated with the integer variables $\{t_s : s = 2,3,\ldots,k-1\}$ can split at t_{k-1} into two optimal sections. One section that provides a best fit on $[t_0, t_{k-1}]$, which is similar to $\underline{\beta}$ with one monotonic section less, and one section on $[t_{k-1}, t_k]$ that is a single monotonic fit to the remaining data, giving $\delta(k,n) = \delta(k-1, t_{k-1}) + \alpha(t_{k-1}, t_k)$ if k is odd and analogously if k is even. We note that the value of t_{k-1} in $[t_{k-2}, t_k]$ that gives $\delta(k,n)$ is independent of the integers $\{t_j : 0 \le j \le t_{k-3}\}$. Hence, it is proved by [8] that the optimization problem of Step 3 can be replaced by a problem which is amenable to dynamic programming.

The implementation of dynamic programming includes several options that are considered by Demetriou [5] and Demetriou and Powell [8]. Demetriou [6], especially, has implemented the method of [5] in Fortran and provided a software package that derives a solution in $O(m^2\nu + k\nu^2)$ computer operations, where ν is the number of local extrema of the data and where an integer p is the index of a local maximum of the sequence $\beta_i^{(j+1)}$, $i = 2,\ldots,m-1$, if $\beta_{p-1}^{(j+1)} \le \beta_p^{(j+1)}$ and $\beta_p^{(j+1)} > \beta_{p+1}^{(j+1)}$, and similarly for a local minimum. Since ν is a fraction of m, the previous complexity is reduced at least by a factor of 4. In practice, however, the method is by far more efficient than the theory indicates.

4 Numerical Results from an Example on the USA Inflation Rate

Studies for different time periods suggest that changes in the growth rate of money are reflected in the inflation rate for a long time period ahead. Among other factors the nature of the lag is important to the decision of specifying a short or a long-term policy, as, for example, it is stated by Karlson [24].

In this section some numerical results illustrate the methods of Sects. 2 and 3 by an application to real quarterly macroeconomic data that considers particular relations for the lags between money and prices. The original source of the data is the International Monetary Fund. Dependent variable is the Continuously Compounded Annual Rate of Change of the GDP Implicit Price Deflator in United States

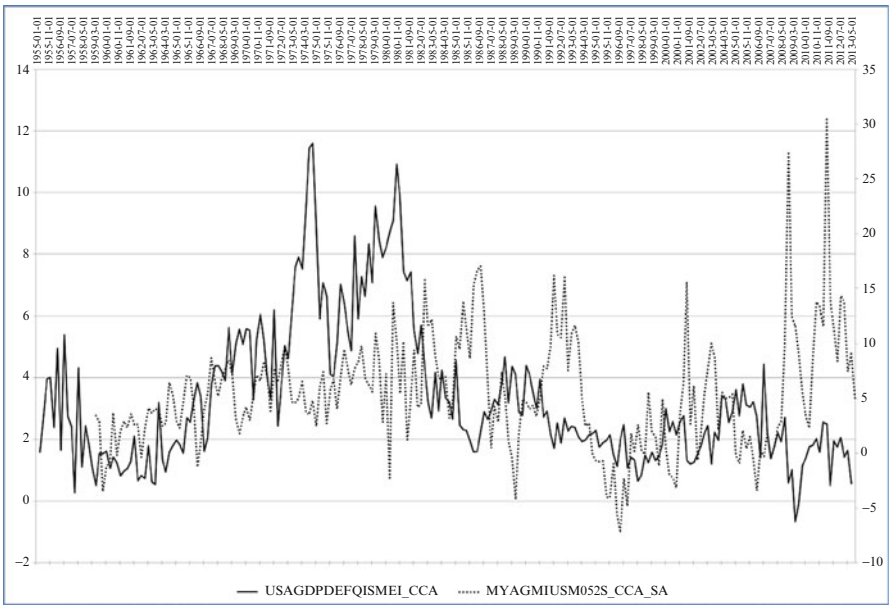


Fig. 1 Time series plots of the quarterly rates of change of the GDP Implicit Price Deflator in United States (*solid line*) for the period 1955-04-01–2013-04-01 and the quarterly rates of change of the M1 for United States (*dotted line*) for the period 1959-04-01–2013-09-01. Both time series are seasonally adjusted. The right-hand side secondary axis corresponds to M1 values

and independent variable is the Continuously Compounded Annual Rate of Change of the Money Supply for United States. The money supply variable is defined of what is known as "M1." The data amount to 217 pairs of observations x and y for the period 1959:Q2–2013:Q2 and are available from the Federal Reserve Bank of St. Louis (see, http://www.research.stlouisfed.org). Variable x is identified with the name MYAGM1USM052S and variable y with the name USAGDPDEFQISMEI in the relevant data base. The data are displayed in Fig. 1.

First we applied the method of Sect. 2 to these data. Specifically, we calculated the coefficients of the distributed-lag model (1) with $m = 17, 21$ subject to the constraints (5) on the components of $\underline{\beta}$ by allowing $r = 1, 2, 3, 4, 5, 6$, and 7. We call r-convex the coefficients so derived. Also, we call r-concave the coefficients derived by a similar calculation subject to the constraints (6). Occasionally, throughout the section, we refer to the coefficients with the term "model."

The actual values of m, r, and the calculated coefficients are given in Tables 1 and 2 for the problems with the r-convex constraints (5) and the r-concave constraints (6), respectively. The coefficients are shown in the third ($r = 1$), fourth ($r = 2$) and so on column of each table. The last column of each table displays the

unconstrained lag coefficients obtained by minimizing (2) for each m. The amount of CPU time to carry out these calculations in double precision arithmetic in a common pc is negligible. All the results are presented in four decimal digits of accuracy. The coefficients of Tables 1 and 2, for $r = 2, 3, 4$, and 5 for each m are displayed in Figs. 2, 3, 4, and 5.

The values of m, r, and the calculated Lagrange multipliers associated with the lag coefficients of Tables 1 and 2 are presented in Tables 3 and 4, respectively. Tables 3 and 4 indicate the dependencies between active constraints and Lagrange multipliers, because a Lagrange multiplier measures the marginal potential change of the value of (2) at an optimal $\underline{\beta}$, when the corresponding constraint is changed ever so slightly. The higher the value of the multiplier, the more sensitive the optimal value of the objective function is to perturbations of the corresponding constraint. A zero Lagrange multiplier in these two tables indicates an inactive constraint.

If the lag coefficients satisfy all the constraints (5) as equations, as for example in the case ($m = 17$, $r = 7$) of Table 3 where all Lagrange multipliers are positive, then all the corresponding coefficients of Table 1 lie on the best fit by a polynomial of degree at most $r-1$. If, as it is actually expected, some constraints in (5) are amply satisfied, then the r-convex lag coefficients lie on a piecewise polynomial curve, where the polynomial pieces are of degree at most $r-1$. The results of [34] show that the r-convex lag coefficients do not deviate far from the polynomial of degree $r-1$ and they do so in a smooth manner alternating above and below the polynomial curve. Hence the r-convex lag model is more flexible than the corresponding polynomial of degree $r-1$, which in fact is Almon's model [1].

We had better look at some details of Table 1 when $m = 17$. The 1-convex lag coefficients, which are presented for $r = 1$, satisfy the monotonicity constraints (8) and consist of four sections of different equal components. In view of Table 3, they are associated with the active constraint indices $\mathcal{A} = \{1, 2, \ldots, 5, 7, 9, \ldots, 15\}$ and the zero Lagrange multipliers λ_6, λ_8, and λ_{16}. The 1-convex model, simple as it is, is no liable to produce an undulating fit but only a monotonically increasing step function. The 2-convex coefficients, in view of the Lagrange multipliers in Table 3 for $r = 2$, are obtained by minimizing (2) subject to the equations $\beta_i - 2\beta_{i+1} + \beta_{i+2} = 0$, $i = 2, 3, \ldots, 14$. We see in Fig. 2 that the 2-convex model is a polygonal line with interior knots at the second and the 16th data point, which are associated with the zero Lagrange multipliers λ_1 and λ_{15} of Table 3. This polygonal line follows the general trend of the unconstrained coefficients. The 3-convex coefficients, in view of the Lagrange multipliers for $r = 3$, are obtained by minimizing (2) subject to the equations $-\beta_i + 3\beta_{i+1} - 3\beta_{i+2} + \beta_{i+3} = 0$, $i = 1, 2, \ldots, 13$, while $-\beta_{14} + 3\beta_{15} - 3\beta_{16} + \beta_{17} = 0.0167 > 0$. Hence, the first 16 coefficients lie on an increasing second degree polynomial on the range $[1, 16]$, while the 17th coefficient, $\beta_{17} = 0.0650$, due to the inactive constraint, lies over the polynomial curve toward the coefficient $\tilde{\beta}_{17} = 0.0819$. The 4-convex coefficients lie on two overlapping cubics that are obtained by minimizing (2) subject to $\beta_i - 4\beta_{i+1} + 6\beta_{i+2} - 4\beta_{i+3} + \beta_{i+4} = 0, i = 2, 3, \ldots, 10$ and $\beta_{13} - 4\beta_{14} + 6\beta_{15} - 4\beta_{16} + \beta_{17} = 0$, while the inactive constraints are associated with the zero Lagrange multipliers λ_1, λ_{11}, and λ_{12}, as we can see in Table 3 for $r = 4$. The 5-convex coefficients lie on two overlapping quartics that are obtained

Table 1 The r-convex and the unconstrained lag coefficients

		r-Convex							Unconstrained
m	β_i	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	lag coefficients
	β_1	0.0231	0.0648	0.0225	0.0685	0.0541	0.0705	0.0642	0.0772
	β_2	0.0231	0.0116	0.0232	0.0095	0.0263	0.0062	0.0168	−0.0020
	β_3	0.0231	0.0143	0.0241	0.0090	0.0123	0.0063	0.0056	0.0015
	β_4	0.0231	0.0170	0.0250	0.0110	0.0084	0.0153	0.0092	0.0190
	β_5	0.0231	0.0197	0.0261	0.0151	0.0111	0.0184	0.0164	0.0348
	β_6	0.0231	0.0224	0.0273	0.0204	0.0178	0.0204	0.0227	−0.0025
	β_7	0.0243	0.0251	0.0286	0.0265	0.0260	0.0240	0.0275	0.0425
	β_8	0.0243	0.0278	0.0301	0.0327	0.0340	0.0299	0.0315	−0.0020
17	β_9	0.0407	0.0305	0.0316	0.0382	0.0405	0.0370	0.0360	0.0728
	β_{10}	0.0407	0.0332	0.0333	0.0426	0.0444	0.0434	0.0408	0.0391
	β_{11}	0.0407	0.0359	0.0351	0.0451	0.0455	0.0470	0.0450	0.0440
	β_{12}	0.0407	0.0386	0.0370	0.0450	0.0438	0.0465	0.0466	0.0203
	β_{13}	0.0407	0.0413	0.0390	0.0419	0.0398	0.0418	0.0439	0.0755
	β_{14}	0.0407	0.0440	0.0411	0.0349	0.0348	0.0349	0.0371	0.0241
	β_{15}	0.0407	0.0467	0.0434	0.0315	0.0326	0.0308	0.0301	0.0269
	β_{16}	0.0407	0.0494	0.0458	0.0398	0.0403	0.0380	0.0344	0.0281
	β_{17}	0.0656	0.0578	0.0650	0.0681	0.0680	0.0693	0.0718	0.0819
	β_1	0.0200	0.0269	0.0069	0.0472	0.0361	0.0480	0.0405	0.0527
	β_2	0.0200	0.0255	0.0132	0.0099	0.0172	0.0080	0.0162	−0.0009
	β_3	0.0200	0.0254	0.0188	0.0016	0.0091	0.0027	0.0061	−0.0034
	β_4	0.0200	0.0254	0.0235	0.0081	0.0087	0.0081	0.0061	0.0148
	β_5	0.0230	0.0254	0.0275	0.0156	0.0134	0.0155	0.0123	0.0341
	β_6	0.0274	0.0254	0.0307	0.0233	0.0207	0.0236	0.0215	0.0078
	β_7	0.0274	0.0253	0.0330	0.0307	0.0289	0.0312	0.0310	0.0367
	β_8	0.0274	0.0253	0.0346	0.0371	0.0364	0.0375	0.0388	0.0067
	β_9	0.0274	0.0253	0.0354	0.0419	0.0420	0.0418	0.0439	0.0778
	β_{10}	0.0274	0.0253	0.0354	0.0444	0.0448	0.0437	0.0454	0.0345
21	β_{11}	0.0274	0.0253	0.0347	0.0440	0.0446	0.0429	0.0436	0.0451
	β_{12}	0.0274	0.0252	0.0331	0.0401	0.0413	0.0396	0.0389	0.0267
	β_{13}	0.0274	0.0252	0.0307	0.0337	0.0353	0.0340	0.0323	0.0753
	β_{14}	0.0274	0.0252	0.0275	0.0261	0.0272	0.0268	0.0249	0.0038
	β_{15}	0.0274	0.0252	0.0236	0.0182	0.0183	0.0188	0.0179	−0.0072
	β_{16}	0.0274	0.0251	0.0188	0.0113	0.0100	0.0114	0.0122	0.0029
	β_{17}	0.0274	0.0251	0.0133	0.0064	0.0041	0.0061	0.0082	0.0445
	β_{18}	0.0274	0.0251	0.0070	0.0047	0.0031	0.0047	0.0057	0.0023
	β_{19}	0.0274	0.0251	−0.0002	0.0072	0.0095	0.0095	0.0084	−0.0123
	β_{20}	0.0274	0.0251	0.0406	0.0397	0.0405	0.0364	0.0342	0.0442
	β_{21}	0.1036	0.1098	0.1291	0.1278	0.1277	0.1286	0.1308	0.1351

by minimizing (2) subject to $-\beta_i+5\beta_{i+1}-10\beta_{i+2}+10\beta_{i+3}-5\beta_{i+4}+\beta_{i+5}=0, i=1,2,\ldots,9$ and $-\beta_{12}+5\beta_{13}-10\beta_{14}+10\beta_{15}-5\beta_{16}+\beta_{17}=0$. And so on for $r=6,7$. We see that the r-convex coefficients for $r\geq 4$ of Table 1 when $m=17$ follow gently the trend of the unconstrained coefficients, as it is also illustrated in Fig. 2. The impact of the constraints to the calculation of the coefficients is shown by the sizes of the multipliers, which in the case $r=4$ are smaller than those of the cases $r=3,5,6,7$. It seems that the 4-convex model in this $m=17$ experiment is the most

Table 2 The r-concave and the unconstrained lag coefficients

		r-Concave							Unconstrained
m	β_i	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	lag coefficients
	β_1	0.0334	0.0190	0.0703	0.0362	0.0683	0.0562	0.0709	0.0772
	β_2	0.0334	0.0209	0.0101	0.0277	0.0130	0.0258	0.0057	−0.0020
	β_3	0.0334	0.0227	0.0056	0.0223	0.0073	0.0110	0.0062	0.0015
	β_4	0.0334	0.0246	0.0118	0.0195	0.0084	0.0072	0.0156	0.0190
	β_5	0.0334	0.0264	0.0176	0.0190	0.0137	0.0108	0.0192	0.0348
	β_6	0.0334	0.0283	0.0228	0.0203	0.0210	0.0183	0.0199	−0.0025
	β_7	0.0334	0.0302	0.0274	0.0231	0.0286	0.0271	0.0231	0.0425
	β_8	0.0334	0.0320	0.0315	0.0269	0.0351	0.0352	0.0295	−0.0020
17	β_9	0.0334	0.0339	0.0351	0.0313	0.0397	0.0412	0.0375	0.0728
	β_{10}	0.0334	0.0358	0.0382	0.0360	0.0421	0.0442	0.0444	0.0391
	β_{11}	0.0334	0.0376	0.0407	0.0405	0.0424	0.0442	0.0477	0.0440
	β_{12}	0.0334	0.0395	0.0427	0.0443	0.0410	0.0418	0.0462	0.0203
	β_{13}	0.0334	0.0414	0.0442	0.0472	0.0389	0.0382	0.0407	0.0755
	β_{14}	0.0334	0.0432	0.0451	0.0487	0.0375	0.0354	0.0341	0.0241
	β_{15}	0.0334	0.0451	0.0455	0.0484	0.0387	0.0361	0.0315	0.0269
	β_{16}	0.0334	0.0469	0.0454	0.0458	0.0449	0.0439	0.0399	0.0281
	β_{17}	0.0334	0.0488	0.0447	0.0407	0.0588	0.0628	0.0677	0.0819
	β_1	0.0286	0.0186	0.0502	0.0049	0.0474	0.0366	0.0498	0.0527
	β_2	0.0286	0.0200	0.0136	0.0150	0.0116	0.0181	0.0059	−0.0009
	β_3	0.0286	0.0214	0.0159	0.0226	0.0043	0.0094	0.0020	−0.0034
	β_4	0.0286	0.0228	0.0181	0.0280	0.0055	0.0082	0.0090	0.0148
	β_5	0.0286	0.0241	0.0202	0.0315	0.0121	0.0124	0.0170	0.0341
	β_6	0.0286	0.0255	0.0221	0.0333	0.0213	0.0197	0.0243	0.0078
	β_7	0.0286	0.0269	0.0239	0.0337	0.0308	0.0283	0.0308	0.0367
	β_8	0.0286	0.0283	0.0255	0.0331	0.0391	0.0365	0.0362	0.0067
	β_9	0.0286	0.0297	0.0270	0.0317	0.0448	0.0429	0.0404	0.0778
	β_{10}	0.0286	0.0301	0.0284	0.0298	0.0470	0.0464	0.0428	0.0345
21	β_{11}	0.0286	0.0305	0.0296	0.0276	0.0456	0.0464	0.0432	0.0451
	β_{12}	0.0286	0.0309	0.0307	0.0256	0.0407	0.0427	0.0410	0.0267
	β_{13}	0.0286	0.0314	0.0317	0.0239	0.0329	0.0356	0.0361	0.0753
	β_{14}	0.0286	0.0318	0.0325	0.0229	0.0233	0.0259	0.0285	0.0038
	β_{15}	0.0286	0.0322	0.0331	0.0228	0.0137	0.0153	0.0190	−0.0072
	β_{16}	0.0286	0.0326	0.0336	0.0240	0.0060	0.0060	0.0092	0.0029
	β_{17}	0.0286	0.0330	0.0340	0.0266	0.0029	0.0008	0.0018	0.0445
	β_{18}	0.0286	0.0335	0.0343	0.0311	0.0075	0.0037	0.0013	0.0023
	β_{19}	0.0286	0.0339	0.0344	0.0376	0.0231	0.0192	0.0140	−0.0123
	β_{20}	0.0286	0.0343	0.0343	0.0466	0.0540	0.0529	0.0488	0.0442
	β_{21}	0.0286	0.0347	0.0342	0.0581	0.1045	0.1114	0.1175	0.1351

successful choice among the r-convex models in following the trend of the unconstrained coefficients. Similar results are obtained for $m = 21$, except that the method employs 21 coefficients instead of 17. In this case, it is clear the 4-convex model gives the best results. Now the 4-convex coefficients lie on a fit that consists of four overlapping cubics that follow smoothly the trend of the unconstrained coefficients all over the range, as it is also illustrated in Fig. 3.

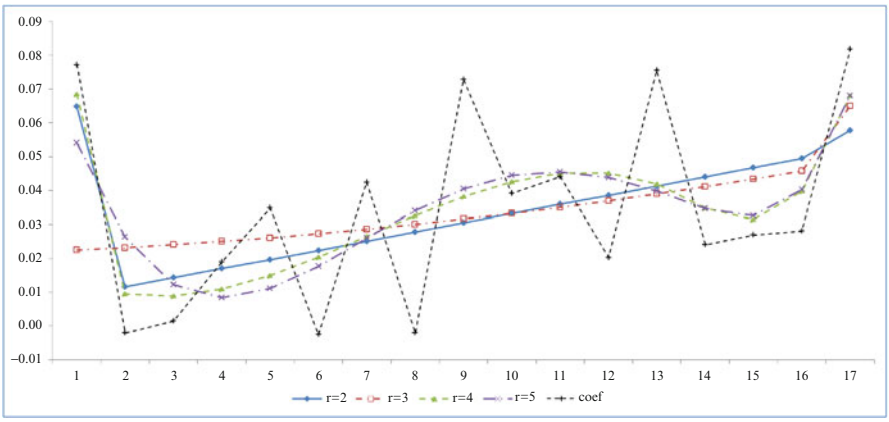


Fig. 2 The unconstrained (*plus sign*) and the r-convex lag coefficients of Table 1, when $m = 17$ and $r = 2, 3, 4$, and 5. A piecewise linear interpolant illustrates the associated coefficients

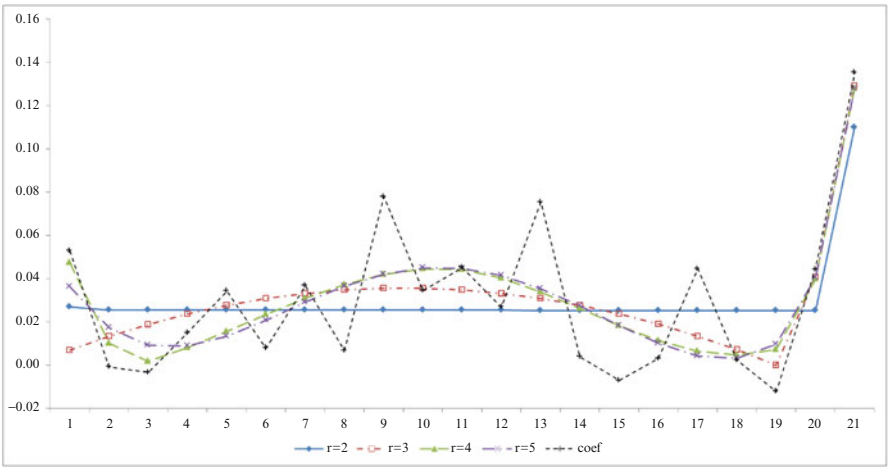


Fig. 3 As in Fig. 2, but $m = 21$

The r-concave coefficients for $r \leq 4$ of Table 2 seem unsuccessful in following this trend, especially at the tails of the unconstrained coefficients. The reason is that not only the approximation by nonpositive differences is not suitable when $r \leq 4$, but also the inactive constraints that occur at the r-concave fits are so rare that these fits do not allow more flexibility than that of a polynomial fit of degree $r - 1$. For example, all the components of the 1-concave model are equal, as opposed to the 1-convex model; the 2-concave model for $m = 17$ and $m = 21$, which is presented in Figs. 4 and 5, respectively, is a straight line fit as opposed to the 2-convex polygonal model; the 3-concave model, except at the left end of the range, gave poor results;

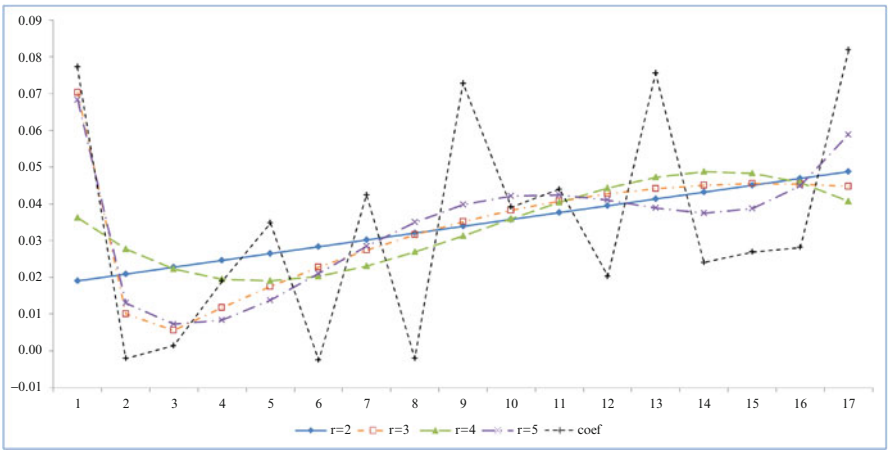


Fig. 4 The unconstrained (*plus sign*) and the r-concave lag coefficients of Table 2, when $m = 17$ and $r = 2, 3, 4$, and 5

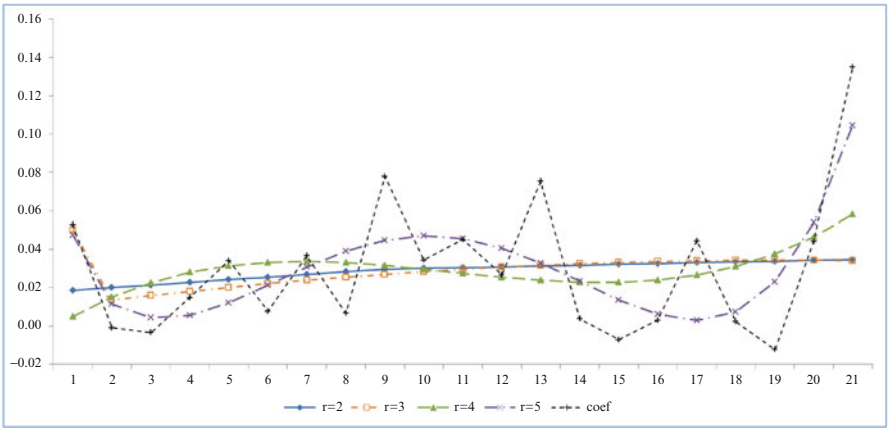


Fig. 5 As in Fig. 4, but $m = 21$

the 4-concave model failed to follow the trend of the unconstrained coefficients at the left half of the range and gave poor results for the right half of the range. However, the r-concave coefficients, for $r \geq 5$, are much closer to the components of $\underline{\tilde{\beta}}$ than those for $r \leq 4$, because the r-concave fit tends to undulate for larger values of r. Still, the values of the Lagrange multipliers are kept large.

Next, we applied the conjugate gradient-type method of Sect. 3 to the data described in the beginning of the section. Therefore we estimated the coefficients of the distributed-lag model with $m = 17, 21$ subject to the piecewise monotonicity constraints (22) on the components of $\underline{\beta}$ for $k = 1, 2, \ldots, 6$, where the first mono-

Table 3 The Lagrange multipliers that correspond to the lag coefficients of Table 1

		r-Convex						
m	λ_i	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
	λ_1	296.58	0	1,634.05	0	6,149.12	1,697.93	71,610.28
	λ_2	318.98	76.90	4,861.62	49.41	22,266.43	0	366,144.86
	λ_3	199.50	375.86	8,743.49	1,192.22	41,321.04	0	902,718.82
	λ_4	134.47	732.77	12,588.86	3,153.40	53,127.46	12,878.60	1,453,646.53
	λ_5	80.12	1,115.96	15,652.99	4,463.18	53,474.10	31,826.18	1,741,209.45
	λ_6	0	1,617.83	17,064.43	4,999.97	41,940.12	51,847.92	1,580,571.26
	λ_7	58.98	1,976.79	16,884.35	3,783.62	27,705.83	50,496.51	1,115,288.38
	λ_8	0	2,395.58	14,654.00	2,587.10	14,324.94	32,132.68	613,846.77
17	λ_9	82.03	2,471.50	11,244.74	1,665.53	4,473.71	15,393.00	249,044.41
	λ_{10}	132.08	2,243.72	7,483.21	803.68	0	8,517.42	52,882.99
	λ_{11}	150.42	1,783.05	4,018.59	0	0	2,494.36	–
	λ_{12}	178.29	1,230.80	1,465.29	0	86.08	–	–
	λ_{13}	221.67	580.09	300.39	23.29	–	–	–
	λ_{14}	123.17	162.89	0	–	–	–	–
	λ_{15}	34.46	0	–	–	–	–	–
	λ_{16}	0	–	–	–	–	–	–
	λ_{17}	–	–	–	–	–	–	–
	λ_1	178.66	0	1,520.79	6.36	5,617.82	1,063.93	112,337.53
	λ_2	173.80	367.40	4,787.05	0	24,097.74	0	654,489.39
	λ_3	47.26	1,327.38	8,885.78	941.59	53,651.32	15,605.84	1,902,073.32
	λ_4	0	2,673.06	12,999.04	2,957.46	85,725.13	90,771.01	3,723,747.68
	λ_5	0	4,210.62	16,378.73	4,740.05	113,230.14	252,693.56	5,502,455.15
	λ_6	16.09	5,800.62	18,315.51	5,103.94	133,710.17	479,660.33	6,420,797.25
	λ_7	119.44	7,201.96	18,557.30	3,094.40	151,648.22	678,318.38	6,071,823.25
	λ_8	172.27	8,462.07	16,536.60	1,044.23	163,510.59	763,000.22	4,624,195.67
	λ_9	477.13	9,046.76	13,048.99	0	164,431.73	694,166.97	2,721,109.06
	λ_{10}	754.75	8,988.84	8,830.08	657.58	149,367.38	501,655.02	1,110,351.23
21	λ_{11}	1,021.84	8,272.88	4,791.60	2,588.90	117,277.33	267,402.70	237,879.61
	λ_{12}	1,227.38	7,012.41	1,817.45	4,879.08	73,098.98	90,975.31	0
	λ_{13}	1,352.69	5,347.66	557.68	5,204.32	32,177.45	12,267.54	14,858.87
	λ_{14}	1,233.82	3,757.00	364.70	3,329.33	7,528.65	0	20,205.95
	λ_{15}	1,029.78	2,411.41	366.66	1,009.03	0	2,301.89	–
	λ_{16}	802.72	1,352.58	170.93	0	54.85	–	–
	λ_{17}	590.81	538.09	0	67.56	–	–	–
	λ_{18}	346.62	53.88	9.76	–	–	–	–
	λ_{19}	67.76	0	–	–	–	–	–
	λ_{20}	0	–	–	–	–	–	–
	λ_{21}	–	–	–	–	–	–	–

tonic section of the fit is increasing. Also we estimated coefficients, where we allowed the first monotonic section to be decreasing. The amount of CPU time to carry out these calculations in single precision arithmetic is negligible.

The tolerance for the termination criterion (28) in Step 4 was set to 10^{-6}. The actual values of m, k, and the calculated coefficients with the increasing option are given in Table 5, while the coefficients with the decreasing option are given in Table 6. The coefficients are shown in the third ($k=1$), fourth ($k=2$), and so on column of each table. The last column of each table displays the unconstrained

Table 4 The Lagrange multipliers that correspond to the lag coefficients of Table 2

		r-Concave						
m	λ_i	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
	λ_1	76.51	653.07	0	4,041.46	0	30,641.37	14,430.64
	λ_2	455.65	1,382.08	0	14,603.60	1,406.95	136,046.25	17,553.30
	λ_3	981.14	1,853.82	214.43	30,428.76	10,973.84	312,424.60	0
	λ_4	1,434.64	2,112.71	442.54	48,758.00	32,745.41	497,491.78	8,058.22
	λ_5	1,841.09	2,109.29	534.64	66,170.59	61,169.00	615,835.25	47,007.34
	λ_6	2,198.91	1,803.41	821.42	78,654.88	87,255.56	610,712.40	129,036.51
	λ_7	2,371.01	1,448.50	1,153.21	84,255.25	97,070.86	500,875.99	161,218.23
	λ_8	2,555.75	859.09	2,116.72	80,253.22	88,742.28	336,025.61	104,777.13
17	λ_9	2,439.73	469.75	3,140.85	67,388.58	67,084.79	174,025.68	27,566.27
	λ_{10}	2,251.48	269.84	3,772.34	48,926.32	40,249.24	60,824.27	3,022.83
	λ_{11}	2,013.19	209.83	3,721.74	29,293.46	16,598.38	11,490.81	–
	λ_{12}	1,725.58	289.84	3,092.94	12,901.62	3,755.66	–	–
	λ_{13}	1,406.44	456.18	1,838.81	3,260.96	–	–	–
	λ_{14}	1,169.72	419.65	632.85	–	–	–	–
	λ_{15}	873.55	217.14	–	–	–	–	–
	λ_{16}	486.74	–	–	–	–	–	–
	λ_{17}	–	–	–	–	–	–	–
	λ_1	59.00	456.66	0	7,036.82	0	33,927.71	10,655.22
	λ_2	376.54	881.22	449.63	28,537.02	2,079.81	172,036.14	0
	λ_3	820.74	997.01	2,263.46	67,499.85	16,793.68	456,105.85	0
	λ_4	1,173.21	934.50	5,733.88	122,916.97	55,911.40	854,672.23	237,572.29
	λ_5	1,452.29	760.16	10,782.88	189,891.40	120,597.74	1,279,736.53	1,037,984.91
	λ_6	1,677.55	494.11	17,341.48	259,823.15	199,140.55	1,632,523.87	2,570,050.48
	λ_7	1,775.89	293.92	24,858.56	322,698.03	268,188.21	1,868,054.30	4,469,092.13
	λ_8	1,904.41	0	33,258.33	366,002.78	313,905.99	1,960,403.04	6,013,750.04
	λ_9	1,745.24	96.85	41,043.27	381,826.35	329,370.07	1,899,263.75	6,510,991.70
	λ_{10}	1,554.97	559.44	46,832.45	367,461.26	315,230.01	1,686,421.42	5,704,141.80
21	λ_{11}	1,367.95	1,301.65	49,537.66	325,462.15	276,942.31	1,342,042.44	3,933,275.53
	λ_{12}	1,195.70	2,236.97	48,618.74	263,660.49	223,403.24	911,134.06	2,016,305.28
	λ_{13}	1,056.27	3,229.46	43,815.49	193,808.72	159,611.75	489,767.11	680,876.45
	λ_{14}	1,084.41	3,873.47	36,113.20	126,459.39	93,854.36	183,118.29	110,027.36
	λ_{15}	1,115.68	4,105.97	26,752.69	69,871.66	39,704.26	34,778.82	–
	λ_{16}	1,136.72	3,910.86	17,092.21	29,516.89	8,716.24	–	–
	λ_{17}	1,155.57	3,266.40	8,612.74	7,146.94	–	–	–
	λ_{18}	1,123.22	2,250.31	2,575.12	–	–	–	–
	λ_{19}	1,045.03	948.85	–	–	–	–	–
	λ_{20}	658.89	–	–	–	–	–	–
	λ_{21}	–	–	–	–	–	–	–

lag coefficients. The underlined numbers indicate the positions of the local extrema, maxima, and minima. The coefficients of Tables 5 and 6, for $k = 1, 2, 3, 4, 5$, and 6, for each m are displayed in Figs. 6, 7, 8, and 9.

The results of Table 5 are as follows. The monotonically increasing components when $(m = 17, 21;\ k = 1)$ consist of four sections of different equal components. The lag coefficients for $k = 1, 3, 5$, apart from slight differences in the fourth decimal place of sporadic values, are the same with the lag coefficients for $k = 2, 4, 6$, respectively. The user may specify whether the first monotonic section in (22) is in-

Table 5 The piecewise monotonic lag coefficients where the first monotonic section is increasing and the unconstrained lag coefficients

m	β_i	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Unconstrained lag coefficients
	β_1	0.0227	0.0227	0.0708	0.0708	0.0704	0.0704	0.0772
	β_2	0.0227	0.0227	0.0120	0.0120	0.0108	0.0108	−0.0020
	β_3	0.0227	0.0227	−0.0034	−0.0034	−0.0036	−0.0036	0.0015
	β_4	0.0227	0.0227	0.0197	0.0197	0.0200	0.0200	0.0190
	β_5	0.0227	0.0227	0.0197	0.0197	0.0200	0.0200	0.0348
	β_6	0.0227	0.0227	0.0197	0.0197	0.0200	0.0200	−0.0025
	β_7	0.0245	0.0245	0.0251	0.0251	0.0200	0.0200	0.0425
	β_8	0.0245	0.0245	0.0251	0.0251	0.0200	0.0200	−0.0020
17	β_9	0.0408	0.0408	0.0411	0.0411	0.0578	0.0578	0.0728
	β_{10}	0.0408	0.0408	0.0411	0.0411	0.0435	0.0435	0.0391
	β_{11}	0.0408	0.0408	0.0411	0.0411	0.0435	0.0435	0.0440
	β_{12}	0.0408	0.0408	0.0411	0.0411	0.0435	0.0435	0.0203
	β_{13}	0.0408	0.0408	0.0411	0.0411	0.0435	0.0435	0.0755
	β_{14}	0.0408	0.0408	0.0411	0.0411	0.0328	0.0328	0.0241
	β_{15}	0.0408	0.0408	0.0411	0.0411	0.0281	0.0281	0.0269
	β_{16}	0.0408	0.0408	0.0411	0.0411	0.0401	0.0401	0.0281
	β_{17}	0.0670	0.0670	0.0627	0.0627	0.0696	0.0696	0.0819
	β_1	0.0199	0.0199	0.0165	0.0165	0.0483	0.0483	0.0527
	β_2	0.0199	0.0199	0.0165	0.0165	0.0100	0.0100	−0.0009
	β_3	0.0199	0.0199	0.0165	0.0165	−0.0082	−0.0082	−0.0034
	β_4	0.0199	0.0199	0.0165	0.0165	0.0149	0.0149	0.0148
	β_5	0.0236	0.0235	0.0215	0.0215	0.0229	0.0229	0.0341
	β_6	0.0274	0.0274	0.0218	0.0218	0.0229	0.0229	0.0078
	β_7	0.0274	0.0274	0.0224	0.0224	0.0229	0.0229	0.0367
	β_8	0.0274	0.0274	0.0224	0.0224	0.0229	0.0229	0.0067
	β_9	0.0274	0.0274	0.0604	0.0604	0.0611	0.0611	0.0778
	β_{10}	0.0274	0.0274	0.0441	0.0441	0.0438	0.0438	0.0345
21	β_{11}	0.0274	0.0274	0.0441	0.0441	0.0438	0.0438	0.0451
	β_{12}	0.0274	0.0274	0.0441	0.0441	0.0438	0.0438	0.0267
	β_{13}	0.0274	0.0274	0.0441	0.0441	0.0438	0.0438	0.0753
	β_{14}	0.0274	0.0274	0.0116	0.0116	0.0118	0.0118	0.0038
	β_{15}	0.0274	0.0274	0.0116	0.0116	0.0118	0.0118	−0.0072
	β_{16}	0.0274	0.0274	0.0116	0.0116	0.0118	0.0118	0.0029
	β_{17}	0.0274	0.0274	0.0116	0.0116	0.0118	0.0118	0.0445
	β_{18}	0.0274	0.0274	0.0116	0.0116	0.0118	0.0118	0.0023
	β_{19}	0.0274	0.0274	−0.0081	−0.0081	−0.0087	−0.0087	−0.0123
	β_{20}	0.0274	0.0274	0.0458	0.0458	0.0484	0.0484	0.0442
	β_{21}	0.1033	0.1033	0.1330	0.1330	0.1286	0.1286	0.1351

creasing or decreasing, but the algorithm can give $\beta_2 < \beta_1$, as for example occurs in Table 5 for $k = 3,4,5,6$ when $m = 17$, by regarding the first monotonic component β_1 as the first monotonic section. Thus in Table 5, we have underlined the number β_1 for $k = 3,4,5,6$ when $m = 17$ and for $k = 5,6$ when $m = 21$. When $m = 17$, a minimum occurs at β_3 for $k = 3,4,5,6$, a maximum at β_9 for $k = 5,6$, and a minimum at β_{15} for $k = 5,6$. It is noticeable that each fit preserves the positions of the extrema as k increases. Similar results are observed when $m = 21$. We see also that

Table 6 The piecewise monotonic lag coefficients where the first monotonic section is decreasing and the unconstrained lag coefficients

m	β_i	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Unconstrained lag coefficients
	β_1	0.0335	0.0708	0.0708	0.0704	0.0704	0.0699	0.0772
	β_2	0.0335	0.0120	0.0120	0.0108	0.0108	0.0103	−0.0020
	β_3	0.0335	−0.0035	−0.0034	−0.0036	−0.0036	−0.0008	0.0015
	β_4	0.0335	0.0197	0.0197	0.0200	0.0200	0.0179	0.0190
	β_5	0.0335	0.0197	0.0197	0.0200	0.0200	0.0179	0.0348
	β_6	0.0335	0.0197	0.0197	0.0200	0.0200	0.0179	−0.0025
	β_7	0.0335	0.0251	0.0251	0.0200	0.0200	0.0392	0.0425
	β_8	0.0335	0.0251	0.0251	0.0200	0.0200	0.0004	−0.0020
17	β_9	0.0335	0.0411	0.0411	0.0578	0.0578	0.0618	0.0728
	β_{10}	0.0335	0.0411	0.0411	0.0435	0.0435	0.0438	0.0391
	β_{11}	0.0335	0.0411	0.0411	0.0435	0.0435	0.0438	0.0440
	β_{12}	0.0335	0.0411	0.0411	0.0435	0.0435	0.0438	0.0203
	β_{13}	0.0335	0.0411	0.0411	0.0435	0.0435	0.0438	0.0755
	β_{14}	0.0335	0.0411	0.0411	0.0328	0.0328	0.0322	0.0241
	β_{15}	0.0335	0.0411	0.0411	0.0281	0.0281	0.0285	0.0269
	β_{16}	0.0335	0.0411	0.0411	0.0401	0.0401	0.0394	0.0281
	β_{17}	0.0335	0.0628	0.0627	0.0696	0.0696	0.0705	0.0819
	β_1	0.0286	0.0528	0.0528	0.0483	0.0483	0.0480	0.0527
	β_2	0.0286	0.0125	0.0125	0.0100	0.0100	0.0101	−0.0009
	β_3	0.0286	−0.0073	−0.0073	−0.0082	−0.0082	−0.0070	−0.0034
	β_4	0.0286	0.0197	0.0198	0.0149	0.0149	0.0119	0.0148
	β_5	0.0286	0.0275	0.0275	0.0229	0.0229	0.0244	0.0341
	β_6	0.0286	0.0275	0.0275	0.0229	0.0229	0.0244	0.0078
	β_7	0.0286	0.0275	0.0275	0.0229	0.0229	0.0318	0.0367
	β_8	0.0286	0.0275	0.0275	0.0229	0.0229	0.0076	0.0067
	β_9	0.0286	0.0275	0.0275	0.0611	0.0611	0.0660	0.0778
	β_{10}	0.0286	0.0275	0.0275	0.0438	0.0438	0.0439	0.0345
21	β_{11}	0.0286	0.0275	0.0275	0.0438	0.0438	0.0439	0.0451
	β_{12}	0.0286	0.0275	0.0275	0.0438	0.0438	0.0439	0.0267
	β_{13}	0.0286	0.0275	0.0275	0.0438	0.0438	0.0439	0.0753
	β_{14}	0.0286	0.0275	0.0275	0.0118	0.0118	0.0115	0.0038
	β_{15}	0.0286	0.0275	0.0275	0.0118	0.0118	0.0115	−0.0072
	β_{16}	0.0286	0.0275	0.0275	0.0118	0.0118	0.0115	0.0029
	β_{17}	0.0286	0.0275	0.0275	0.0118	0.0118	0.0115	0.0445
	β_{18}	0.0286	0.0275	0.0275	0.0118	0.0118	0.0115	0.0023
	β_{19}	0.0286	0.0275	0.0275	−0.0087	−0.0087	−0.0050	−0.0123
	β_{20}	0.0286	0.0275	0.0275	0.0484	0.0484	0.0468	0.0442
	β_{21}	0.0286	0.0999	0.0999	0.1286	0.1286	0.1280	0.1351

the last monotonic section for $(m = 17, 21;\ k = 2, 4, 6)$ has degenerated to the last component β_m.

The results of Table 6 are as follows. The monotonically decreasing components when $(m = 17, 21;\ k = 1)$ are all equal, which indicates that this model is less successful than the corresponding model of Table 5. The lag coefficients for $k = 2, 4$, apart from slight differences, are the same with the lag coefficients for $k = 3, 5$, respectively. When $m = 17$, the coefficients have a minimum at β_3 for $k = 2, 3, 4, 5, 6$, a maximum at β_9 and a minimum at β_{15} for $k = 4, 5, 6$, and a maximum at β_7 as well

as a minimum at β_8 for $k=6$. Similar results are observed when $m=21$. We see also that the last monotonic section when $(m=17,21;\ k=3,5)$ has degenerated to β_m.

It is remarkable that the extrema $\{\beta_{t_j}:\ j=1,2,\ldots,k-1\}$ of the piecewise monotonic coefficient estimates approach $k-1$ out of the m unconstrained coefficients. If a suitable value of k is not known in advance, then the user may apply the piecewise monotonicity constraints that are incorporated in the conjugate gradient algorithm for a sequence of integers k. On the other hand, it would have been sensible to give k a value, after there had been derived some information from the unconstrained lag coefficients. Further, we should note that the cases $(m=17;k=2,4)$ and $(m=21,k=4)$ of Table 6, apart from slight differences, present the same results as the cases $(m=17;k=3,5)$ and $(m=21,k=5)$ of Table 5, respectively. This remark suggests that the piecewise monotonic constraints can be employed either with the increasing or with the decreasing option for the first monotonic section, because as k increases the piecewise monotonicity algorithm through the course may allow a monotonic section to degenerate to a single component, which in turn can remedy an initially unsuccessful choice of the first monotonic section.

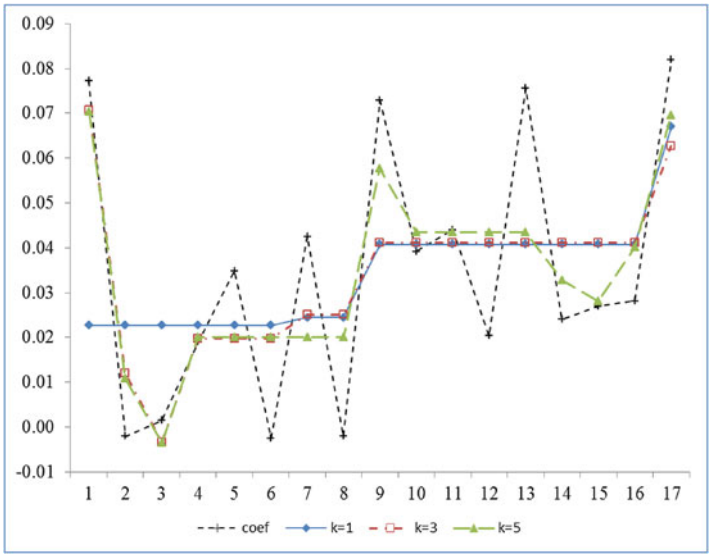


Fig. 6 With $m=17$, the unconstrained (*plus sign*) and the piecewise monotonic lag coefficients (first section is increasing) with $k=1(2)$, $k=3(4)$, and $k=5(6)$ of Table 5

A comparison of the $k=1$ case of Tables 5 and 6 with the $r=1$ case of Tables 1 and 2 respectively shows that the monotonic coefficients of Tables 5 and 6, obtained by the conjugate gradient type method of Sect. 3 when $k=1$, are quite close to the monotonic coefficients of Tables 1 and 2 obtained by the method of Sect. 2 when $r=1$.

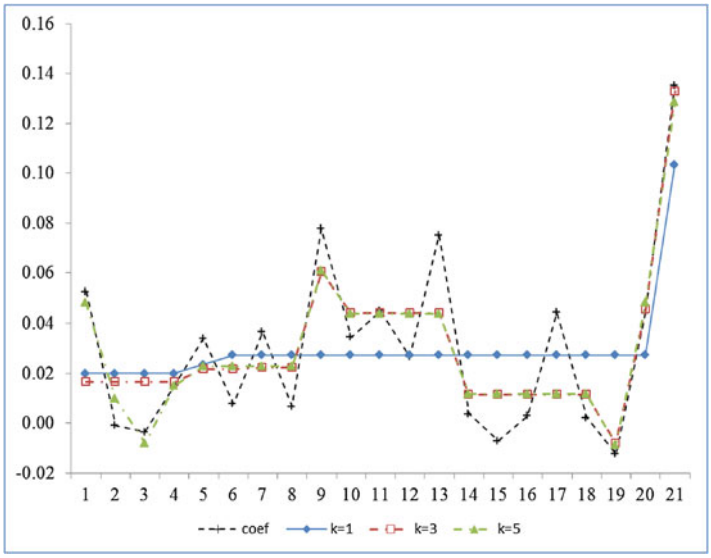


Fig. 7 With $m = 21$, the unconstrained (*plus sign*) and the piecewise monotonic lag coefficients (first section is increasing) with $k = 1(2)$, $k = 3(4)$, and $k = 5(6)$ of Table 5

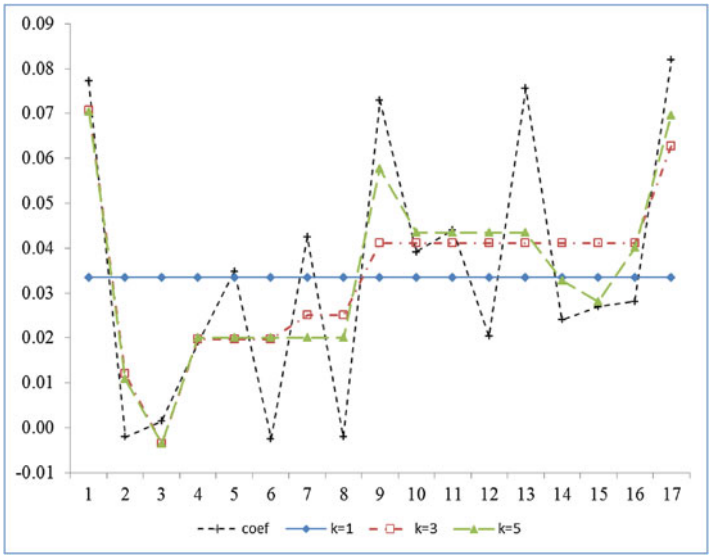


Fig. 8 With $m = 17$, the unconstrained (*plus sign*) and the piecewise monotonic lag coefficients (first section is decreasing) with $k = 1, 3(2)$ and $k = 5(4)$ of Table 6

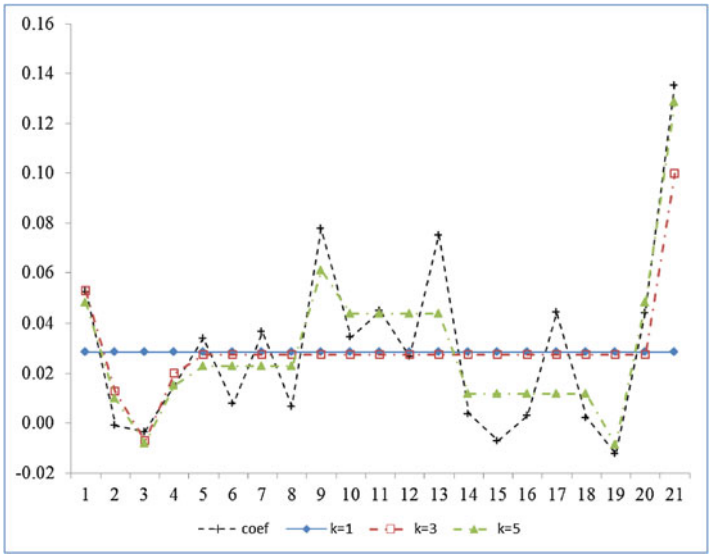


Fig. 9 With $m = 21$, the unconstrained (*plus sign*) and the piecewise monotonic lag coefficients (first section is decreasing) with $k = 1, 3(2)$ and $k = 5(4)$ of Table 6

5 Discussion

We have considered two methods for calculating distributed-lag coefficients subject to sign conditions on consecutive differences of the coefficient estimates and have presented an application of these methods to real macroeconomic data on money and prices that gives attention to particular relations for the lag coefficients. Further, we have included some relations that suggest conflicting strategies, in order to shed light to the behavior of these methods and help decide for suitable relations.

The first method is a strictly convex quadratic programming calculation subject to nonnegative differences of order r of the lag coefficients. The method is efficient computationally, because the matrices that occur are banded and positive definite due to the Toeplitz structure of the constraint functions. Two important practical questions concern the size of m and the choice of r. The size of m can be selected by statistical methods. The choice of order r should depend on properties of the underlying relation. If, however, the choice of r is a matter of experimentation, the user may try iteratively some values of r, as suggested by Cullinan [2] and Vassiliou and Demetriou [34]. Some modeling advantages of using this method are that it achieves a rather weak representation of the lag coefficients, it obtains well recognized structures which take the form of monotonicity, convexity, and r-convexity for $r \geq 3$, and it provides estimations of higher rates of change of the underlying relation. In the example considered, the r-convex lag coefficients were able to follow the general trend of the (unknown) unconstrained lag coefficients in a rather smooth manner.

The second method estimates piecewise monotonic lag coefficients by an iterative procedure that combines the conjugate gradient method with a piecewise monotonicity data approximation method. The procedure is both efficient computationally and competent to its modeling task. In addition, it overcomes the multicollinearity problem that frequently occurs in the practice of distributed-lag calculations. The obtained piecewise monotonicity model provides a weak, nonetheless very realistic representation of the lag coefficients, a property that is highly desirable by the experts. Moreover, the choice of the prior knowledge parameter k gives the estimation of the lag coefficients valuable flexibility.

The authors have developed Fortran versions of the algorithms that would be helpful for obtaining particular relations in real problem applications. Furthermore, there is room for much empirical analysis as well as for comparisons with other distributed-lag methods. Our methods may be useful, because they are driven by properties such as r-convexity and piecewise monotonicity that allow a wide range of assumptions about the lags. Besides, these properties do not occur in other distributed-lag methods.

Acknowledgements This work was partially supported by the University of Athens under Research Grant 11105.

References

1. Almon, S.: The distributed lag between capital appropriations and expenditures. Econometrica **33**(1), 178–196 (1965)
2. Cullinan, M.P.: Data smoothing using non-negative divided differences and ℓ_2 approximation. IMA J. Numer. Anal. **10**, 583–608 (1990)
3. Cullinan, M.P., Powell, M.J.D.: Data smoothing by divided differences. In: Watson, G.A. (ed.) Proceedings of the Dundee Conference on Numerical Analysis, 1981. LNIM, No. 912, pp. 26–37. Springer, Berlin (1982)
4. DeLeeuw, F.: The demand for capital goods by manufacturers: a study of quarterly time series. Econometrica **30**, 407–423 (1962)
5. Demetriou, I.C.: Discrete piecewise monotonic approximation by a strictly convex distance function. Math. Comput. **64**(209), 157–180 (1995)
6. Demetriou, I.C.: Algorithm 863: L2WPMA, a fortran 77 package for weighted least squares piecewise monotonic data approximation. ACM Trans. Math. Softw. **33**(1), 1–19 (2007)
7. Demetriou, I.C., Lipitakis, E.A.: Certain positive definite submatrices that arise from binomial coefficient matrices. Appl. Numer. Math. **36**, 219–229 (2001)
8. Demetriou, I.C., Powell, M.J.D.: Least squares smoothing of univariate data to achieve piecewise monotonicity. IMA J. Numer. Anal. **11**, 411–432 (1991)
9. Demetriou, I.C., Powell, M.J.D.: The minimum sum of squares change to univariate data that gives convexity. IMA J. Numer. Anal. **11**, 433–448 (1991)
10. Demetriou, I.C., Vassiliou, E.E.: A distributed lag estimator with piecewise monotonic coefficients. In: Ao, S.I., Gelman, L., Hukins, D.W.L., Hunter, A., Korsunsky, A.M. (eds.) Proceedings of the World Congress on Engineering, London, 2–4 July 2008, International Association of Engineers, vol. II, pp. 1088–1094 (2008)
11. Demetriou, I.C., Vassiliou, E.E.: An algorithm for distributed lag estimation subject to piecewise monotonic coefficients. IAENG Int. J. Appl. Math. **39**(1), 82–91 (2009)

12. Dhrymes, P.J., Klein, L.R., Steiglitz, K.: Estimation of distributed lags. Int. Econ. Rev. **11**(2), 235–250 (1970)
13. Fisher, I.: Note on a short-cut method for calculating distributed lags. Bulletin de l'Institut International de Statistique. **29**, 323–328 (1937)
14. Fletcher, R.: Practical Methods of Optimization, 2nd edn. Wiley, Chichester (2003)
15. Fletcher, R., Reeves, C.M.: Function minimization by conjugate gradients. Comput. J. **7**, 149–154 (1964)
16. Goldstein, A.A.: Constructive Real Analysis. Harper, New York (1967)
17. Golub, G., van Loan, C.F.: Matrix Computations, 2nd edn. The John Hopkins University Press, Baltimore/London (1989)
18. Hannan, E.J.: The estimation of relations involving distributed lags. Econometrica **33**(1), 206–224 (1965)
19. Hildebrand, F.B.: Introduction to Numerical Analysis, 2nd edn. Dover Publication, New York (1974)
20. Hildreth, C.: Point estimates of ordinates of concave functions. J. Am. Stat. Assoc. **49**, 598–619 (1954)
21. Jorgenson, D.: Rational distributed lag functions. Econometrica **34**(1), 135–149 (1966)
22. Kailath, T. (ed.): Linear Least-Squares Estimation. Benchmark Papers in Electrical Engineering and Computer Science, vol. 17, pp. 10–45. Dowden, Hutchinson and Ross, Inc., Stroudsburg (1977)
23. Karlin, S.: Total Positivity, vol. 1. Stanford University Press, Stanford (1968)
24. Karlson, K.M.: The lag from money to prices. Report Federal Reserve Bank of St Louis (October 1980)
25. Katsaggelos, A.K.: Iterative image restoration algorithms. Opt. Eng. **28**(7), 735–748 (1989)
26. Koyck, L.: Distributed Lags and Investment Analysis. North-Holland, Amsterdam (1954)
27. Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. SIAM Publications, Philadelphia (1995)
28. Levinson, N.: The wiener RMS error criterion in filter design and prediction. J. Math. Phys. **25**, 261–278 (1947)
29. Maddala, G.S., Lahiri, K.: Introduction to Econometrics, 4th edn. Wiley, Chichester (2010)
30. Powell, M.J.D.: Convergence properties of algorithms for nonlinear optimization. SIAM Rev. **28**, 487–500 (1986)
31. Robertson, T., Wright, F.T., Dykstra, R.L.: Order Restricted Statistical Inference. Wiley, Chichester (1988)
32. Shiller, R.J.: A distributed lag estimator derived from smoothness priors. Econometrica **41**(4), 775–788 (1973)
33. Solow, R.M.: On a family of lag distributions. Econometrica **28**(2), 393–406 (1960)
34. Vassiliou, E.E., Demetriou, I.C.: A linearly distributed lag estimator with *r*-convex coefficients. Comput. Stat. Data Anal. **54**(11), 2836–2849 (2010)

Poincaré-Type Inequalities for Green's Operator on Harmonic Forms

Shusen Ding and Yuming Xing

1 Introduction

The purpose of this paper is to derive the Poincaré-type inequalities with unbounded factors for Green's operator applied to the solutions of the nonlinear elliptic differential equation $d^{\star}A(x,du) = B(x,du)$, which is called the nonhomogeneous A-harmonic equation for differential forms in $\mathbb{R}^n$, $n \geq 2$, where A and B are operators satisfying certain conditions. Furthermore, we prove both local and global Poincaré inequalities with Orlicz norms for Green's operator applied to differential forms in $L^{\varphi}(m)$-averaging domains. Our new results are extensions of L^p norm inequalities for Green's operator and can be used to estimate the norms of differential forms or the norms of other operators, such as the projection operator. The Poincaré-type inequalities have been widely studied and used in PDEs, analysis, and the related areas, and different versions of the Poincaré-type inequalities have been established during the recent years, see [1, 4–6, 8–12]. We all know that Green's operator is one of the key operators which is widely used in many areas, such as analysis and PDEs. The study of the above equation just started in recent years, see [1, 6, 8, 16]. However, the investigation of the homogeneous A-harmonic equation has been well developed and many applications in the related fields, including potential theory and nonlinear elasticity, have been found, see [13–15, 20–23]. In many situations, we often need to evaluate the integrals with unbounded factors. For instance, if the object P_1 with mass m_1 is located at the origin and the object P_2 with mass m_2 is located at (x,y,z) in $\mathbb{R}^3$, then, Newton's Law of Gravitation states that the magnitude of the gravitational force between two objects P_1 and P_2 is $|\mathbf{F}| = m_1 m_2 G/d^2(P_1,P_2)$,

S. Ding (✉)
Department of Mathematics, Seattle University, Seattle, WA 98122, USA
e-mail: sding@seattleu.edu

Y. Xing
Department of Mathematics, Harbin Institute of Technology, Harbin 150001, People's Republic of China
e-mail: xyuming@hit.edu.cn

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_7,

where $d(P_1,P_2)=\sqrt{x^2+y^2+z^2}$ is the distance between P_1 and P_2, and G is the gravitational constant. Hence, we need to deal with an integral whenever the integrand contains $|\mathbf{F}|$ as a factor and the integral domain includes the origin. Moreover, in calculating an electric field, we will evaluate the integral $E(y)=\frac{1}{4\pi\varepsilon_0}\int_D \rho(x)\frac{y-x}{\|y-x\|^3}dx$, where $\rho(x)$ is a charge density and x is the integral variable. The integrand is unbounded if $y\in D$. This is our motivation to prove the Poincaré-type inequalities for Green's operator G with unbounded factors.

In this paper, we always assume that M is a bounded, convex domain and B is a ball in $\mathbb{R}^n$, $n\geq 2$. Let σB be the ball with the same center as B and with $\mathrm{diam}(\sigma B)=\sigma\mathrm{diam}(B)$, $\sigma>0$. We do not distinguish the balls from cubes in this paper. We use $|E|$ to denote the Lebesgue measure of the set E. We say w is a weight if $w\in L^1_{\mathrm{loc}}(\mathbb{R}^n)$ and $w>0$ a.e. Differential forms are extensions of functions in $\mathbb{R}^n$. For example, the function $u(x_1,x_2,\ldots,x_n)$ is called a 0-form. A differential k-form $u(x)$ is generated by $\{dx_{i_1}\wedge dx_{i_2}\wedge\cdots dx_{i_k}\}$, $k=1,2,\ldots,n$, that is, $u(x)=\sum_I u_I(x)dx_I=\sum u_{i_1i_2\cdots i_k}(x)dx_{i_1}\wedge dx_{i_2}\wedge\cdots\wedge dx_{i_k}$, where $I=(i_1,i_2,\ldots,i_k)$, $1\leq i_1<i_2<\cdots<i_k\leq n$. Let $\wedge^l=\wedge^l(\mathbb{R}^n)$ be the set of all l-forms in $\mathbb{R}^n$, $D'(M,\wedge^l)$ be the space of all differential l-forms on M, and $L^p(M,\wedge^l)$ be the l-forms $u(x)=\sum_I u_I(x)dx_I$ on M satisfying $\int_M|u_I|^p<\infty$ for all ordered l-tuples I, $l=1,2,\ldots,n$. We denote the exterior derivative by $d:D'(M,\wedge^l)\to D'(M,\wedge^{l+1})$ for $l=0,1,\ldots,n-1$, and define the Hodge star operator $\star:\wedge^k\to\wedge^{n-k}$ as follows. If $u=u_{i_1i_2\cdots i_k}(x_1,x_2,\ldots,x_n)dx_{i_1}\wedge dx_{i_2}\wedge\cdots\wedge dx_{i_k}=u_I dx_I$, $i_1<i_2<\cdots<i_k$ is a differential k-form, then $\star u=\star(u_{i_1i_2\cdots i_k}dx_{i_1}\wedge dx_{i_2}\wedge\cdots\wedge dx_{i_k})=(-1)^{\Sigma(I)}u_I dx_J$, where $I=(i_1,i_2,\ldots,i_k)$, $J=\{1,2,\ldots,n\}-I$, and $\Sigma(I)=\frac{k(k+1)}{2}+\sum_{j=1}^k i_j$. The Hodge codifferential operator $d^\star:D'(M,\wedge^{l+1})\to D'(M,\wedge^l)$ is given by $d^\star=(-1)^{nl+1}\star d\star$ on $D'(M,\wedge^{l+1})$, $l=0,1,\ldots,n-1$. We write $\|u\|_{s,M}=(\int_M|u|^s)^{1/s}$ and $||u||_{s,M,w}=(\int_M|u|^s w(x)dx)^{1/s}$, where $w(x)$ is a weight. Let $\mathcal{W}(\wedge^l\Omega)=\{u\in L^1_{\mathrm{loc}}(\wedge^l\Omega):u \text{ has generalized gradient}\}$. As usual, the harmonic l-field is defined by $\mathcal{H}(\wedge^l\Omega)=\{u\in\mathcal{W}(\wedge^l\Omega):du=d^\star u=0,u\in L^p \text{ for some } 1<p<\infty\}$. The orthogonal complement of $\mathcal{H}$ in L^1 is defined by $\mathcal{H}^\perp=\{u\in L^1:<u,h>=0 \text{ for all } h\in\mathcal{H}\}$. Green's operator G is defined as $G:C^\infty(\wedge^l\Omega)\to\mathcal{H}^\perp\cap C^\infty(\wedge^l\Omega)$ by assigning $G(u)$ be the unique element of $\mathcal{H}^\perp\cap C^\infty(\wedge^l\Omega)$ satisfying Poisson's equation $\Delta G(u)=u-H(u)$, where H is either the harmonic projection or sometimes the harmonic part of u and Δ is the Laplace–Beltrami operator, see [18, 21] for more properties of Green's operator. We always use G to denote Green's operator in this paper.

We consider the nonhomogeneous A-harmonic equation for differential forms

$$d^\star A(x,du)=B(x,du), \tag{1}$$

where the mappings $A(x,\xi):M\times\wedge^l(\mathbb{R}^n)\to\wedge^l(\mathbb{R}^n)$ and $B(x,\xi):M\times\wedge^l(\mathbb{R}^n)\to\wedge^{l-1}(\mathbb{R}^n)$ are measurable with respect to x and ξ, and satisfy the conditions:

$$|A(x,\xi)|\leq a|\xi|^{p-1},\quad A(x,\xi)\cdot\xi\geq|\xi|^p \text{ and } |B(x,\xi)|\leq b|\xi|^{p-1} \tag{2}$$

for almost every $x\in M$ and all $\xi\in\wedge^l(\mathbb{R}^n)$, where $a,b>0$ are constants and $1<p<\infty$ is a fixed exponent associated with (1). A solution to (1) is an element of

the Sobolev space $W^{1,p}_{\mathrm{loc}}(M,\wedge^{l-1})$ such that $\int_M A(x,du)\cdot d\varphi + B(x,du)\cdot\varphi = 0$ for all $\varphi \in W^{1,p}_{\mathrm{loc}}(M,\wedge^{l-1})$ with compact support. Let $A : M\times\wedge^l(\mathbb{R}^n)\to\wedge^l(\mathbb{R}^n)$ be defined by $A(x,\xi) = \xi|\xi|^{p-2}$ with $p>1$. Then, A satisfies the required conditions and $d^*A(x,du) = 0$ reduces to the p-harmonic equation

$$d^*(du|du|^{p-2}) = 0 \tag{3}$$

for differential forms. In case that u is a function (0-form), (3) becomes the usual p-harmonic equation $\mathrm{div}(\nabla u|\nabla u|^{p-2}) = 0$. A differential form u is called a harmonic form if u satisfies some version of the A-harmonic equation. Much progress has been made recently in the investigation of different versions of the A-harmonic equations, see [5, 6, 8, 13, 15–17].

The operator K_y with the case $y = 0$ was first introduced by Cartan in [3]. Then, it was extended to the following version in [14]. To each $y\in M$ there corresponds a linear operator $K_y : C^\infty(M,\wedge^l)\to C^\infty(M,\wedge^{l-1})$ defined by $(K_yu)(x;\xi_1,\dots,\xi_{l-1}) = \int_0^1 t^{l-1}u(tx+y-ty;x-y,\xi_1,\dots,\xi_{l-1})dt$. A homotopy operator $T : C^\infty(M,\wedge^l)\to C^\infty(M,\wedge^{l-1})$ is defined by $Tu = \int_M\phi(y)K_yudy$, where $\phi\in C_0^\infty(M)$ is normalized so that $\int_M\phi(y)dy = 1$. The l-form $u_M\in D'(M,\wedge^l)$ is defined by

$$u_M = |M|^{-1}\int_M u(y)dy,\ l=0,\ \text{ and }\ u_M = d(Tu),\ l=1,2,\dots,n \tag{4}$$

for all $u\in L^p(M,\wedge^l)$, $1\le p<\infty$. Furthermore, we have

$$u = d(Tu)+T(du), \tag{5}$$

$$\|Tu\|_{s,M}\le C(s,n,M)\mathrm{diam}(M)\|u\|_{s,M} \tag{6}$$

for any differential form u.

2 Local Inequalities

We first introduce the following lemmas that will be used in this paper.

Lemma 1 ([8]). *Let u be a solution to the nonhomogeneous A-harmonic equation (1) in M and let $\sigma>1$, $0<s,t<\infty$. Then, there exists a constant C, depending only on σ,n,a,b,s, and t, such that*

$$\|du\|_{s,B}\le C(n,M)|B|^{(t-s)/st}\|du\|_{t,\sigma B}$$

for all balls or cubes B with $\sigma B\subset M$.

Using the same method developed in the proof of Propositions 5.15 and 5.17 in [18], we can prove that for any closed ball $\overline{B} = B\cup\partial B$, we have

$$\|dd^*G(u)\|_{s,\overline{B}}+\|d^*dG(u)\|_{s,\overline{B}}+\|dG(u)\|_{s,\overline{B}}+\|d^*G(u)\|_{s,\overline{B}}+\|G(u)\|_{s,\overline{B}}\le C(s)\|u\|_{s,\overline{B}}. \tag{7}$$

Note that for any Lebesgue measurable function f defined on a Lebesgue measurable set E with $|E| = 0$, we have $\int_E f dx = 0$. Thus, $\|u\|_{s,\partial B} = 0$ and $\|dd^*G(u)\|_{s,\partial B} + \|d^*dG(u)\|_{s,\partial B} + \|dG(u)\|_{s,\partial B} + \|d^*G(u)\|_{s,\partial B} + \|G(u)\|_{s,\partial B} = 0$ since $|\partial B| = 0$. Therefore, we obtain

$$\begin{aligned}
&\|dd^*G(u)\|_{s,B} + \|d^*dG(u)\|_{s,B} + \|dG(u)\|_{s,B} + \|d^*G(u)\|_{s,MB} + \|G(u)\|_{s,B} \\
&= \|dd^*G(u)\|_{s,\overline{B}} + \|d^*dG(u)\|_{s,\overline{B}} + \|dG(u)\|_{s,\overline{B}} + \|d^*G(u)\|_{s,\overline{B}} + \|G(u)\|_{s,\overline{B}} \\
&\leq C(s)\|u\|_{s,\overline{B}} \\
&= C(s)\|u\|_{s,B}.
\end{aligned}$$

Hence, we have the following lemma.

Lemma 2. *Let u be a smooth differential form defined in M and $1 < s < \infty$. Then, there exists a positive constant $C = C(s)$, independent of u, such that*

$$\|dd^*G(u)\|_{s,B} + \|d^*dG(u)\|_{s,B} + \|dG(u)\|_{s,B} + \|d^*G(u)\|_{s,MB} + \|G(u)\|_{s,B} \leq C(s)\|u\|_{s,B}$$

for any ball $B \subset M$.

Now, we prove the following local elementary Poincaré-type inequality for Green's operator applied to differential forms.

Lemma 3. *Let $u \in L^s_{\mathrm{loc}}(M, \wedge^l)$, $l = 1, 2, \ldots, n$, $1 < s < \infty$ and G be Green's operator. Then, there exists a constant $C = C(n, s, M)$, independent of u, such that*

$$\|G(u) - (G(u))_B\|_{s,B} \leq C(n, s, M)\mathrm{diam}(B)\|du\|_{s,B}$$

for all balls B with $B \subset M$.

Proof. Replacing u by $G(u)$ in (5), then using (6) over a ball B, we obtain

$$\|G(u) - (G(u))_B\|_{s,B} = \|T(d(G(u)))\|_{s,B} \leq C_1(s, n, \Omega)\mathrm{diam}(B)\|d(G(u))\|_{s,B} \quad (8)$$

for any differential form u. Since G commutes with d (see [13]), using Lemma 2, we have

$$\|d(G(u))\|_{s,B} = \|G(d(u))\|_{s,B} \leq C_2(s)\|du\|_{s,B}. \quad (9)$$

Combining (8) and (9), we have

$$\|G(u) - (G(u))_B\|_{s,B} \leq C_3(s, n, \Omega)\mathrm{diam}(B)\|du\|_{s,B}.$$

we have completed the proof of Lemma 3.

Theorem 1. *Let $du \in L^s_{\mathrm{loc}}(\Omega, \wedge^l)$, $l = 1, 2, \ldots, n$, $1 < s < \infty$, be a solution of the non-homogeneous A-harmonic equation in a bounded domain Ω, G be Green's operator. Then, there exists a constant C, independent of u, such that*

$$\left(\int_B |G(u) - (G(u))_B|^s \frac{1}{|x - x_B|^\alpha} dx\right)^{1/s} \leq C(n, s, \alpha, \lambda, \Omega)|B|^\gamma \left(\int_{\sigma B} |du|^s \frac{1}{|x - x_B|^\lambda} dx\right)^{1/s} \quad (10)$$

for all balls B with $\sigma B \subset \Omega$ and any real numbers α and λ with $\alpha > \lambda \geq 0$, where $\gamma = \frac{1}{n} - \frac{\alpha-\lambda}{ns}$ and x_B is the center of ball B and $\sigma > 1$ is a constant.

Proof. Choose $\varepsilon \in (0,1)$ such that $\varepsilon n < \alpha - \lambda$ and let $B \subset \Omega$ be any ball with center x_B and radius r_B. Set $t = s/(1-\varepsilon)$, then, $t > s$. Write $\beta = t/(t-s)$, from the Hölder inequality and Lemma 3, we have

$$\begin{aligned}
&\Big(\int_B \Big(|G(u) - (G(u))_B|\Big)^s \frac{1}{|x-x_B|^\alpha} dx\Big)^{1/s} \\
&= \Big(\int_B \Big(|G(u) - (G(u))_B| \frac{1}{|x-x_B|^{\alpha/s}}\Big)^s dx\Big)^{1/s} \\
&\leq \|G(u) - (G(u))_B\|_{t,B} \Big(\int_B \Big(\frac{1}{|x-x_B|}\Big)^{t\alpha/(t-s)} dx\Big)^{(t-s)/st} \\
&= \|G(u) - (G(u))_B\|_{t,B} \Big(\int_B |x-x_B|^{-\alpha\beta} dx\Big)^{1/\beta s} \\
&\leq C_1(n,s,\Omega)\mathrm{diam}(B)\|du\|_{t,\nu B} \||x-x_B|^{-\alpha}\|_{\beta,B}^{1/s},
\end{aligned} \tag{11}$$

where $\nu > 1$ is a constant. We may assume that $x_B = 0$. Otherwise, we just move the center of the ball to the origin by a simple transformation. Therefore, for any $x \in B$, $|x-x_B| \geq |x| - |x_B| = |x|$. Using the polar coordinate substitution, we have

$$\int_B |x-x_B|^{-\alpha\beta} dx \leq C_2(n,s,\alpha) \int_0^{r_B} \rho^{-\alpha\beta} \rho^{n-1} d\rho \leq \frac{C_2(n,s,\alpha)}{n-\alpha\beta} (r_B)^{n-\alpha\beta}. \tag{12}$$

Select $m = nst/(ns + \alpha t - \lambda t)$, then $0 < m < s$. By the reverse Hölder inequality (Lemma 1), we find that

$$\|du\|_{t,\nu B} \leq C_3(n,s,\alpha,\lambda,\Omega)|B|^{\frac{m-t}{mt}} \|du\|_{m,\sigma B}, \tag{13}$$

where $\sigma > \nu > 1$ is a constant. Using the Hölder inequality again yields

$$\begin{aligned}
\|du\|_{m,\sigma B} &= \Big(\int_{\sigma B} \Big(|du||x-x_B|^{-\lambda/s}|x-x_B|^{\lambda/s}\Big)^m dx\Big)^{1/m} \\
&\leq \Big(\int_{\sigma B} \Big(|du||x-x_B|^{-\lambda/s}\Big)^s dx\Big)^{1/s} \Big(\int_{\sigma B} \Big(|x-x_B|^{\lambda/s}\Big)^{\frac{ms}{s-m}} dx\Big)^{\frac{s-m}{ms}} \\
&\leq \Big(\int_{\sigma B} |du|^s |x-x_B|^{-\lambda} dx\Big)^{1/s} C_4(n,s,\alpha,\Omega)(\sigma r_B)^{\lambda/s + n(s-m)/ms} \\
&\leq C_5(n,s,\alpha,\Omega) \Big(\int_{\sigma B} |du|^s |x-x_B|^{-\lambda} dx\Big)^{1/s} (r_B)^{\lambda/s + n(s-m)/ms}.
\end{aligned} \tag{14}$$

Note that

$$\mathrm{diam}(B) \cdot |B|^{1+\frac{1}{t}-\frac{1}{m}} = |B|^{\frac{1}{n}+\frac{1}{t}-\frac{ns+\alpha t-\lambda t}{nst}} = |B|^{\frac{1}{n}-\frac{\alpha-\lambda}{ns}}. \tag{15}$$

Substituting (12)–(14) in (11) and using (15), we have

$$\left(\int_B (|G(u)-(G(u))_B|)^s \frac{1}{|x-x_B|^\alpha}dx\right)^{1/s} \le C_5(n,s,\alpha,\lambda,\Omega)|B|^\gamma \left(\int_{\sigma B}|du|^s|x-x_B|^{-\lambda}dx\right)^{1/s}.$$

We have completed the proof of Theorem 1.

Since $\frac{1}{d(x,\partial\Omega)} \le \frac{1}{r_B-|x|}$ for any $x \in B$, where r_B is the radius of ball $B \subset \Omega$, using the same method developed in the proof of Theorem 1, we obtain the following Poincaré-type inequality for Green's operator with unbounded factors.

Theorem 2. *Let $du \in L^s_{\mathrm{loc}}(\Omega,\wedge^l)$, $l=1,2,\ldots,n$, $1<s<\infty$, be a solution of the nonhomogeneous A-harmonic equation in a bounded domain Ω, G be Green's operator. Then, there exists a constant C, independent of u, such that*

$$\left(\int_B |G(u)-(G(u))_B|^s \frac{1}{d^\alpha(x,\partial\Omega)}dx\right)^{1/s} \le C(n,s,\alpha,\lambda,\Omega)|B|^\gamma \left(\int_{\sigma B}|du|^s \frac{1}{|x-x_B|^\lambda}dx\right)^{1/s} \tag{16}$$

for all balls B with $\sigma B \subset \Omega$ and any real numbers α and λ with $\alpha > \lambda \ge 0$, where $\gamma = \frac{1}{n} - \frac{\alpha-\lambda}{ns}$ and x_B is the center of ball B and $\sigma > 1$ is a constant.

3 Inequalities in John Domains

Finally, we are ready to prove the global Poincaré-type inequalities for Green's operator with unbounded factors in John domains.

Definition 1. A proper subdomain $\Omega \subset \mathbb{R}^n$ is called a δ-John domain, $\delta > 0$, if there exists a point $x_0 \in \Omega$ which can be joined with any other point $x \in \Omega$ by a continuous curve $\gamma \subset \Omega$ so that

$$d(\xi,\partial\Omega) \ge \delta|x-\xi|$$

for each $\xi \in \gamma$. Here $d(\xi,\partial\Omega)$ is the Euclidean distance between ξ and $\partial\Omega$.

Lemma 4 (Covering Lemma [17]). *Each Ω has a modified Whitney cover of cubes $\mathcal{V} = \{Q_i\}$ such that*

$$\cup_i Q_i = \Omega, \quad \sum_{Q_i \in \mathcal{V}} \chi_{\sqrt{\frac{5}{4}}Q_i} \le N\chi_\Omega$$

and some $N > 1$, and if $Q_i \cap Q_j \neq \emptyset$, then there exists a cube R (this cube need not be a member of $\mathcal{V}$) in $Q_i \cap Q_j$ such that $Q_i \cup Q_j \subset NR$. Moreover, if Ω is δ-John, then there is a distinguished cube $Q_0 \in \mathcal{V}$ which can be connected with every cube $Q \in \mathcal{V}$ by a chain of cubes $Q_0, Q_1, \ldots, Q_k = Q$ from $\mathcal{V}$ and such that $Q \subset \rho Q_i$, $i = 0,1,2,\ldots,k$, for some $\rho = \rho(n,\delta)$.

Theorem 3. *Let $u \in D'(\Omega,\wedge^0)$ be a solution of the nonhomogeneous A-harmonic equation (1) and s be a fixed exponent associated with the nonhomogeneous*

A-harmonic equation. Then, there exists a constant $C(n,N,s,\alpha,\lambda,Q_0,\Omega)$, *independent of u, such that*

$$\left(\int_\Omega |G(u)-(G(u))_{Q_0}|^s \frac{1}{d^\alpha(x,\partial\Omega)}dx\right)^{1/s} \le C(n,N,s,\alpha,\lambda,Q_0,\Omega)\left(\int_\Omega |du|^s g(x)dx\right)^{1/s} \tag{17}$$

for any bounded and convex δ*-John domain* $\Omega \subset \mathbb{R}^n$, *where* $g(x)=\sum_i \chi_{Q_i}\frac{1}{|x-x_{Q_i}|^\lambda}$. *Here* α *and* λ *are constants with* $0\le\lambda<\alpha<\min\{n,s+\lambda-n\}$, $s+\lambda>n$, *and the fixed cube* $Q_0\subset\Omega$, *the cubes* $Q_i\subset\Omega$, *and the constant* $N>1$ *appeared in Lemma 4,* x_{Q_i} *is the center of* Q_i.

Proof. Assume that $\mu(x)$ and $\mu_1(x)$ are the Radon measures induced by $d\mu=\frac{1}{d^\alpha(x,\partial\Omega)}dx$ and $d\mu_1(x)=g(x)dx$, respectively. We have

$$\mu(Q)=\int_Q \frac{1}{d^\alpha(x,\partial\Omega)}dx \ge \int_Q \frac{1}{(\mathrm{diam}(\Omega))^\alpha}dx = M(n,\alpha,\Omega)|Q|, \tag{18}$$

where $M(n,\alpha,\Omega)$ is a positive constant. Using the elementary inequality $(a+b)^s\le 2^s(|a|^s+|b|^s)$, $s\ge 0$, we find that

$$\begin{aligned}
&\left(\int_\Omega |G(u)-(G(u))_{Q_0}|^s\frac{1}{d^\alpha(x,\partial\Omega)}dx\right)^{1/s}=\left(\int_{\cup Q}|G(u)-(G(u))_{Q_0}|^s d\mu\right)^{1/s}\\
&\le \left(\sum_{Q\in\mathcal{V}}\left(2^s\int_Q|G(u)-(G(u))_Q|^s d\mu+2^s\int_Q|(G(u))_Q-(G(u))_{Q_0}|^s d\mu\right)\right)^{1/s}\\
&\le C_1(s)\left(\left(\sum_{Q\in\mathcal{V}}\int_Q|G(u)-(G(u))_Q|^s d\mu\right)^{1/s}+\left(\sum_{Q\in\mathcal{V}}\int_Q|(G(u))_Q-(G(u))_{Q_0}|^s d\mu\right)^{1/s}\right.
\end{aligned} \tag{19}$$

for a fixed $Q_0\subset\Omega$. The first sum in (19) can be estimated by using Theorem 2

$$\begin{aligned}
&\sum_{Q\in\mathcal{V}}\int_Q|G(u)-(G(u))_Q|^s d\mu\\
&\le C_2(n,s,\alpha,\lambda,\Omega)\sum_{Q\in\mathcal{V}}|Q|^{\gamma s}\int_{\rho Q}|du|^s d\mu_1\\
&\le C_3(n,s,\alpha,\lambda,\Omega)|\Omega|^{\gamma s}\sum_{Q\in\mathcal{V}}\int_Q\left(|du|^s d\mu_1\right)\chi_Q\\
&\le C_4(n,s,\alpha,\lambda,\Omega)|\Omega|^{\gamma s}\int_\Omega|du|^s d\mu_1\\
&\le C_5(n,s,\alpha,\lambda,\Omega)\int_\Omega|du|^s d\mu_1.
\end{aligned} \tag{20}$$

To estimate the second sum in (19), we use the property of δ-John domain. Fix a cube $Q\in\mathcal{V}$ and let $Q_0,Q_1,\ldots,Q_k=Q$ be the chain in Lemma 4

$$|(G(u))_Q-(G(u))_{Q_0}|\le\sum_{i=0}^{k-1}|(G(u))_{Q_i}-(G(u))_{Q_{i+1}}|. \tag{21}$$

The chain $\{Q_i\}$ also has property that, for each i, $i = 0, 1, \ldots, k-1$, with $Q_i \cap Q_{i+1} \neq \emptyset$, there exists a cube D_i such that $D_i \subset Q_i \cap Q_{i+1}$ and $Q_i \cup Q_{i+1} \subset ND_i$, $N > 1$. Then,

$$|D_i| \le |Q_i \cap Q_{i+1}| \le \max\{|Q_i|, |Q_{i+1}|\} \le |Q_i \cup Q_{i+1}| \le |ND_i| \le C_6(N)|D_i|$$

which gives

$$\frac{\max\{|Q_i|, |Q_{i+1}|\}}{|Q_i \cap Q_{i+1}|} \le \frac{\max\{|Q_i|, |Q_{i+1}|\}}{|D_i|} \le C_6(N).$$

For such D_j, $j = 0, 1, \ldots, k-1$, set $D = \min\{|D_0|, |D_1|, \ldots, |D_{k-1}|\}$. Then

$$\frac{\max\{|Q_i|, |Q_{i+1}|\}}{|Q_i \cap Q_{i+1}|} \le \frac{\max\{|Q_i|, |Q_{i+1}|\}}{|D|} \le C_7(N). \tag{22}$$

Using (18), (22), and Theorem 2, we obtain

$$\begin{aligned}
&|(G(u))_{Q_i} - (G(u))_{Q_{i+1}}|^s \\
&= \frac{1}{\mu(Q_i \cap Q_{i+1})} \int_{Q_i \cap Q_{i+1}} |(G(u))_{Q_i} - (G(u))_{Q_{i+1}}|^s \frac{dx}{d^\alpha(x, \partial\Omega)} \\
&\le C_8(n, \alpha, \Omega) \frac{1}{|Q_i \cap Q_{i+1}|} \int_{Q_i \cap Q_{i+1}} |(G(u))_{Q_i} - (G(u))_{Q_{i+1}}|^s \frac{dx}{d^\alpha(x, \partial\Omega)} \\
&\le C_8(n, \alpha, \Omega) \frac{C_7(N)}{\max\{|Q_i|, |Q_{i+1}|\}} \int_{Q_i \cap Q_{i+1}} |(G(u))_{Q_i} - (G(u))_{Q_{i+1}}|^s d\mu \\
&\le C_9(n, N, \alpha, \Omega) \sum_{j=i}^{i+1} \frac{1}{|Q_j|} \int_{Q_j} |G(u) - (G(u))_{Q_j}|^s d\mu \\
&\le C_{10}(n, N, s, \alpha, \lambda, \Omega) \sum_{j=i}^{i+1} \frac{|Q_j|^{\gamma s}}{|Q_j|} \int_{\rho Q_j} |du|^s d\mu_1 \\
&= C_{10}(n, N, s, \alpha, \lambda, \Omega) \sum_{j=i}^{i+1} |Q_j|^{\gamma s - 1} \int_{\rho Q_j} |du|^s d\mu_1.
\end{aligned} \tag{23}$$

Since $Q \subset NQ_j$ for $j = i, i+1$, $0 \le i \le k-1$, from (23), we have

$$\begin{aligned}
&|(G(u))_{Q_i} - (G(u))_{Q_{i+1}}|^s \chi_Q(x) \\
&\le C_{11}(n, N, s, \alpha, \lambda, \Omega) \sum_{j=i}^{i+1} \chi_{NQ_j}(x) |Q_j|^{\gamma s - 1} \int_{\rho Q_j} |du|^s d\mu_1 \\
&\le C_{12}(n, N, s, \alpha, \lambda, \Omega) \sum_{j=i}^{i+1} \chi_{NQ_j}(x) |\Omega|^{\gamma s - 1} \int_{\rho Q_j} |du|^s d\mu_1.
\end{aligned} \tag{24}$$

We know that $|\Omega|^{\gamma-1/s} < \infty$ since Ω is bounded and $\gamma - \frac{1}{s} = \frac{1}{n} + \frac{\lambda}{ns} - \frac{1}{s} - \frac{\alpha}{ns} > 0$ when $\alpha < s + \lambda + n(s-1)$. Thus, from $(a+b)^{1/s} \le 2^{1/s}(|a|^{1/s} + |b|^{1/s})$, (21) and (24), it follows that

$$|(G(u))_Q - (G(u))_{Q_0}|\chi_Q(x) \le C_{13}(n,N,s,\alpha,\lambda,\Omega) \sum_{D\in\mathcal{V}} \Big(\int_{\rho D} |du|^s d\mu_1\Big)^{1/s} \cdot \chi_{ND}(x)$$

for every $x \in \mathbb{R}^n$. Then,

$$\begin{aligned} &\sum_{Q\in\mathcal{V}} \int_Q |(G(u))_Q - (G(u))_{Q_0}|^s d\mu \\ &\le C_{13}(n,N,s,\alpha,\lambda,\Omega) \int_{\mathbb{R}^n} \Big| \sum_{D\in\mathcal{V}} \Big(\int_{\rho D} |du|^s d\mu_1\Big)^{1/s} \chi_D(x)\Big|^s d\mu. \end{aligned}$$

Notice that

$$\sum_{D\in\mathcal{V}} \chi_D(x) \le \sum_{D\in\mathcal{V}} \chi_{\rho D}(x) \le N\chi_\Omega(x).$$

Using elementary inequality $|\sum_{i=1}^M t_i|^s \le M^{s-1}\sum_{i=1}^M |t_i|^s$, we finally have

$$\begin{aligned} &\sum_{Q\in\mathcal{V}} \int_Q |(G(u))_Q - (G(u))_{Q_0}|^s d\mu \\ &\le C_{14}(n,N,s,\alpha,\lambda,\Omega) \int_{\mathbb{R}^n} \Big(\sum_{D\in\mathcal{V}} \Big(\int_{\rho D} |du|^s d\mu_1\Big) \chi_D(x)\Big) d\mu \\ &= C_{14}(n,N,s,\alpha,\lambda,\Omega) \sum_{D\in\mathcal{V}} \Big(\int_{\rho D} |du|^s d\mu_1\Big) \\ &\le C_{15}(n,N,s,\alpha,\lambda,\Omega) \int_\Omega |du|^s d\mu_1. \end{aligned} \tag{25}$$

Substituting (20) and (25) in (19), we have proved Theorem 3.

4 Inequalities with Orlicz Norms

In this section, we first prove the local Poincaré inequalities for Green's operator applied to differential forms. Then, we extend the local Poincaré-type inequalities into the global cases in the $L^\varphi(m)$-averaging domains, which are the extension of John domains and L^s-averaging domain, see [7, 19]. A continuously increasing function $\varphi : [0,\infty) \to [0,\infty)$ with $\varphi(0) = 0$ is called an Orlicz function. The Orlicz space $L^\varphi(\Omega)$ consists of all measurable functions f on Ω such that $\int_\Omega \varphi\left(\frac{|f|}{\lambda}\right) dx < \infty$ for some $\lambda = \lambda(f) > 0$. $L^\varphi(\Omega)$ is equipped with the nonlinear Luxemburg functional

$$\|f\|_{\varphi(\Omega)} = \inf\ \{\lambda > 0 : \int_\Omega \varphi\left(\frac{|f|}{\lambda}\right) dx \le 1\}. \tag{26}$$

A convex Orlicz function φ is often called a Young function. If φ is a Young function, then $\|\cdot\|_\varphi$ defines a norm in $L^\varphi(\Omega)$, which is called the Luxemburg norm.

Definition 2 ([2]). We say a Young function φ lies in the class $G(p,q,C)$, $1 \le p < q < \infty$, $C \ge 1$, if (i) $1/C \le \varphi(t^{1/p})/\Phi(t) \le C$ and (ii) $1/C \le \varphi(t^{1/q})/\Psi(t) \le C$ for all $t > 0$, where Φ is a convex increasing function and Ψ is a concave increasing function on $[0,\infty)$

From [2], each of φ, Φ, and Ψ in above definition is doubling in the sense that its values at t and $2t$ are uniformly comparable for all $t > 0$, and the consequent fact that

$$C_1 t^q \le \Psi^{-1}(\varphi(t)) \le C_2 t^q, \;\; C_1 t^p \le \Phi^{-1}(\varphi(t)) \le C_2 t^p, \tag{27}$$

where C_1 and C_2 are constants. Also, for all $1 \le p_1 < p < p_2$ and $\alpha \in \mathbb{R}$, the function $\varphi(t) = t^p \log_+^\alpha t$ belongs to $G(p_1,p_2,C)$ for some constant $C = C(p,\alpha,p_1,p_2)$. Here $\log_+(t)$ is defined by $\log_+(t) = 1$ for $t \le e$; and $\log_+(t) = \log(t)$ for $t > e$. Particularly, if $\alpha = 0$, we see that $\varphi(t) = t^p$ lies in $G(p_1,p_2,C)$, $1 \le p_1 < p < p_2$.

We first prove the following generalized Poincaré inequality that will be used to establish the global inequality.

Theorem 4. *Let φ be a Young function in the class $G(p,q,C)$, $1 \le p < q < \infty$, $C \ge 1$, Ω be a bounded domain and $q(n-p) < np$. Assume that $u \in D'(\Omega, \wedge^l)$ is any differential l-form, $l = 0,1,\ldots,n-1$, and $\varphi(|du|) \in L^1_{\mathrm{loc}}(\Omega,m)$. Then, there exists a constant C, independent of u, such that*

$$\int_B \varphi\left(|G(u) - (G(u))_B|\right) dm \le C \int_B \varphi\left(|du|\right) dm \tag{28}$$

for all balls B with $B \subset \Omega$.

Proof. Using Jensen's inequality for Ψ^{-1}, (8), and noticing that φ and Ψ are doubling, we obtain

$$\begin{aligned}
\int_B \varphi\Big(|G(u) - (G(u))_B|\Big) dm &= \Psi\Big(\Psi^{-1}\Big(\int_B \varphi\left(|G(u) - (G(u))_B|\right) dm\Big)\Big) \\
&\le \Psi\Big(\int_B \Psi^{-1}\Big(\varphi\left(|G(u) - (G(u))_B|\right)\Big) dm\Big) \\
&\le \Psi\Big(C_7 \int_B |G(u) - (G(u))_B|^q dm\Big) \\
&\le C_8 \varphi\Big(\Big(C_7 \int_B |G(u) - (G(u))_B|^q dm\Big)^{1/q}\Big) \\
&\le C_9 \varphi\Big(\Big(\int_B |G(u) - (G(u))_B|^q dm\Big)^{1/q}\Big).
\end{aligned} \tag{29}$$

If $1 < p < n$, by assumption, we have $q < \frac{np}{n-p}$. Using the Poincaré-type inequality for differential forms $G(u)$

$$\begin{aligned}
&\left(\int_B |G(u) - (G(u))_B|^{np/(n-p)} dm\right)^{(n-p)/np} \\
&\leq C_2 \left(\int_B |d(G(u))|^p dm\right)^{1/p} \\
&\leq C_2 \left(\int_B |G(du)|^p dm\right)^{1/p} \\
&\leq C_2 \left(\int_B |du|^p dm\right)^{1/p}, \qquad (30)
\end{aligned}$$

we find that

$$\left(\int_B |G(u) - (G(u))_B|^q dm\right)^{1/q} \leq C_3 \left(\int_B |du|^p dm\right)^{1/p}. \qquad (31)$$

Note that the L^p-norm of $|G(u) - (G(u))_B|$ increases with p and $\frac{np}{n-p} \to \infty$ as $p \to n$, and it follows that (31) still holds when $p \geq n$. Since φ is increasing, from (28) and (31), we obtain

$$\int_B \varphi\Big(|G(u) - (G(u))_B|\Big) dm \leq C_1 \varphi\Big(C_3 \Big(\int_B |du|^p dm\Big)^{1/p}\Big). \qquad (32)$$

Applying (32), (i) in Definition 2, Jensen's inequality, and noticing that φ and Φ are doubling, we have

$$\begin{aligned}
\int_B \varphi\Big(|G(u) - (G(u))_B|\Big) dm &\leq C_1 \varphi\Big(C_3 \Big(\int_B |du|^p dm\Big)^{1/p}\Big) \\
&\leq C_1 \Phi\Big(C_4 \Big(\int_B |du|^p dm\Big)\Big) \\
&\leq C_5 \int_B \Phi(|du|^p) dm. \qquad (33)
\end{aligned}$$

Using (i) in Definition 2 again yields

$$\int_B \Phi(|du|^p) dm \leq C_6 \int_B \varphi(|du|) dm. \qquad (34)$$

Combining (33) and (34), we obtain

$$\int_B \varphi\Big(|G(u) - (G(u))_B|\Big) dm \leq C_7 \int_B \varphi(|du|) dm. \qquad (35)$$

The proof of Theorem 4 has been completed.

Since each of φ, Φ, and Ψ in Definition 2 is doubling, from the proof of Theorem 4 or directly from (28), we have

$$\int_B \varphi\left(\frac{|G(u)-(G(u))_B|}{\lambda}\right) dm \le C \int_{\sigma B} \varphi\left(\frac{|du|}{\lambda}\right) dm \tag{36}$$

for all balls B with $\sigma B \subset \Omega$ and any constant $\lambda > 0$. From (26) and (36), the following Poincaré inequality with the Luxemburg norm

$$\|G(u)-(G(u))_B\|_{\varphi(B)} \le C\|du\|_{\varphi(\sigma B)} \tag{37}$$

holds under the conditions described in Theorem 4.

Using Lemma 3.7.2 with $w(x) = 1$ in [1], we have the following Poincaré-type inequality for the composition of Δ and G.

Lemma 5. *Let $u \in D'(\Omega, \wedge^l)$, $l = 0, 1, \cdots, n-1$, be an A-harmonic tensor on Ω. Assume that $\rho > 1$ and $1 < s < \infty$. Then, there exists a constant C, independent of u, such that*

$$\|\Delta G(u) - (\Delta G(u))_B\|_{s,B} \le C\mathrm{diam}(B)\|du\|_{s,\rho B} \tag{38}$$

for any ball B with $\rho B \subset \Omega$.

Using Lemma 5 and the method developed in the proof of Theorem 4, we can prove the following version of Poincaré-type inequality for the composition of Δ and G.

Theorem 5. *Let φ be a Young function in the class $G(p,q,C)$, $1 \le p < q < \infty$, $C \ge 1$, Ω be a bounded domain and $q(n-p) < np$. Assume that $u \in D'(\Omega, \wedge^l)$ is any differential l-form, $l = 0, 1, \ldots, n-1$, and $\varphi(|du|) \in L^1_{\mathrm{loc}}(\Omega, m)$. Then, there exists a constant C, independent of u, such that*

$$\int_B \varphi\left(|\Delta G(u) - (\Delta G(u))_B|\right) dm \le C \int_B \varphi\left(|du|\right) dm \tag{39}$$

for all balls B with $B \subset \Omega$.

Now we extend Theorem 1 into the global cases in the following $L^{\varphi}(m)$-averaging domains.

Definition 3 ([7]). Let φ be an increasing convex function on $[0, \infty)$ with $\varphi(0) = 0$. We call a proper subdomain $\Omega \subset \mathbb{R}^n$ an $L^{\varphi}(m)$-averaging domain, if $m(\Omega) < \infty$ and there exists a constant C such that

$$\int_{\Omega} \varphi(\tau|u - u_{B_0}|) dm \le C \sup_{B \subset \Omega} \int_B \varphi(\sigma|u - u_B|) dm \tag{40}$$

for some ball $B_0 \subset \Omega$ and all u such that $\varphi(|u|) \in L^1_{loc}(\Omega, m)$, where τ, σ are constants with $0 < \tau < \infty$, $0 < \sigma < \infty$ and the supremum is over all balls $B \subset \Omega$.

From the above definition we see that L^s-averaging domains and $L^s(m)$-averaging domains are special $L^\varphi(m)$-averaging domains when $\varphi(t)=t^s$ in Definition 3. Also, uniform domains and John domains are very special $L^\varphi(m)$-averaging domains, see [1, 7, 19] for more results about domains.

Theorem 6. *Let φ be a Young function in the class $G(p,q,C)$, $1\le p<q<\infty$, $C\ge 1$, Ω be a bounded $L^\varphi(m)$-averaging domain and $q(n-p)<np$. Assume that $u\in D'(\Omega,\wedge^0)$ and $\varphi(|du|)\in L^1(\Omega,m)$. Then, there exists a constant C, independent of u, such that*

$$\int_\Omega \varphi\left(|G(u)-(G(u))_{B_0}|\right)dm\le C\int_\Omega \varphi\left(|du|\right)dm, \tag{41}$$

where $B_0\subset\Omega$ is some fixed ball.

Proof. From Definition 3, (28) and noticing that φ is doubling, we have

$$\begin{aligned}\int_\Omega \varphi\left(|G(u)-(G(u))_{B_0}|\right)dm &\le C_1\sup_{B\subset\Omega}\int_B \varphi\left(|G(u)-(G(u))_B|\right)dm\\ &\le C_1\sup_{B\subset\Omega}\left(C_2\int_{\sigma B}\varphi\left(|du|\right)dm\right)\\ &\le C_1\sup_{B\subset\Omega}\left(C_2\int_{\Omega}\varphi\left(|du|\right)dm\right)\\ &\le C_3\int_\Omega \varphi\left(|du|\right)dm. \end{aligned}\tag{42}$$

We have completed the proof of Theorem 6.

Similar to the local case, the following global Poincaré inequality with the Orlicz norm

$$\|G(u)-(G(u))_{B_0}\|_{\varphi(\Omega)}\le C\|du\|_{\varphi(\Omega)} \tag{43}$$

holds if all conditions in Theorem 6 are satisfied.

Choosing $\varphi(t)=t^p\log_+^\alpha t$ in Theorem 6, we obtain the following Poincaré inequalities with the $L^p(\log_+^\alpha L)$-norms.

Corollary 1. *Let $\varphi(t)=t^p\log_+^\alpha t$, $1\le p_1<p<p_2$ and $\alpha\in\mathbb{R}$ and Ω be a bounded $L^\varphi(m)$-averaging domain and $p_2(n-p_1)<np_1$. Assume that $u\in D'(\Omega,\wedge^0)$ and $\varphi(|du|)\in L^1(\Omega,m)$. Then, there exists a constant C, independent of u, such that*

$$\int_\Omega |G(u)-(G(u))_{B_0}|^p\log_+^\alpha\left(|G(u)-(G(u))_{B_0}|\right)dm\le C\int_\Omega |du|^p\log_+^\alpha\left(|du|\right)dm, \tag{44}$$

where $B_0\subset\Omega$ is some fixed ball.

Note that (43) can be written as the following version with the Luxemburg norm

$$\|G(u)-(G(u))_{B_0}\|_{L^p(\log_+^\alpha L)(\Omega)}\le C\|du\|_{L^p(\log_+^\alpha L)(\Omega)}$$

provided the conditions in Corollary 1 are satisfied.

Remark (i) If u is a differential function (0-form) in a domain $\Omega \subset \mathbb{R}^n$, (1) for differential forms reduces to

$$\text{div}A(x, \nabla u) = B(x, \nabla u), \quad x \in \Omega \tag{45}$$

which is called the nonhomogeneous A-harmonic equation for functions. If the operator $B = 0$, (45) becomes

$$\text{div}A(x, \nabla u) = 0, \quad x \in \Omega \tag{46}$$

which is called the homogeneous A-harmonic equation for functions. In the case that the operator $A(x, \xi) = \xi |\xi|^{p-2}$ in (46) with $p > 1$, the homogeneous A-harmonic equation (46) reduces to the usual p-harmonic equation for functions

$$\text{div}(\nabla u |\nabla u|^{p-2}) = 0. \tag{47}$$

Let $p = 2$ in (47), we obtain the Laplace equation $\Delta u = 0$ for functions in $\Omega \subset \mathbb{R}^n$. Hence, each of Eqs. (45)–(47) and the equation $\Delta u = 0$ is the special case of the nonhomogeneous A-harmonic equation (1). All results obtained in this paper are still true for solutions of (45)–(47), that is, each theorem proved in this paper holds for A-harmonic functions and p-harmonic functions. Especially, in Sect. 4, u does not need to be a solution of any version of the A-harmonic equations. (ii) When dealing with the integral of the vector field $\mathbf{F} = \nabla f$, we will face the singular integral if the potential function f contains a singular factor, such as the potential energy in physics. We believe that the Poincaré-type inequalities with singular factors will find more applications in many fields in mathematics and physics.

References

1. Agarwal, R.P., Ding, S., Nolder, C.A.: Inequalities for Differential Forms. Springer, New York (2009)
2. Buckley, S.M., Koskela, P.: Orlicz-Hardy inequalities. Ill. J. Math. **48**, 787–802 (2004)
3. Cartan, H.: Differential Forms. Houghton Mifflin, Boston (1970)
4. Chanillo, S., Wheeden, R.L.: Weighted Poincaré and Sobolev inequalities and estimates for weighted Peano maximal functions. Am. J. Math. **107**, 1191–1226 (1985)
5. Ding, S.: Integral estimates for the Laplace-Beltrami and Green's operators applied to differential forms. Zeitschrift fur Analysis und ihre Anwendungen (J. Anal. Appl.) **22**, 939–957 (2003)
6. Ding, S.: Two-weight Caccioppoli inequalities for solutions of nonhomogeneous A-harmonic equations on Riemannian manifolds. Proc. Am. Math. Soc. **132**, 2367–2375 (2004)
7. Ding, S.: $L^{\varphi}(\mu)$-averaging domains and the quasihyperbolic metric. Comput. Math. Appl. **47**, 1611–1618 (2004)
8. Ding, S., Nolder, C.A.: Weighted Poincaré-type inequalities for solutions to the A-harmonic equation. Ill. J. Math. **2**, 199–205 (2002)
9. Franchi, B., Wheeden, R.L.: Some remarks about Poincaré type inequalities and representation formulas in metric spaces of homogeneous type. J. Inequal. Appl. **3**(1), 65–89 (1999)
10. Franchi, B., Gutiérrez, C.E., Wheeden, R.L.: Weighted Sobolev-Poincaré inequalities for Grushin type operators. Commun. Partial Differ. Equ. **19**, 523–604 (1994)

11. Franchi, B., Lu, G., Wheeden, R.L.: Weighted Poincaré inequalities for Hörmander vector fields and local regularity for a class of degenerate elliptic equations. Potential theory and degenerate partial differential operators (Parma). Potential Anal. **4**, 361–375 (1995)
12. Franchi, B., Pérez, C., Wheeden, R.L.: Sharp geometric Poincaré inequalities for vector fields and non-doubling measures. Proc. Lond. Math. Soc. **80**, 665–689 (2000)
13. Iwaniec, T.: p-harmonic tensors and quasiregular mappings. Ann. Math. (2) **136**(3), 589–624 (1992)
14. Iwaniec, T., Lutoborski, A.: Integral estimates for null Lagrangians. Arch. Ration. Mech. Anal. **125**, 25–79 (1993)
15. Liu, B.: $A_r^{\lambda}(\Omega)$-weighted imbedding inequalities for A-harmonic tensors. J. Math. Anal. Appl. **273**(2), 667–676 (2002)
16. Liu, B.: L^p-estimates for the solutions of A-harmonic equations and the related operators. In: The Proceedings of the 6th International Conference on Differential Equations and Dynamical Systems (2008)
17. Nolder, C.A.: Hardy-Littlewood theorems for A-harmonic tensors. Ill. J. Math. **43**, 613–631 (1999)
18. Scott, C.: L^p-theory of differential forms on manifolds. Trans. Am. Math. Soc. **347**, 2075–2096 (1995)
19. Staples, S.G.: L^p-averaging domains and the Poincaré inequality. Ann. Acad. Sci. Fenn. Ser. AI Math. **14**, 103–127 (1989)
20. Stroffolini, B.: On weakly A-harmonic tensors. Stud. Math. **114**(3), 289–301 (1995)
21. Warner, F.W.: Foundations of differentiable manifolds and Lie groups. Springer, New York (1983)
22. Xing, Y.: Weighted integral inequalities for solutions of the A-harmonic equation. J. Math. Anal. Appl. **279**, 350–363 (2003)
23. Xing, Y.: Two-weight imbedding inequalities for solutions to the A-harmonic equation. J. Math. Anal. Appl. **307**, 555–564 (2005)

The Robustness Concern in Preference Disaggregation Approaches for Decision Aiding: An Overview

Michael Doumpos and Constantin Zopounidis

1 Introduction

Managers, analysts, policy makers, and regulators are often facing multiple technical, socio-economic, and environmental objectives, goals, criteria, and constraints, in a complex and ill-structured decision-making framework, encountered in all aspects of the daily operation of firms, organizations, and public entities. Coping with such a diverse and conflicting set of decision factors poses a significant burden to the decision process when ad hoc empirical procedures are employed.

Multiple criteria decision aid (MCDA) has evolved into a major discipline in operations research/management science, which is well-suited for problem structuring, modeling, and analysis in this context. MCDA provides a wide arsenal of methodologies and techniques that enable the systematic treatment of decision problems under multiple criteria, in a rigorous yet flexible manner, taking into consideration the expertise, preferences, and judgment policy of the decision makers (DMs) involved. The MCDA framework is applicable in a wide range of different types of decision problems, including deterministic and stochastic problems, static and dynamic problems, as well as in situations that require the consideration of fuzzy and qualitative data of either small or large scale, by a single DM or a group of DMs. A comprehensive overview of the recent advances in the theory and practice of MCDA can be found in the book of Zopounidis and Pardalos [68].

M. Doumpos (✉)
School of Production Engineering and Management, Technical University of Crete, University Campus, 73100 Chania, Greece
e-mail: mdoumpos@dpem.tuc.gr

C. Zopounidis
School of Production Engineering and Management, Technical University of Crete, University Campus, 73100 Chania, Greece

Audencia Group, Nantes School of Management, Nantes, France
e-mail: kostas@dpem.tuc.gr

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_8,

Similarly to other OR and management science modeling approaches, MCDA techniques are also based on assumptions and estimates on the characteristics of the problem, the aggregation of the decision criteria, and the preferential system of the DM. Naturally, such assumptions and estimates incorporate uncertainties, fuzziness, and errors, which affect the results and recommendations provided to the DM. As a result, changes in the decision context, the available data, or a reconsideration of the decision criteria and the goals of the analysis, may ultimately require a very different modeling approach leading to completely different outputs. Thus, even if the results may be judged satisfactory when modeling and analyzing the problem, their actual implementation in practice often leads to new challenges not taken previously into consideration.

In this context, robustness analysis has emerged as a major research issue in MCDA. Robustness analysis seeks to address the above issues through the introduction of a new modeling paradigm based on the idea that the multicriteria problem structuring and criteria aggregation process should not be considered in the context of a well-defined, strict set of conditions, assumptions, and estimates, but rather to seek to provide satisfactory outcomes even in cases where the decision context is altered.

Vincke [61] emphasized that robustness should not be considered in the restrictive framework of stochastic analysis (see also [34] for a discussion in the context of discrete optimization) and distinguished between robust solutions and robust methods. He further argued that although robustness is an appealing property, it is not a sufficient condition to judge the quality of a method or a solution. Roy [45], on the other hand, introduced the term *robustness concern* to emphasize that robustness is taken into consideration a priori rather than a posteriori (as is the case of sensitivity analysis). In the framework of Roy, the robustness concern is raised by *vague approximations* and *zones of ignorance* that cause the formal representation of a problem to diverge from the real-life context, due to: (1) the way imperfect knowledge is treated, (2) the inappropriate preferential interpretation of certain types of data (e.g., transformations of qualitative attributes), (3) the use of modeling parameters to grasp complex aspects of reality, and (4) the introduction of technical parameters with no concrete meaning. An recent example of robustness in the context of multiobjective linear programming can be found in Georgiev et al. [18]. The framework for robust decision aid has some differences compared to the traditional approach to robustness often encounter in other OR areas. A discussion of these differences (and similarities) can be found in Hites et al. [28].

The robustness concern is particularly important in the context of the preference disaggregation approach of MCDA, which is involved with the inference of preferential information and decision models from data. Disaggregation techniques are widely used to facilitate the construction of multicriteria evaluation models, based on simple information that can the DM can provide [30], without requiring the specification of complex parameters whose concept is not clearly understood by the DMs. In this chapter we provide an overview of the robustness concern in the preference disaggregation context, covering the issues and factors that affect the robustness of disaggregation methods, the approaches that have been proposed to

deal with robustness in this area, and the existing connections with concepts and methodologies from the area of statistical learning.

The rest of the chapter is organized as follows. Section 2 presents the context of preference disaggregation analysis (PDA) with examples from ordinal regression and classification problems. Section 3 discusses the concept of robustness in disaggregation methods and some factors that affect it, whereas Sect. 4 overviews the different approaches that have been proposed to obtain robust recommendations and models in PDA. Section 5 presents the statistical learning perspective and discusses its connections to the MCDA disaggregation framework. Finally, Sect. 6 concludes the chapter and proposes some future research directions.

2 Preference Disaggregation Analysis

2.1 General Framework

A wide class of MCDA problems requires the evaluation of a discrete set of alternatives (i.e., ways of actions, options) $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ described on the basis of n evaluation criteria. The DM may be interested in choosing the best alternatives, ranking the alternatives from the best to the worst, or classifying them into predefined performance categories.

In this context, the construction of an evaluation model that aggregates the performance criteria and provides recommendations in one of the above forms, requires some preferential information by the DM (e.g., the relative importance of the criteria). This information can be specified either through interactive, structured communication sessions between the analyst and the DM or it can be inferred from a sample of representative decision examples provided by the DM. PDA adopts the latter approach, which is very convenient in situations where, due to cognitive or time limitations, the DM is unwilling or unable to provide the analyst with specific information on a number of technical parameters (which are often difficult to understand) required to formulate the evaluation model.

PDA provides a general methodological framework for the development of multicriteria evaluation models using examples of decisions taken by a DM (or a group of DMs), so that DM's system of preferences is represented in the models as accurately as possible. The main input used in this process is a reference set of alternatives evaluated by the DM (decision examples). The reference set may consist of past decisions, a subset of the alternatives under consideration, or a set of fictitious alternatives which can be easily judged by the DM [30]. Depending on the decision problematic, the evaluation of the reference alternatives may be expressed by defining an order structure (total, weak, partial, etc.) or by classifying them into appropriate classes.

Formally, let $\mathcal{D}(X')$ denote the DM's evaluation of a set X' consisting of m reference alternatives described over n criteria (the description of alternative i on criterion

k will henceforth be denoted by x_{ik}). The DM's evaluation is assumed to be based (implicitly) on a decision model f_β defined by some parameters β, which represent the actual preferential system of the DM. Different classes of models can be considered. Typical examples include:

- Value functions defined such that $V(\mathbf{x}_i) > V(\mathbf{x}_j)$ if alternative i is preferred over alternative j and $V(\mathbf{x}_i) = V(\mathbf{x}_j)$ in cases of indifference [33]. The parameters of a value function model involve the criteria trade-offs and the form of the marginal value functions.
- Outranking relations defined such that $\mathbf{x}_i S \mathbf{x}_j$ if alternative i is at least as good as alternative j. The parameters of an outranking model may involve the weights of the criteria, as well as preference, indifference, and veto thresholds, etc. (for details see [44, 60]).
- "If ... then ..." decision rules [19]. In this case the parameters of the model involve the conditions and the conclusions associated to each rule.

The objective of PDA is to infer the "optimal" parameters $\hat{\beta}^*$ that approximate, as accurately as possible, the actual preferential system of the DM as represented in the unknown set of parameters β, i.e.:

$$\hat{\beta}^* = \arg\min_{\hat{\beta} \in \mathcal{A}} \|\hat{\beta} - \beta\| \tag{1}$$

where $\mathcal{A}$ is a set of feasible values for the parameters $\hat{\beta}$. With the obtained parameters, the evaluations performed with the corresponding decision model $f_{\hat{\beta}^*}$ will be consistent with the evaluations actually performed by the DM for any set of alternatives.

Problem (1), however, cannot be solved explicitly because β is unknown. Instead, an empirical estimation approach is employed using the DM's evaluation of the reference alternatives to proxy β. Thus, the general form of the optimization problem is expressed as follows:

$$\hat{\beta}^* = \arg\min_{\hat{\beta} \in \mathcal{A}} L[\mathcal{D}(X'), \hat{\mathcal{D}}(X')] \tag{2}$$

where $\hat{\mathcal{D}}(X')$ denotes the recommendations of the model $f_{\hat{\beta}}$ for the alternatives in X' and $L(\cdot)$ is a function that measures the differences between $\mathcal{D}(X')$ and $\hat{\mathcal{D}}(X')$.

2.2 *Inferring Value Function Models for Ordinal Regression and Classification Problems*

The general framework of PDA is materialized in several MCDA methods that enable the development of decision models in different forms [14, 50, 67]. To facilitate the exposition we shall focus on functional models expressed in the form of additive value functions, which have been widely used in MCDA.

A general multiattribute value function aggregates all the criteria into an overall performance index V (global value) defined such that:

$$\begin{aligned} V(\mathbf{x}_i) > V(\mathbf{x}_j) &\Leftrightarrow \mathbf{x}_i \succ \mathbf{x}_j \\ V(\mathbf{x}_i) = V(\mathbf{x}_j) &\Leftrightarrow \mathbf{x}_i \sim \mathbf{x}_j \end{aligned} \tag{3}$$

where $\succ$ and $\sim$ denote the preference and indifference relations, respectively. A value function may be expressed in different forms, depending on the criteria independence conditions [33]. Due to its simplicity, the most widely used form of value function is the additive one:

$$V(\mathbf{x}_i) = \sum_{k=1}^{n} w_k v_k(x_{ik}) \tag{4}$$

where w_k is the (nonnegative) trade-off constant of criterion k (the trade-offs are often normalized to sum up to one) and $v_k(\cdot)$ is the marginal value functions of the criterion, usually scaled such that $v_k(x_{k*}) = 0$ and $v_k(x_k^*) = 1$, where x_{k*} and x_k^* are the least and the most preferred levels of criterion k, respectively.

Such a model can be used to rank a set of alternatives or to classify them in predefined groups. In the ranking case, the relationships (3) provide a straightforward way to compare the alternatives. In the classification case, the simplest approach is to define an ordinal set of groups $G_1, G_2, \ldots, G_q$ on the value scale with the following rule:

$$t_\ell < V(\mathbf{x}_i) < t_{\ell-1} \Leftrightarrow \mathbf{x}_i \in G_\ell \tag{5}$$

where $t_1 > t_2 \cdots > t_{q-1}$ are thresholds that distinguish the groups. Alternative classification rules can also be employed such as the example-based approach of Greco et al. [21] or the hierarchical model of Zopounidis and Doumpos [66].

The construction of a value function from a set of reference examples can be performed with mathematical programming formulations. For example, in an ordinal regression setting, the DM's defines a weak-order of the alternatives in the reference set, by ranking them from the best (alternative $\mathbf{x}_1$) to the worst one (alternative $\mathbf{x}_m$). Then, the general form of the optimization problem for inferring a decision model from the data can be expressed as in the case of the UTA method [29] as follows:

$$\begin{aligned} \min \quad & \sigma_1 + \sigma_2 + \cdots + \sigma_m \\ \text{s.t.} \quad & \sum_{k=1}^{n} w_k[v_k(x_{ik}) - v_k(x_{i+1,k})] + \sigma_i - \sigma_{i+1} \geq \delta && \forall \mathbf{x}_i \succ \mathbf{x}_{i+1} \\ & \sum_{k=1}^{n} w_k[v_k(x_{ik}) - v_k(x_{i+1,k})] + \sigma_i - \sigma_{i+1} = 0 && \forall \mathbf{x}_i \sim \mathbf{x}_{i+1} \\ & w_1 + w_2 + \cdots + w_n = 1 \\ & v_k(x_{ik}) - v_k(x_{jk}) \geq 0 && \forall x_{ik} \geq x_{jk} \\ & v_k(x_{k*}) = 0,\ v_k(x_k^*) = 1 && k = 1, \ldots, n \\ & w_k,\ v_k(x_{ik}),\ \sigma_i \geq 0, && \forall i, k \end{aligned} \tag{6}$$

where $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ and $\mathbf{x}_* = (x_{1*}, \ldots, x_{n*})$ represent the ideal and anti-ideal alternatives, respectively. The solution of this optimization problem provides a value function that reproduces the DM's ranking of the reference alternatives as accurately as possible. The differences between the model's recommendations and the DM's weak-order are measured by the error variables $\sigma_1, \ldots, \sigma_m$, which are defined through the first two constraints (with δ being a small positive constant). The third constraint normalizes the trade-off constants, whereas the fourth constraint ensures that the marginal value functions and non-decreasing (assuming that the criteria are expressed in maximization form).

For classification problems, the optimization formulation for inferring a classification model from the reference examples using the threshold-based rule (5) can be expressed as follows:

$$
\begin{aligned}
\min \quad & \sum_{\ell=1}^{q} \frac{1}{m_\ell} \sum_{\mathbf{x}_i \in G_\ell} (\sigma_i^+ + \sigma_i^-) & & \\
\text{s.t.} \quad & \sum_{k=1}^{n} w_k v_k(x_{ik}) + \sigma_i^+ \geq t_\ell + \delta & & \forall \mathbf{x}_i \in G_\ell, \ell = 1, \ldots, q-1 \\
& \sum_{k=1}^{n} w_k v_k(x_{ik}) - \sigma_i^- \leq t_\ell - \delta & & \forall \mathbf{x}_i \in G_\ell, \ell = 2, \ldots, q \\
& t_\ell - t_{\ell+1} \geq \varepsilon & & \ell = 1, \ldots, q-2 \\
& w_1 + w_2 + \cdots + w_n = 1 & & \\
& v_k(x_{ik}) - v_k(x_{jk}) \geq 0 & & \forall x_{ik} \geq x_{jk} \\
& v_k(x_{k*}) = 0, \, v_k(x_k^*) = 1 & & k = 1, \ldots, n \\
& w_k, \, \sigma_i^+, \, \sigma_i^- \geq 0 & & \forall i, k
\end{aligned}
\tag{7}
$$

The objective function minimizes the total weighted classification error, where the weights are defined on the basis of the number of reference alternatives from each class $(m_1, \ldots, m_q)$. The error variables σ^+ and σ^- are defined through the first two constraints as the magnitude of the violations of the classification rules, whereas the third constraint ensures that the class thresholds are non-increasing (with ε being a small positive constant).

For the case of an additive value function, the above optimization problems can be re-expressed in linear programming form with a piecewise linear modeling of the marginal values function (see for example [29]).

3 Robustness in Preference Disaggregation Approaches

The quality of models resulting from disaggregation techniques is usually described in terms of their accuracy, which can be defined as the level of agreement between the DM's evaluations and the outputs of the inferred model. For instance, in ordinal regression problems rank correlation coefficients (e.g., the Kendall's τ

or Spearman's ρ) can be used for this purpose, whereas in classification problems the classification accuracy rate and the area under the receiver operating characteristic curve are commonly used measures. Except for accuracy-related measures, however, the robustness of the inferred model is also a crucial feature. Recent experimental studies have shown that robustness and accuracy are closely related [59]. However, accuracy measurements are done ex-post and rely on the use of additional test data, while robustness is taken into consideration ex-ante, thus making it an important issue that is taken into consideration before a decision model is actually put into practical use.

The robustness concern in the context of PDA arises because in most cases multiple alternative decision models can be inferred in accordance with the information embodied in the set of reference decision examples that a DM provides. This is particularly true for reference sets that do not contain inconsistencies, but it is also relevant when inconsistencies do exist (in the PDA context, inconsistencies are usually resolved algorithmically or interactively with the DM before the final model is built; see for instance [41]). With a consistent reference set, the error variables in formulations (6)–(7) become equal to zero and consequently these optimization models reduce to a set of feasible linear constraints. Each solution satisfying these constraints corresponds to a different decision model and even though all the corresponding feasible decision models provide the same outputs for the reference set, their recommendations can differ significantly when the models are used to perform evaluations for other alternatives.

For instance, consider the example data of Table 1 for a classification problem where a DM classified six references alternatives in two categories, under three evaluation criteria. Assuming a linear weighted average model of the form $V(\mathbf{x}_i) = w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3}$, with $w_1 + w_2 + w_3 = 1$ and $w_1, w_2, w_3 \geq 0$, the model would be consistent with the classification of the alternatives if $V(\mathbf{x}_i) \geq V(\mathbf{x}_j) + \delta$ for all $i = 1,2,3$ and $j = 4,5,6$, where δ is a small positive constant (e.g., $\delta = 0.01$). Figure 1 illustrates graphically the set of values for the criteria trade-offs that comply with the classification of the reference alternatives (the shaded area defined by the corner points A–E). It is evident that very different trade-offs provide the same results for the reference data. For example, the trade-off w_1 of the first criterion may vary anywhere from zero to one, whereas w_2 may vary from zero up to 0.7.

The size of the polyhedron defined by a set of feasible constraints of formulations such as (6) and (7) depends on a number of factors, but the two most important can be identified to be the adequacy of set of reference examples and the complexity of the selected decision modeling form. The former is immediately related to the quality of the information on which model inference is based. Vetschera et al. [59] performed an experimental analysis to investigate how the size of the reference set affects the robustness and accuracy of the resulting multicriteria models in classification problems. They found that small reference sets (e.g., with a limited number of alternatives with respect to the number of criteria) lead to decision models that are neither robustness nor accurate. Expect for its size other characteristics of the reference set are also relevant. These may involve the existence of noisy data, outliers, the existence of correlated criteria, etc. [12].

Table 1 An illustrative classification problem

	Criteria			
Alternatives	x_1	x_2	x_3	Classification
$\mathbf{x}_1$	7	1	8	G_1
$\mathbf{x}_2$	4	5	8	G_1
$\mathbf{x}_3$	10	4	2	G_1
$\mathbf{x}_4$	2	4	1	G_2
$\mathbf{x}_5$	4	1	1	G_2
$\mathbf{x}_6$	1	2	5	G_2

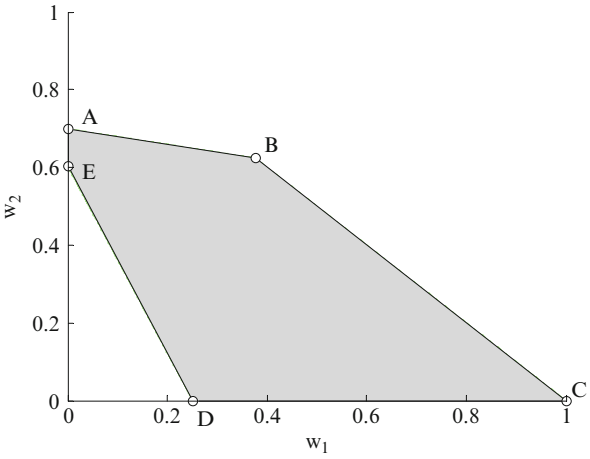


Fig. 1 The feasible set for the criteria trade-offs that are compatible with the classification of the example data of Table 1

The complexity of the inferred decision model is also an issue that is related to its robustness. Simpler models (e.g., a linear value function) are more robust compared to more complex nonlinear models. The latter are defined by a larger number of parameters and as a result the inference procedure becomes less robust and more sensitive to the available data. For instance, Fig. 2 illustrates a two-class classification problem with two criteria (which correspond to the axes of the figure). The linear classification model (green line) is robust; with the available data only marginal changes can be made in this model (separating line) without affecting its classification results for the data shown in the figure. On the other hand, a nonlinear model (blue line) is not robust, particularly in the areas where the data are sparse (i.e., the upper left and lower right parts of the graph). Therefore, care should be given to the selection of the appropriate modeling taking into account both the DM's

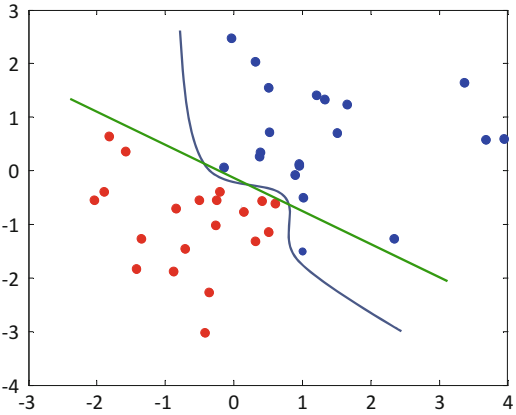


Fig. 2 A linear vs a nonlinear classification model

system of preferences as well as the available data. This issue has been studied extensively in areas such as the statistical learning theory [47, 56, 57].

4 Robust Disaggregation Approaches

The research in the area of building robust multicriteria decision models and obtaining robust recommendations with disaggregation techniques can be classified into three main directions. The first involves approaches that focus on describing the set of feasible decision models with analytic or simulation techniques. The second direction focuses on procedures for formulating robust recommendations through multiple acceptable decision models, whereas a third line of research has focused on techniques for selecting the most characteristic (representative) model from the set of all models compatible with the information provided by the reference set. The following subsections discuss these approaches in more detail.

4.1 Describing the Set of Acceptable Decision Models

The DM's evaluations for the reference alternatives provide information on the set of acceptable decision models that comply with these evaluations. Searching for different solutions within this feasible set and measuring its size provides useful information on the robustness of the results. Analytic and simulation-based techniques have been used for this purpose, focusing on convex polyhedral sets for which the analysis is computationally feasible. As explained in the previous section, for decision models which are linear with respect to their parameters (such as additive

value functions) the set of acceptable decision models is a convex polyhedron. The same applies to the other types of decision models with some simplifications on the parameters that are inferred (see, for example, [40]).

Jacquet-Lagrèze and Siskos [29] were the first to emphasize that the inference of a decision model through optimization formulations such as the ones described in Sect. 2.2 may not be robust thus suggesting that the existence of multiple optimal solutions (or even alternative near-optimal ones in the cases of inconsistent reference sets) should be carefully explored. The approach they suggested was based on a heuristic post-optimality procedure seeking to identify some characteristic alternative models corresponding to corner points of the feasible polyhedron. In the context of inferring an ordinal regression decision model, this approach is implemented in two phases. First, problem (6) is solved and its optimal objective function value F^* (total sum of errors) is recorded. In the second phase, $2n$ additional optimization problems are solved by maximizing and minimizing the trade-offs of the criteria (one at a time), while ensuring that the new solutions do not yield an overall error larger than $F^*(1+\alpha)$, where α is a small percentage of F^*. While this heuristic approach does not fully describe the polyhedron that defines the parameters of the decision model, it does give an indication of how much the relative importance of the criteria deviates within the polyhedron. Based on this approach, Grigoroudis and Siskos [24] developed a measure to assess the stability and robustness of the inferred model as the normalized standard deviation of the results obtained from the post-optimality analysis.

Despite their simplicity, post-optimality techniques provide only a limited partial view of the complete set of models that are compatible with the DM's preferences. A more thorough analysis requires the implementation of computationally intensive analytic or simulation approaches. Following the former direction, Vetschera [58] developed a recursive algorithm for computing the volume of the polyhedron that is derived from preferential constraints in the case of a linear evaluation model, but the algorithm was applicable to rather small problems (e.g., up to 20 alternatives and 6 criteria). Similar, but computationally more efficient algorithms, are available in the area of computational geometry, but they have not yet been employed in the context of MCDA. For instance, Lovász and Vempala [38] presented a fast algorithm for computing the volume of a convex polyhedron, which combines simulated annealing with multi-phase Monte Carlo sampling.

The computational difficulties of analytic techniques have led to the adoption of simulation approaches, which have gained much interest in the context of robust decision aiding. Originally used for sensitivity analysis [7] and decision aiding in stochastic environments [37], simulation techniques have been recently employed to facilitate the formulation of robust recommendations under different decision modeling forms. For instance, Tervonen et al. [52] used such an approach in order to formulate robust recommendations with the ELECTRE TRI multicriteria classification method [16], whereas Kadziński and Tervonen [31, 32] used a simulation-based approach to enhance the results of robust analytic techniques obtained with additive value models in the context of ranking and classification problems.

Simulation-based techniques were first based on rejection sampling schemes. Rejection sampling is a naïve approach under which a random model is constructed (usually from a uniform distribution [46]) and tested against the DM's evaluations for the reference alternatives. The model is accepted only if it is compatible with the DM's evaluations and rejected otherwise. However, the rejection rate increases rapidly with the dimensionality of the polyhedron (as defined by the number of the model's parameters). As a result the sampling of feasible solutions becomes intractable for problems of realistic complexity. Hit-and-run algorithms [35, 53] are particularly useful in reducing the computational burden, thus enabling the efficient sampling from high-dimensional convex regions.

4.2 Robust Decision Aid with a Set of Decision Models

Instead of focusing on the identification of different evaluation models that can be inferred from a set of reference decision examples through heuristic, analytic, or simulation approaches, a second line of research has been concerned with how robust recommendations can be formulated by aggregating the outputs of different models and exploiting the full information embodied in a given set of decision instances.

Siskos [49] first introduced the idea of building preference relations based on a set of decision models inferred with a preference disaggregation approach for ordinal regression problems. In particular, he presented the construction of a fuzzy preference relation based on the results of a post-optimality procedure. The fuzzy preference relation allows the evaluation of the alternatives through the aggregation of the outputs of multiple characteristic models (additive value functions) inferred from a set of decision instances.

Recently, this idea has been further extended to consider not only a subset of acceptable models but all models that can be inferred from a given reference set (without actually identifying them). Following this approach and in an ordinal regression setting, Greco et al. [20] defined necessary and possible preference relations on the basis of the DM's evaluations on a set of reference alternatives, as follows:

- Weak necessary preference relation: $\mathbf{x}_i \succsim^N \mathbf{x}_j$ if $V(\mathbf{x}_i) \geq V(\mathbf{x}_j)$ for all decision models $V(\cdot)$ compatible with the DM's evaluations on a set of reference alternatives.
- Weak possible preference relation: $\mathbf{x}_i \succsim^P \mathbf{x}_j$ if $V(\mathbf{x}_i) \geq V(\mathbf{x}_j)$ for at least one decision model $V(\cdot)$ compatible with the DM's evaluations on a set of reference alternatives.

From these basic relations preference, indifference, and incomparability relations can be built allowing the global evaluation of any alternative using the full information provided by the reference examples. The above relations can be checked through the solution of simple optimization formulations, without actually requiring the enumeration of all decision models that can be inferred from the reference

examples. This approach was also used for multicriteria classification problems [21] as well as for outranking models [10, 22] and nonadditive value models [1].

4.3 Selecting a Representative Decision Model

Having an analytic or simulation-based characterization of all compatible models (e.g., with approaches such as the ones described in the previous subsections) provides the DM with a comprehensive view of the range of possible recommendations that can be formed on the basis of a set of models implied from some decision examples. On the other hand, a single representative model is easier to use as it only requires the DM to "plug-in" the data for any alternative into a functional, relational, or symbolic model. Furthermore, the aggregation of all evaluation criteria in a single decision model enables the DM to get insight into the role of the criteria and their effect on the recommendations formulated through the model [23].

In the above context several approaches have been introduced to infer a single decision model that best represents the information provided by a reference set of alternatives. Traditional disaggregation techniques such as the family of the UTA methods [50] use post-optimality techniques based on linear programming in order to build a representative additive value function defined as an average solution of some characteristic models compatible with the DM's judgments, defined by maximizing and minimizing the criteria trade-offs. Such an averaging approach provides a proxy of the center of the feasible region.

However, given that only a very few number of corner points are identified with this heuristic post-optimality process (at most $2n$ corner points), it is clear that the average solution is only a very rough "approximation" of the center of the polyhedron. Furthermore, the optimizations performed during the post-optimality analysis may not lead to unique results. For instance, consider again the classification example discussed in Sect. 3 and its graphical illustration in Fig. 1 for the feasible set for the criteria trade-offs which are compatible with the DM's classification of the reference alternatives (Table 1). The maximization of the trade-off constant w_1 leads to corner point C, the maximization of w_2 leads to point A, whereas the maximization of w_3 (which corresponds to the minimization of $w_1 + w_2$) leads to point D. However, the minimization of the two trade-offs does not lead to uniquely defined solutions. For instance, the minimization of w_1 may lead to point A or point E, the minimization of w_2 leads either to C or D, and the minimization of w_3 (i.e., the maximization of $w_1 + w_2$) may lead to points B or C. Thus, depending on which corner solutions are obtained, different average decision models can be constructed. Table 2 lists the average criteria trade-offs corresponding to different centroid solutions. It is evident that the results vary significantly depending on the obtained post-optimality results.

A number of alternative approaches have been proposed to address the ambiguity in the results of the above post-optimality process. Beuthe and Scannella [4] presented different post-optimality criteria in an ordinal regression setting to improve

Table 2 The post-optimality approach for constructing a centroid model within the polyhedron of acceptable models for the data of Table 1

	Post-optimality solutions							
max w_1	C	C	C	C	C	C	C	C
min w_1	E	A	E	A	E	A	E	A
max w_2	A	A	A	A	A	A	A	A
min w_2	D	D	C	C	D	D	C	C
max w_3 (min $w_1 + w_2$)	D	D	D	D	D	D	D	D
min w_3 (max $w_1 + w_2$)	B	B	B	B	C	C	C	C
	Centroid solutions							
w_1	0.31	0.31	0.44	0.44	0.42	0.42	0.54	0.54
w_2	0.32	0.34	0.32	0.34	0.22	0.23	0.22	0.23
w_3	0.37	0.35	0.24	0.23	0.37	0.35	0.24	0.23

the discriminatory power of the resulting evaluating model. Similar criteria were also proposed by Doumpos and Zopounidis [12] for classification problems.

Alternative optimization formulations have also been introduced allowing the construction of robust decision models without requiring the implementation of post-optimality analyses. Following this direction, Doumpos and Zopounidis [13] presented simple modifications of traditional optimization formulations (such as the ones discussed in Sect. 2.2) on the grounds of the regularization principle which is widely used in data mining and statistical learning [57]. Experimental results on artificial data showed that new formulations can provide improved results in ordinal regression and classification problems. On the other hand, Bous et al. [5] proposed a nonlinear optimization formulation for ordinal regression problems that enables the construction of an evaluation model through the identification of the analytic center of the polyhedron form by the DM's evaluations on some reference decision instances. Despite its nonlinear character, the proposed optimization model is easy to solve with existing iterative algorithms. In a different framework, Greco et al. [23] considered the construction of a representative model through an interactive process, which is based on the grounds of preference relations inferred from the full set of models compatible with the DM's evaluations [20]. During the proposed interactive process, different targets are formulated, which can be used by the DM as criteria for specifying the most representative evaluation model.

5 Connections with Statistical Learning

5.1 Principles of Data Mining and Statistical Learning

Similarly to disaggregation analysis, statistical learning and data mining are also involved with learning from examples [25, 26]. Many advances have been made within these fields for regression, classification, and clustering problems. Recently there has been a growing interest among machine learning researchers towards preference modeling and decision-making. Some interest has also been developed by MCDA researchers on exploiting the advances in machine learning.

Hand et al. [25] define data mining as "*the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner*." Statistical learning plays an important role in the data mining process, by describing the theory that underlies the identification of such relationships and providing the necessary algorithmic techniques. According to Vapnik [56, 57] the process of learning from examples includes three main components:

1. A set X of data vectors $\mathbf{x}$ drawn independently from a probability distribution $P(\mathbf{x})$. This distribution is assumed to be unknown, thus implying that there is no control on how the data are observed [51].
2. An output y from a set Y, which is defined for every input $\mathbf{x}$ according to an unknown conditional distribution function $P(y \mid \mathbf{x})$. This implies that the relationship between the input data and the outputs is unknown.
3. A learning method (machine), which is able to assign a function $f_\beta : X \to Y$, where β are some parameters of the unknown function.

The best function f_β is the one that best approximates the actual outputs, i.e., the one that minimizes:

$$\int L[y, f_\beta(\mathbf{x})] dP(\mathbf{x}, y) \tag{8}$$

where $L[y, f_\beta(\mathbf{x})]$ is a function of the differences between the actual output y and the estimate $f_\beta(\mathbf{x})$,[1] and $P(\mathbf{x}, y) = P(\mathbf{x})P(y \mid \mathbf{x})$ is the joint probability distribution of $\mathbf{x}$ and y. However, this joint distribution is unknown and the only available information is contained in a training set of m objects $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, which are assumed to be generated independently from this unknown distribution. Thus, the objective (8) is substituted by an empirical risk estimate:

$$\frac{1}{m} \sum_{i=1}^{m} L[y_i, f_\beta(\mathbf{x}_i)] \tag{9}$$

[1] The specification of the loss function L depends on the problem under consideration. For instance, in a regression setting it may correspond to the mean squared error, whereas in a classification context it may represent the accuracy rate.

For a class of functions f_β of a given complexity, the minimization of (9) leads to the minimization of an upper bound for (8).

A comparison of (2) and (9) shows that PDA and statistical learning are concerned with similar problems from different perspectives and focus (for a discussion of the similarities and differences of the two fields see [14, 62]).

5.2 Regularization and Robustness in Learning Machines

In the context of data mining and statistical learning, robustness is a topic of fundamental importance and is directly linked to the theory in these fields. Robustness in this case has a slightly different interpretation compared to its use in MCDA. In particular, from a data mining/statistical learning perspective robustness involves the ability of a prediction model (or learning algorithm) to retain its structure and provide accurate results in cases where the learning process is based on data that contain imperfections (i.e., errors, outliers, noise, missing data, etc.). Given that the robustness of a prediction model is related to its complexity, statistical learning has been founded on a rigorous theoretical framework that connects robustness, complexity, and the empirical risk minimization approach.

The foundations of this theoretical framework are based on Tikhonov's regularization principle [54], which involves systems of linear equations of the form $\mathbf{Ax} = \mathbf{b}$. When the problem is ill-posed, such a system of equations may not have a solution and the inverse of matrix $\mathbf{A}$ may exhibit instabilities (i.e., $\mathbf{A}$ may be singular or ill-conditioned). In such cases, a numerically robust solution can be obtained through the approximate system $\mathbf{Ax} \approx \mathbf{b}$, such that the following function is minimized:

$$\|\mathbf{Ax} - \mathbf{b}\|^2 + \lambda \|\mathbf{x}\|^2 \tag{10}$$

where $\lambda > 0$ is a regularization parameter that defines the trade-off between the error term $\|\mathbf{Ax} - \mathbf{b}\|^2$ and the "size" of the solution (thus controlling the solution for changes in $\mathbf{A}$ and $\mathbf{b}$).

With the introduction of statistical learning theory Vapnik [56] developed a general framework that uses the above idea to relate the complexity and accuracy of learning machines. In particular, Vapnik showed that under a binary loss function[2], the expected error $E(\beta)$ of a decision model defined by some parameters β, is bounded (with probability $1 - \alpha$) by:

$$E(\beta) \leq E_{\text{emp}}(\beta) + \sqrt{\frac{h[\log(2m/h) + 1] - \log(\alpha/4)}{m}} \tag{11}$$

where E_{emp} is the empirical error of the model as defined by Eq. (9) and h is the Vapnik–Chervonenkis dimension, which represents the complexity of the model.

[2] Although this is not a restricted assumption, as the theory is general enough to accommodate other loss functions as well.

When the size of the training data set in relation to the complexity of the model is large (i.e., when $m/h \gg 1$), then the second term in the left-hand side of (11) decreases and the expected error is mainly defined by the empirical error. On the other hand, when $m/h \ll 1$ (i.e., the number of training observations is too low compared to the model's complexity), then the second term increases and thus becomes relevant for the expected error of the model.

This fundamental result constitutes the basis for developing decision and prediction models in classification, regression, and clustering tasks. For instance, assume a binary classification setting where a linear model $f(\mathbf{x}) = \mathbf{w}\mathbf{x} - \gamma$ should be developed to distinguish between a set of positive and negative observations. In this context, it can be shown that if the data belong in a ball of radius R, the complexity parameter h of a model with $\|\mathbf{w}\| \leq L$ (for some $L > 0$) is bounded as follows [56, 57]:

$$h \leq \min\{L^2R^2, n\} + 1 \tag{12}$$

Thus, with a training set consisting of m positive and negative observations ($y = 1$ and $y_i = -1$, respectively), the optimal model that minimizes the expected error can be obtained from the solution of the following convex quadratic program:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\sigma_i & & \\ \text{s.t.} \quad & y_i(\mathbf{w}\mathbf{x}_i - \gamma) + \sigma_i \geq 1 & & \forall i = 1,\ldots,m \\ & \sigma_i \geq 0 & & \forall i = 1,\ldots,m \\ & \mathbf{w}, \gamma \in \mathbb{R} & & \end{aligned} \tag{13}$$

The objective function of this problem is in accordance with the Tikhonov regularization function (10). In particular, the sum of classification errors $\sigma_1, \ldots, \sigma_m$ is used as a substitute for the error term $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ in (10), whereas the regularization parameter λ in (10) is set equal to $0.5/C$. The minimization of $\|\mathbf{w}\|^2$ in the objective function of the above problem corresponds to the minimization of the complexity bound (12), which in turn leads to the minimization of the second term in the error bound (11). On the other hand, the minimization of the sum of the classification errors corresponds to the minimization of the empirical error E_{emp}.

This framework is not restricted to linear models, but it also extends to nonlinear models of arbitrary complexity and it is applicable to multi-class problems [6], regression problems [9, 39], and clustering problems [2]. Similar, principles and approaches have also been used for other types of data mining models such as neural networks [17].

The development of data mining and statistical learning models with optimization with mathematical programming techniques has received much attention [43]. In this context, robust model building has been considered from the perspective of robust optimization. Bertsimas et al. [3] expressed a robust optimization model in the following general form:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}, \mathbf{u}_i) \leq \mathbf{0} \qquad \forall \mathbf{u}_i \in \mathcal{U}_i,\ i = 1, \ldots, m \\ & \mathbf{x} \in \mathbb{R} \end{aligned} \tag{14}$$

where $\mathbf{x}$ is the vector of decision variables, $\mathbf{u}_i \in \mathbb{R}^k$ are perturbation vectors associated with the uncertainty in the parameters that define the constraints, and $\mathcal{U}_i \subseteq \mathbb{R}^k$ are uncertainty sets in which the perturbations are defined (for an overview of the theory and applications of robust optimization in design problems see [36]). For instance, a robust linear program can be expressed as follows:

$$\begin{aligned} \min \quad & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{a}_i^\top \mathbf{x} \leq b_i \qquad \forall \mathbf{a}_i \in \mathcal{U}_i,\ i = 1, \ldots, m \\ & \mathbf{x} \in \mathbb{R} \end{aligned} \tag{15}$$

where the coefficients of the decision variables in the constraints take values from the uncertainty sets $\mathcal{U}_i \subseteq \mathbb{R}^n$. Thus, a constraint $\mathbf{a}_i^\top \mathbf{x} \leq b_i$ is satisfied for every $\mathbf{a}_i \in \mathcal{U}_i$ if and only if $\max_{\mathbf{a}_i \in \mathcal{U}_i} \{\mathbf{a}_i^\top \mathbf{x}\} \leq b_i$.

The framework of robust optimization has been used to develop robust decision and prediction models in the context of statistical learning. For instance, assuming that the data for observation i are subject to perturbations defined by a stochastic vector δ_i from some distribution, bounded such that $\|\delta_i\|^2 \leq \eta_i$, the constraints of problem (13) can be re-written as:

$$y_i[\mathbf{w}(\mathbf{x}_i + \delta_i) - \gamma] + \sigma_i \geq 1 \tag{16}$$

Such methodologies for developing robust learning machines have been presented in several works (see, for instance, [48, 55, 63, 65]). Caramanis et al. [8] as well as Xu and Mannor [64] provide comprehensive overviews of robust optimization in the context of statistical learning and data mining.

5.3 Applications in MCDA Disaggregation Approaches

The principles and methodologies available in the areas of data mining and statistical/machine learning have recently attracted interest for the development of enhanced approaches in MCDA. In this context, Herbrich et al. [27] explored how the modeling approach described in the previous section can be used to develop value function models in ordinal regression problems and analyzed the generalization ability of such models in relation to the value differences between alternatives in consecutive ranks.

Evgeniou et al. [15] also examined the use of the statistical learning paradigm in an ordinal regression setting. They showed that the development of a linear value function model of the form $V(x) = \mathbf{wx}$ that minimizes $\|\mathbf{w}\|^2$ leads to robust results,

as the obtained model corresponds to the center of the largest sphere that can be inscribed by preferential constraints of the form $\mathbf{w}(\mathbf{x}_i - \mathbf{x}_j) \geq 1$ for pairs of alternatives such that $\mathbf{x}_i \succ \mathbf{x}_j$.

Doumpos and Zopounidis [13] followed a similar approach for the development of additive functions using the L_1 norm for the vector of parameters. Thus, they augmented the objective function of problems (6)–(7) considering not only the error variables, but also the complexity of the resulting value function. Through this approach, they described the relationship between the accuracy of the decision model and the quality of the information provided by the reference data. Empirical analyses on ranking and classification problems showed that the new formulations provide results that best describe the DM's preferences, are more robust to changes of the reference data, and have higher generalization performance compared to existing PDA approaches. A similar approach for constructing additive value functions was also proposed by Dembczynski et al. [11] who combined a statistical learning algorithm with a decision rule approach for classification problems.

Except for functional decision models, similar approaches have also been used for relational models, which are based on pairwise comparisons between the alternatives. For instance, Waegeman et al. [62] used a kernel approach for constructed outranking decision models and showed that such an approach is general enough to accommodate (as special cases) a large class of different types of decision models, including value functions and the Choquet integral. Pahikkala et al. [42] extended this approach to intransitive decision models.

6 Conclusions and Future Perspectives

PDA techniques greatly facilitate the development of multicriteria decision aiding models, requiring the DM to provide minimal information without asking for the specification of complex technical parameters which are often not well-understood by DMs in practice. However, using such a limited amount of data should be done with care in order to derive meaningful and really useful results.

Robustness is an important issue in this context. Addressing the robustness concern enables the formulation of recommendations and results that are valid under different conditions with respect to the modeling conditions and the available data. In this chapter we discussed the main aspects of robustness in PDA techniques and provided an up-to-date overview of the different lines of research and the related advances that have been introduced in this area. We also discussed the statistical learning perspective for developing robust and accurate decision models, which has adopted a different point of view in the analysis of robustness compared to MCDA.

Despite their different philosophies, PDA and statistical learning share common features and their connections could provide further improved approaches to robust decision aiding. Future research should also focus on the further theoretical and empirical analysis of the robustness properties of PDA formulations, the introduction of meaningful measures for assessing robustness, and the development of methodologies to improve the robustness of models and solutions in decision aid.

Acknowledgements This research has been co-financed by the European Union (European Social Fund) and Greek national funds through the Operational Program "Education and Lifelong Learning."

References

1. Angilella, S., Greco, S., Matarazzo, B.: Non-additive robust ordinal regression: a multiple criteria decision model based on the Choquet integral. Eur. J. Oper. Res. **201**(1), 277–288 (2010)
2. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. J. Mach. Learn. Res. **2**, 125–137 (2001)
3. Bertsimas, D., Brown, D.B., Caramanis, C.: Theory and applications of robust optimization. SIAM Rev. **53**(3), 464–501 (2011)
4. Beuthe, M., Scannella, G.: Comparative analysis of UTA multicriteria methods. Eur. J. Oper. Res. **130**(2), 246–262 (2001)
5. Bous, B., Fortemps, Ph., Glineur, F., Pirlot, M.: ACUTA: a novel method for eliciting additive value functions on the basis of holistic preference statements. Eur. J. Oper. Res. **206**(2), 435–444 (2010)
6. Bredensteiner, E.J., Bennett, K.P.: Multicategory classification by support vector machines. Comput. Optim. Appl. **12**(1–3), 53–79 (1999)
7. Butler, J., Jia, J., Dyer, J.: Simulation techniques for the sensitivity analysis of multi-criteria decision models. Eur. J. Oper. Res. **103**(3), 531–546 (1997)
8. Caramanis, C., Mannor, S., Xu, H.: Robust optimization in machine learning. In: Sra, S., Nowozin, S., Wright, S. (eds.) Optimization for Machine Learning, pp. 369–402. MIT Press, Cambridge (2011)
9. Chu, W., Keerthi, S.S.: Support vector ordinal regression. Neural Comput. **19**(3), 792–815 (2007)
10. Corrente, S., Greco, S., Słowiński, R.: Multiple criteria hierarchy process with ELECTRE and PROMETHEE. Omega **41**(5), 820–846 (2013)
11. Dembczynski, K., Kotlowski, W., Słowiński, R.: Additive preference model with piecewise linear components resulting from dominance-based rough set approximations. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L., Zurada, J. (eds.) Artificial Intelligence and Soft Computing - ICAISC 2006. Lecture Notes in Computer Science, vol. 4029, pp. 499–508. Springer, Berlin/Heidelberg (2006)
12. Doumpos, M., Zopounidis, C.: Multicriteria Decision Aid Classification Methods. Kluwer Academic, Dordrecht (2002)
13. Doumpos, M., Zopounidis, C.: Regularized estimation for preference disaggregation in multiple criteria decision making. Comput. Optim. Appl. **38**(1), 61–80 (2007)
14. Doumpos, M., Zopounidis, C.: Preference disaggregation and statistical learning for multicriteria decision support: a review. Eur. J. Oper. Res. **209**(3), 203–214 (2011)
15. Evgeniou, T., Boussios, C., Zacharia, G.: Generalized robust conjoint estimation. Mark. Sci. **24**(3), 415–429 (2005)
16. Figueira, J.R., Greco, S., Roy, B., Słowiński, R.: ELECTRE methods: main features and recent developments. In: Zopounidis, C., Pardalos, P.M. (eds.) Handbook of Multicriteria Analysis, pp. 51–89. Springer, New York (2010)
17. Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/variance dilemma. Neural Comput. **4**(1), 1–58 (1992)
18. Georgiev, P.G., The Luc, D., Pardalos, P.M.: Robust aspects of solutions in deterministic multiple objective linear programming. Eur. J. Oper. Res. **229**(1), 29–36 (2013)
19. Greco, S., Matarazzo, B., Słowiński, R.: Rough sets theory for multicriteria decision analysis. Eur. J. Oper. Res. **129**(1), 1–47 (2001)

20. Greco, S., Mousseau, V., Słowiński, R.: Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. Eur. J. Oper. Res. **191**(2), 416–436 (2008)
21. Greco, S., Mousseau, V., Słowiński, R.: Multiple criteria sorting with a set of additive value functions. Eur. J. Oper. Res. **207**(3), 1455–1470 (2010)
22. Greco, S., Kadziński, M., Mousseau, V., Słowiński, R.: ELECTRE^{GKMS}: robust ordinal regression for outranking methods. Eur. J. Oper. Res. **214**(1), 118–135 (2011)
23. Greco, S., Kadziński, M., Słowiński, R.: Selection of a representative value function in robust multiple criteria sorting. Comput. Oper. Res. **38**(11), 1620–1637 (2011)
24. Grigoroudis, E., Siskos, Y.: Preference disaggregation for measuring and analysing customer satisfaction: the MUSA method. Eur. J. Oper. Res. **143**(1), 148–170 (2002)
25. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press, Cambridge (2001)
26. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2001)
27. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Smola, A.J., Bartlett, P.L., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 115–132. MIT Press, Cambridge (2000)
28. Hites, R., De Smet, Y., Risse, N., Salazar-Neumann, M., Vincke, P.: About the applicability of MCDA to some robustness problems. Eur. J. Oper. Res. **174**(1), 322–332 (2006)
29. Jacquet-Lagrèze, E., Siskos, Y.: Assessing a set of additive utility functions for multicriteria decision making: the UTA method. Eur. J. Oper. Res. **10**, 151–164 (1982)
30. Jacquet-Lagrèze, E., Siskos, Y.: Preference disaggregation: 20 years of MCDA experience. Eur. J. Oper. Res. **130**, 233–245 (2001)
31. Kadziński, M., Tervonen, T.: Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. Eur. J. Oper. Res. **228**(1), 69–180 (2013)
32. Kadziński, M., Tervonen, T.: Stochastic ordinal regression for multiple criteria sorting problems. Decis. Support Syst. **55**(1), 55–66 (2013)
33. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge University Press, Cambridge (1993)
34. Kouvelis, P., Yu, G.: Robust Discrete Optimization and Its Applications. Kluwer Academic, Dordrecht (1997)
35. Kroese, D.P., Taimre, T., Botev, Z.I.: Handbook of Monte Carlo Methods. Wiley, New York (2011)
36. Kurdila, A.J., Pardalos, P.M., Zabarankin, M.: Robust Optimization-Directed Design. Springer, New York (2006)
37. Lahdelma, R., Salminen, P.: SMAA-2: stochastic multicriteria acceptability analysis for group decision making. Oper. Res. **49**(3), 444–454 (2001)
38. Lovász, L., Vempala, S.: Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm. J. Comput. Syst. Sci. **72**(2), 392–417 (2006)
39. Mangasarian, O.L., Musicant, D.R.: Robust linear and support vector regression. IEEE Trans. Pattern Anal. Mach. Intell. **22**(9), 950–955 (2000)
40. Mousseau, V., Figueira, J., Naux, J.-Ph.: Using assignment examples to infer weights for ELECTRE TRI method: some experimental results. Eur. J. Oper. Res. **130**(2), 263–275 (2001)
41. Mousseau, V., Figueira, J., Dias, L., Gomes da Silva, C., Clímaco, J.: Resolving inconsistencies among constraints on the parameters of an MCDA model. Eur. J. Oper. Res. **147**(1), 72–93 (2003)
42. Pahikkala, T., Waegeman, W., Tsivtsivadze, W., De Baets, B., Salakoski, T.: Learning intransitive reciprocal relations with kernel methods. Eur. J. Oper. Res. **206**(3), 676–685 (2010)
43. Pardalos, P.M., Hansen, P.: Data Mining and Mathematical Programming. American Mathematical Society, Providence (2008)
44. Roy, B.: The outranking approach and the foundations of ELECTRE methods. Theory Decis. **31**, 49–73 (1991)
45. Roy, B.: Robustness in operational research and decision aiding: a multi-faceted issue. Eur. J. Oper. Res. **200**(3), 629–638 (2010)

46. Rubinstein, R.Y.: Generating random vectors uniformly distributed inside and on the surface of different regions. Eur. J. Oper. Res. **10**(2), 205–209 (1982)
47. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2002)
48. Shivaswamy, P.K., Bhattacharyya, C., Smola, A.J.: Second order cone programming approaches for handling missing and uncertain data. J. Mach. Learn. Res. **6**, 1283–1314 (2006)
49. Siskos, J.: A way to deal with fuzzy preferences in multicriteria decision problems. Eur. J. Oper. Res. **10**(3), 314–324 (1982)
50. Siskos, Y., Grigoroudis, E.: New trends in aggregation-disaggregation approaches. In: Zopounidis, C., Pardalos, P.M. (eds.) Handbook of Multicriteria Analysis, pp. 189–214. Springer, Berlin/Heidelberg (2010)
51. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
52. Tervonen, T., Figueira, J.R., Lahdelma, R., Dias, J.A., Salminen, P.: A stochastic method for robustness analysis in sorting problems. Eur. J. Oper. Res. **192**(1), 236–242 (2009)
53. Tervonen, T., van Valkenhoef, G., Basturk, N., Postmus, D.: Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur. J. Oper. Res. **224**(3), 552–559 (2013)
54. Tikhonov, A.N., Goncharsky, A.V., Stepanov, V.V., Yagola, A.G.: Numerical Methods for the Solution of Ill-Posed Problems. Springer, Dordrecht (1995)
55. Trafalis, T.B., Gilbert, R.C.: Robust support vector machines for classification and computational issues. Optim. Methods Softw. **22**(1), 187–198 (2007)
56. Vapnik, V.N.: An overview of statistical learning theory. IEEE Trans. Neural Netw. **10**(5), 988–999 (1999)
57. Vapnik, V.N.: The Nature of Statistical Learning Theory, 2nd edn. Springer, New York (2000)
58. Vetschera, R.: A recursive algorithm for volume-based sensitivity analysis of linear decision models. Comput. Oper. Res. **24**(4), 477–491 (1997)
59. Vetschera, R., Chen, Y., Hipel, K.W., Marc Kilgour, D.: Robustness and information levels in case-based multiple criteria sorting. Eur. J. Oper. Res. **202**(3), 841–852 (2010)
60. Vincke, Ph.: Multicriteria Decision Aid. Wiley, New York (1992)
61. Vincke, Ph.: Robust solutions and methods in decision-aid. J. Multi-Criteria Decis. Anal. **8**(3), 181–187 (1999)
62. Waegeman, W., De Baets, B., Boullart, B.: Kernel-based learning methods for preference aggregation. 4OR **7**, 169–189 (2009)
63. Xanthopoulos, P., Guarracino, M.R., Pardalos, P.M.: Robust generalized eigenvalue classifier with ellipsoidal uncertainty. Ann. Oper. Res. **216**(1), 327–342 (2014)
64. Xu, H., Mannor, S.: Robustness and generalization. Mach. Learn. **86**, 391–423 (2012)
65. Xu, H., Caramanis, C., Mannor, S.: Robustness and regularization of support vector machines. J. Mach. Learn. Res. **10**, 1485–1510 (2009)
66. Zopounidis, C., Doumpos, M.: Building additive utilities for multi-group hierarchical discrimination: the M.H.Dis method. Optim. Methods Softw. **14**(3), 219–240 (2000)
67. Zopounidis, C., Doumpos, M.: Multicriteria classification and sorting methods: a literature review. Eur. J. Oper. Res. **138**(2), 229–246 (2002)
68. Zopounidis, C., Pardalos, P.M.: Handbook of Multicriteria Analysis. Springer, Berlin/Heidelberg (2010)

Separation of Finitely Many Convex Sets and Data Pre-classification

Manlio Gaudioso, Jerzy Grzybowski, Diethard Pallaschke, and Ryszard Urbański

1 Introduction

Separation of sets has been for long time an interesting research area for mathematicians. Basic concepts of classification theory are linear separability of sets, separation margin, and kernel transformations. They have provided the theoretical background in constructing powerful classification tools such as SVM (Support Vector Machine) and extensions.

Starting from the pioneering works by Rosen [20] and Mangasarian [15, 16] and under the impulse of Vapnik's theory [25], many scientists from the Mathematical Programming community have given in recent years valuable contributions. Accurate presentations of the field can be found in the books by Cristianini and Shawe-Taylor [6, 7] and by Schölkopf et al. [23].

More recent techniques based on non-smooth optimization have been studied by Bagirov et al. [3, 4], Astorino and Gaudioso [1, 2], Demyanov et al. [8, 9], and Rubinov [21].

A different approach to the separation of two sets was proposed by Grzybowski et al. [11] and Astorino and Gaudioso [1] and Gaudioso et al. [10] which leads to a non-smooth optimization problem. It is based on the method of separating two

M. Gaudioso
Dipartimento di Elettronica, Informatica e Sistemistica (DEIS),
Universita della Calabria, 87036 Arcavacata di Rende, Italy
e-mail: gaudioso@deis.unical.it

J. Grzybowski • R. Urbański
Faculty of Mathematics and Computer Science, Adam Mickiewicz University,
Umultowska 87, 61-614 Poznań, Poland
e-mail: jgrz@amu.edu.pl; rich@amu.edu.pl

D. Pallaschke (✉)
Institute of Operations, University of Karlsruhe (KIT), Kaiserstr. 12, 76128 Karlsruhe, Germany
e-mail: diethard.pallaschke@kit.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_9,

compact convex sets by an other one. In this paper we generalize these results to the case of finitely many convex sets.

The paper is organized as follows. We begin with a survey on basic properties of the family of bounded closed convex sets in a topological vector space. Then we prove a separation theorem for closed bounded convex sets and present a generalization of the Demyanov difference in locally convex vector spaces. Finally we show an application of the separation theorem to data classification.

2 The Semigroup of Closed Bounded Convex Sets

For a Hausdorff topological vector space (X,τ) let us denote by $\mathcal{A}(X)$ the set of all nonempty subsets of X, by $\mathcal{B}^*(X)$ the set of all nonempty bounded subsets of X, by $\mathcal{C}(X)$ the set of all nonempty closed convex subsets of X, by $\mathcal{B}(X) = \mathcal{B}^*(X) \cap \mathcal{C}(X)$ the set of all bounded closed convex sets of X, and by $\mathcal{K}(X)$ the set of all nonempty compact convex subsets of X. (Note that we consider only vector spaces over the reals). Recall that for $A,B \in \mathcal{A}(X)$ the *algebraic sum* is defined by $A+B = \{x = a+b \mid a \in A \text{ and } b \in B\}$, and for $\lambda \in \mathbb{R}$ and $A \in \mathcal{A}(X)$ the *multiplication* is defined by $\lambda A = \{x = \lambda a \mid a \in A\}$.

The *Minkowski sum* for $A,B \in \mathcal{A}(X)$ is defined by

$$A \dot{+} B = \operatorname{cl}(\{x = a+b \mid a \in A \text{ and } b \in B\}),$$

where $\operatorname{cl}(A) = \bar{A}$ denotes the closure of $A \subset X$ with respect to τ. For compact convex sets, the Minkowski sum coincides with the algebraic sum, i.e., for $A,B \in \mathcal{K}(X)$ we have $A \dot{+} B = A+B$. In quasidifferential calculus of Demyanov and Rubinov [8] pairs of bounded closed convex sets are considered. More precisely: For a Hausdorff topological vector space X two pairs $(A,B),(C,D) \in \mathcal{B}^2(X) = \mathcal{B}(X) \times \mathcal{B}(X)$ are called *equivalent* if $B \dot{+} C = A \dot{+} D$ holds and $[A,B]$ denotes the equivalence class represented by the pair $(A,B) \in \mathcal{B}^2(X)$. An ordering among equivalence classes is given by $[A,B] \leq [C,D]$ if and only if $A \dot{+} D \subset B \dot{+} C$. This is the ordering on the Minkowski–Rådström–Hörmander space and is independent of the choice of the representatives.

For $A \in \mathcal{B}(X)$ we denote by $\operatorname{ext}(A)$ the set of its extreme points and by $\exp(A)$ the set of its exposed points (see [18]). Next, for $A,B \in \mathcal{A}(X)$ we define: $A \vee B = \operatorname{cl} \operatorname{conv}(A \cup B)$, where $\operatorname{conv}(A \cup B)$ denotes the convex hull of $A \cup B$. We will use the abbreviation $A \dot{+} B \vee C$ for $A \dot{+} (B \vee C)$ and $C+d$ instead of $C+\{d\}$ for all bounded closed convex sets $A,B,C \in \mathcal{A}(X)$ and a point $d \in X$.

A distributivity relation between the Minkowski sum and the maximum operation is expressed by the *Pinsker Formula* (see [19]) which is stated in a more general form in [18] as:

Proposition 1. *Let (X,τ) be a Hausdorff topological vector space, $A,B,C \in \mathcal{A}(X)$ and C be a convex set. Then*

$$(A \dot{+} C) \vee (B \dot{+} C) = C \dot{+} (A \vee B).$$

The Minkowski–Rådström–Hörmander Theorem on the cancellation property for bounded closed convex subsets in Hausdorff topological vector spaces states that for $A,B,C \in \mathcal{B}(X)$ the inclusion $A \dot{+} B \subseteq B \dot{+} C$ implies $A \subseteq C$. A generalization which is due to Urbański [24] (see also [18]) states:

Theorem 1. *Let X be a Hausdorff topological vector space. Then for any $A \in \mathcal{A}(X)$, $B \in \mathcal{B}^*(X)$, and $C \in \mathcal{C}(X)$ the inclusion*

$$A + B \subseteq C \dot{+} B \quad \textit{implies} \quad A \subseteq C. \qquad (\textbf{\textit{olc}})$$

This implies that $\mathcal{B}(X)$ endowed with the Minkowski sum " $\dot{+}$" and the ordering induced by inclusion is a commutative ordered *semigroup* (i.e., an ordered set endowed with a group operation, without having inverse elements), which satisfies the order cancellation law and contains $\mathcal{K}(X)$ as a sub-semigroup.

3 The Separation Law for Closed Bounded Convex Sets

The separation of two bounded closed convex sets by an other bounded closed set is extensively explained in [18]. In this section we discuss this separation concept more detailed and generalize it to the case of finitely many bounded closed convex sets. We begin with a general principle for the separation concepts of sets.

3.1 The Separation Concept of Martinez-Legaz and Martinón

Although any separation concept for two sets is intuitively clear, this is not so obvious for the separation of arbitrary finitely many sets. The following fundamental principle for a separation concepts had been recently formulated by Martinez-Legaz and Martinón in [17], namely

> A subset S separates a finite family of nonempty subsets $(A_i)_{i\in I}$ if S separates the sets A_i and A_j for every $i,j \in I$ with $i \neq j$.

Now a general separation concept which satisfies this principle is given by a slight modification of the definition of set separation as stated in [18, Defintion 4.5.1]:

Definition 1. Let X be a topological vector space, I a finite index set, and $S,A_i \in \mathcal{B}(X),\ i \in I$. Then we say that the set S *properly separates* the sets $A_i,\ i \in I$ if and only if for every collection $a_i \in A,\ i \in I$ there exist real numbers $0 < \alpha_i$ with $\sum_{i\in I} \alpha_i = 1$ and $\sum_{i\in I} \alpha_i a_i \in S$.

An obvious weakening of the proper separation concept leads to:

Definition 2. Let X be a topological vector space, I a finite index set, and $S, A_i \in \mathcal{B}(X),\ i \in I$. We say that the set S *separates* the sets $A_i,\ i \in I$ if and only if $(\text{conv}\{a_i \mid i \in I\}) \cap S \neq \emptyset$ for every collection $a_i \in A,\ i \in I$ (Fig. 1).

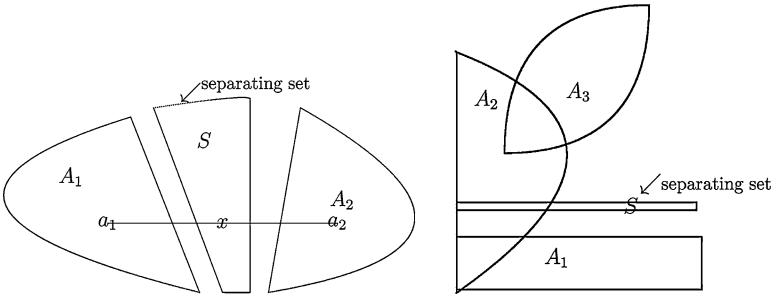


Fig. 1 Proper separation of two sets (*left*) and proper separation of three sets (*right*)

Now we prove that the concept of proper separation of sets satisfies the fundamental principle on set-separation of Martinez-Legaz and Martinón [17]:

Proposition 2. *Let X be a topological vector space, I a finite index set, and $S, A_i \in \mathcal{B}(X),\ i \in I$. If S properly separates the sets A_i and A_j for every $i, j \in I$ with $i \neq j$, then S properly separates the sets $A_i,\ i \in I$.*

Proof. Let $S, A_i \in \mathcal{B}(X),\ i \in I$ be given, where I consists of k elements and assume that S properly separates all pairs of sets A_i and A_j with $i, j \in I$ and $i \neq j$. Then there exist for every $a_i \in A_i$ and $a_j \in A_j$ real numbers $\alpha_{ij} > 0$ with $z_{ij} = \alpha_{ij} a_i + (1 - \alpha_{ij}) a_j \in S$. Put $\sigma = \frac{1}{\binom{k}{2}}$ then the convex combination $\sigma \sum_{\substack{i,j \in I \\ i \neq j}} z_{ij} \in S$ has only nonzero coefficients, i.e., S properly separates the sets $A_i,\ i \in I$.

3.2 The Algebraic Separation Law

We will use the notation $\bigvee_{i \in I} \{a_i\}$ for $\text{conv}\{a_i \mid i \in I\}$ and write

$$\sum_{i=1}^{k} A_i \ = \ A_1 \dot{+} A_2 \dot{+} \cdots \dot{+} A_k.$$

For the weaker concept of separation we have the following algebraic characterization:

Theorem 2. *Let X be a topological vector space, I a finite index set, and $S, A_i \in \mathcal{B}(X)$, $i \in I$. Then S separates the sets A_i, $i \in I$ if and only if*

$$\sum_{i\in I} A_i \subset \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} S.$$

Proof. Necessity: Let $a_i \in A_i$, $i \in I$ be given. Then there exist $\alpha_i \geq 0, \sum_{i\in I}\alpha_i = 1$ such that $\sum_{i\in I}\alpha_i a_i \in S$. Therefore,

$$\begin{aligned}\sum_{i\in I} a_i &= \sum_{i\in I}\left(\sum_{k\in I\setminus\{i\}}\alpha_k\right) a_i + \sum_{i\in I}\alpha_i a_i \\ &= \sum_{i\in I}\alpha_i\left(\sum_{k\in I\setminus\{i\}} a_k\right) + \sum_{i\in I}\alpha_i a_i \\ &\in \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} S,\end{aligned}$$

which proves the necessity.

Sufficiency: Now fix any $a_i \in A_i$, $i \in I$. Then it follows from the assumption

$$\sum_{i\in I} A_i \subset \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} S$$

that for every $i \in I$

$$a_i + \sum_{k\in I\setminus\{i\}} A_k \subset \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} S,$$

which means:

$$\sum_{k\in I\setminus\{i\}} A_k \subset \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} (S - a_i), \quad i \in I.$$

From the Pinsker rule we get:

$$\begin{aligned}\bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) &\subset \bigvee_{i\in I}\left[\bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} (S - a_i)\right] \\ &= \bigvee_{i\in I}\left(\sum_{k\in I\setminus\{i\}} A_k\right) \dot{+} \bigvee_{i\in I}(S - a_i)\end{aligned}$$

and gives by the order cancellation law that $0 \in \bigvee_{i\in I}(S - a_i)$.

Now again by the Pinsker rule we get $0 \in \bigvee_{i\in I}(S-a_i) = S \dot{+} \bigvee_{i\in I}\{-a_i\}$, which implies that $(\operatorname{conv}\{a_i \mid i \in I\}) \cap S \neq \emptyset$.

Remark 1. For the sake of completeness, let us add the following two items:

- Parallel to the notation of *separation* the notation of *shadowing* is also used (see [18, pp. 67 and 77]) to express the same property. Namely the physical interpretation of the *separation by sets* is as follows: *if the sets A, B, S are considered as celestial and A shines, then S separates A and B if and only if B lies in the shadow of S* (see Fig. 2).
- In [12] the following equivalence is proved:
  Let X be a topological vector space, I a finite index set, and $S, A_i \in \mathcal{B}(X)$, $i \in I$. Then S separates the sets A_i, $i \in I$ if and only if $\inf_{i\in I}[A_i, 0] \leq [S, 0]$ *in the sense of the ordering among equivalence classes in the Minkowski–Rådström–Hörmander space.*

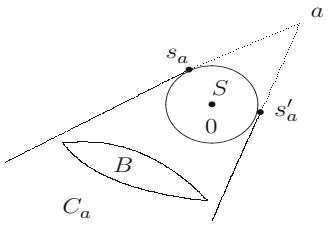


Fig. 2 Illustration to Remark 1

4 The Demyanov Difference

Demyanov original subtraction $A \ddot{-} B$ (see [22]) of compact convex subsets in finite dimensional space is defined with the help of the Clarke subdifferential (see [5]) of the difference of support functions, i.e.,

$$A \ddot{-} B = \partial_{\mathrm{cl}}(p_A - p_B)\Big|_0,$$

where p_A and p_B are the support functions of A and B, i.e., $p_A(x) = \max_{a\in A} \langle a, x\rangle$

This can be equivalently formulated by

$$A \ddot{-} B = \overline{\operatorname{conv}}\{a - b | a \in A, b \in B, a + b \in \exp(A + B)\},$$

where $\exp(A + B)$ are the exposed points of $A + B$. For the proof see [22] and note that every exposed point of $A + B$ is the unique sum of an exposed point of A with an exposed point of B.

To extend the definition of the difference $A \ddot{-} B$ to locally convex vector spaces, the set of exposed points will be replaced by the set of extremal points.

Definition 3. Let (X, τ) be a locally convex vector space and $\mathcal{K}(X)$ the family of all nonempty compact convex subsets of X. Then for $A, B \in \mathcal{K}(X)$, the set

$$A \ddot{-} B = \overline{\text{conv}}\{a - b | a \in A, b \in B, a + b \in \text{ ext}(A + B)\} \in \mathcal{K}(X)$$

is called the *Demyanov difference* of A and B.

This is a canonical generalization of the above definition, because for every $A, B \in \mathcal{K}(X)$ every extremal point $z \in \text{ ext}(A + B)$ has a unique decomposition $z = x + y$ into the sum of two extreme points $x \in \text{ ext}(A)$ and $y \in \text{ ext}(B)$ (see [14, Proposition 1]).

Since in the finite dimensional case the exposed points are dense in the set of extreme points of a compact convex set, this definition coincides with the original definition of the Demyanov difference in finite dimensional spaces.

The Demyanov difference in finite dimensional spaces possesses many important properties. Some of them hold also for its generalization (see [13]):

Proposition 3. *Let X be a locally convex vector space and $A, B, C \in \mathcal{K}(X)$. The Demyanov difference has the following properties:*

(D1) *If $A = B + C$, then $C = A \ddot{-} B$.*
(D2) $(A \ddot{-} B) + B \supseteq A$.
(D3) *If $B \subseteq A$, then $0 \in A \ddot{-} B$.*
(D4) $(A \ddot{-} B) = -(B \ddot{-} A)$.
(D5) $A \ddot{-} C \subset (A \ddot{-} B) + (B \ddot{-} C)$.

From property (D2) of the above proposition follows immediately:

Theorem 3. *Let X be a locally convex vector space, I a finite index set, and $S, A_i \in \mathcal{K}(X)$, $i \in I$. Then the Demyanov difference*

$$S = \left(\sum_{i \in I} A_i\right) \ddot{-} \bigvee_{i \in I} \left(\sum_{k \in I \setminus \{i\}} A_k\right)$$

separates the sets A_i, $i \in I$.

Corollary 1. *Let $A_1, A_2, \ldots, A_k \in \mathcal{K}(\mathbb{R}^n)$ be given. Then for the Demyanov difference holds*

$$\left(\sum_{i=1}^{k} A_i\right) \ddot{-} \bigvee_{i=1}^{k} \left(\sum_{\substack{j=1 \\ j \neq i}}^{k} A_j\right) = \partial_{cl} P\big|_0,$$

where $\partial_{cl} P\big|_0$ is the Clarke subdifferential of $P = \min\left\{p_{A_1}, p_{A_1}, \ldots, p_{A_k}\right\}$, at $0 \in \mathbb{R}^n$, i.e., the minimum of the support functions of the sets A_i.

Proof. This follows immediately from the definition of the Demyanov difference for the finite dimensional case (see [22]) and the formula

$$\left(\sum_{i=1}^{k} p_{A_1}\right) - \max\left\{\sum_{\substack{j=1\\ j\neq i}}^{k} p_{A_1} \mid i \in \{1,\ldots,k\}\right\} = \min\left\{p_{A_1}, p_{A_1}, \ldots, p_{A_k}\right\},$$

which completes the proof.

5 Data Pre-classification

The separation law for finitely many convex sets gives a possibility to classify the elements of $A_1, A_2, \ldots, A_k \in \mathcal{K}(\mathbb{R}^n)$ by $(k+1)$ different types as:

- TYPE 1) $----$ the elements of the set $A_i \setminus S$ for $i \in \{1,..,k\}$,
- TYPE $(k+1)$) $----$ the set $S \cap \left(\bigcup_{i=1}^{k} A_i\right)$,

where $S \in \mathcal{K}(\mathbb{R}^n)$ is a separating set for $A_1, A_2, \ldots, A_k \in \mathcal{K}(\mathbb{R}^n)$

This is illustrated in Fig. 3 for $k = 3$ and $n = 2$.

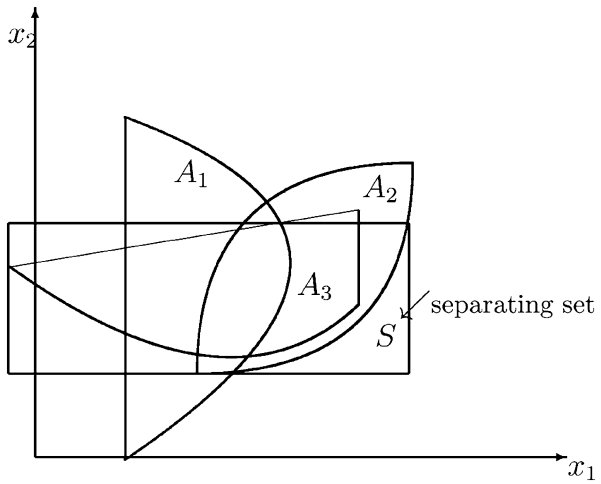


Fig. 3 A separating rectangle S of minimal volume for the sets A_1, A_2, and A_3

Now, by the separation law $S \in \mathcal{K}(\mathbb{R}^n)$ separates the set $A_1, A_2, \ldots, A_k \in \mathcal{K}(\mathbb{R}^n)$ if and only if

$$\sum_{i=1}^{k} A_i \subset \bigvee_{i=1}^{k} \left(\sum_{\substack{j=1 \\ j \neq i}}^{k} A_j \right) \dot{+} S$$

holds, which is in finite dimension equivalent to

$$\left(\sum_{i=1}^{k} p_{A_1} \right) - \max \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{k} p_{A_1} \mid i \in \{1, \ldots, k\} \right\} \leq p_S$$

and finally equivalent to

$$\min \left\{ p_{A_1}, p_{A_1}, \ldots, p_{A_k} \right\} \leq p_S. \qquad (*)$$

For the case of polytopes, this gives a possibility of constructing separating sets with the help of convex optimization problems, as for instance:

Let us assume that $A_1, A_2, \ldots, A_k \in \mathcal{K}(\mathbb{R}^n)$ are polytopes given by

$$A_r = \operatorname{conv} \left\{ a_1^r, \ldots, a_{l_r}^r \right\}, r \in \{1, \ldots, k\}$$

and that we are looking for a separating set of the form $S = \operatorname{conv} \left\{ s_1, \ldots, s_p \right\}$.

Since $p_{A_r}(x) = \max_{1 \leq j \leq l_k} \langle a_j^k, x \rangle$, $r \in \{1, .., k\}$ and $p_{S(x)} = \max_{1 \leq j \leq p} \langle s_j, x \rangle$ equation $(^*)$ implies, that for every point $x \in \mathbb{R}^n$ the following inequality hold:

$$\min \left\{ \max_{1 \leq j \leq l_1} \langle a_j^1, x \rangle, \max_{1 \leq j \leq l_2} \langle a_j^2, x \rangle, \ldots, \max_{1 \leq j \leq l_k} \langle a_j^k, x \rangle \right\} \leq \max_{1 \leq j \leq p} \langle s_j, x \rangle$$

holds.

One way to construct a convex optimization problem consists in finding a suitable finite point set $T \subset \mathbb{R}^n$ called a *test set* for the constraints. Then for the determination of a separating set S with minimal volume the following optimization problem can be used:

$$\min \operatorname{Vol}(S) = \Phi(s_1, \ldots, s_r)$$

under

$$\min \left\{ \max_{1 \leq j \leq l_1} \langle a_j^1, x \rangle, \max_{1 \leq j \leq l_2} \langle a_j^2, x \rangle, \ldots, \max_{1 \leq j \leq l_k} \langle a_j^k, x \rangle \right\} \leq \max_{1 \leq j \leq p} \langle s_j, x \rangle, \; x \in T.$$

References

1. Astorino, A., Gaudioso, M.: Polyhedral separability through successive LP. J. Optim. Theory **112**, 265–293 (2002)
2. Astorino, A., Gaudioso, M., Gorgone, E., Pallaschke, D.: Data preprocessing in semi-supervised SVM classification. Optimization **60**, 143–151 (2011)
3. Bagirov, A.M., Ferguson, B., Ivkovic, S., Saunders, G., Yearwood, J.: New algorithms for multi-class cancer diagnosis using tumor gene expression signatures. Bioinformatics **19**, 1800–1807 (2003)
4. Bagirov, A.M., Karasözen, B., Sezer, M.: Discrete gradient method: derivative-free method for nonsmooth optimization. J. Optim. Theory **137**, 317–334 (2008)
5. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
6. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
7. Cristianini, N., Shawe-Taylor, J.: Kernel Models for Pattern Analysis. Cambridge University Press, Cambridge (2004)
8. Demyanov, V.F., Rubinov, A.M.: Quasidifferential Calculus. Optimization Software Inc., Publications Division, New York (1986)
9. Demyanov, V.F., Astorino, A., Gaudioso, M.: Nonsmooth problems in mathematical diagnostics. In: Advances in Convex Analysis and Global Optimization (Pythagorion, 2000). Nonconvex Optimization and Applications, vol. 54, pp. 11–30. Kluwer Academic, Dordrecht (2001)
10. Gaudioso, M., Gorgone, E., Pallaschke, D.: Separation of convex sets by Clarke subdifferential. Optimization **59**, 1199–1210 (2011)
11. Grzybowski, J., Pallaschke, D., Urbański, R.: Data pre-classification and the separation law for closed bounded convex sets. Optim. Methods Softw. **20**, 219–229 (2005)
12. Grzybowski, J., Pallaschke, D., Urbański, R.: Reduction of finite exhausters. J. Glob. Optim. **46**, 589–601 (2010)
13. Grzybowski, J., Pallaschke, D., Urbánski, R.: Demyanov difference in infinite dimensional spaces. In: Demyanov, V.F., Pardalos, P.M., Batsyn, M. (eds.) Proceedings of the Conference on Constructive Nonsmooth Analysis and Related Topics. Optimization and Its Applications, vol. 87, pp. 13–24. Springer, New York (2013)
14. Husain, T., Tweddle, I.: On the extreme points of the sum of two compact convex sets. Math. Ann. **188**, 113–122 (1970)
15. Mangasarian, O.L.: Linear and nonlinear separation of patterns by linear programming. Oper. Res. **13**, 444–452 (1965)
16. Mangasarian, O.L.: Multi-surface method of pattern separation. IEEE Trans. Inf. Theory **IT-14**, 801–807 (1968)
17. Martinez-Legaz, J.-E., Martinón, A.: On the infimum of a quasiconvex vector function over an intersection. TOP **20**, 503–516 (2012)
18. Pallaschke, D., Urbański, R.: Pairs of Compact Convex Sets—Fractional Arithmetic with Convex Sets. Mathematics and Its Applications, vol. 548. Kluwer Academic, Dordrecht (2002)
19. Pinsker, A.G.: The space of convex sets of a locally convex space. Trudy Leningrad Eng. Econ. Inst. **63**, 13–17 (1966)
20. Rosen, J.B.: Pattern separation by convex programming. J. Math. Anal. Appl. **10**, 123–134 (1965)
21. Rubinov, A.M.: Abstract convexity, global optimization and data classification. Opsearch **38**, 247–265 (2001)
22. Rubinov, A.M., Akhundov, I.S.: Differences of compact sets in the sense of Demyanov and its application to non-smooth-analysis. Optimization **23**, 179–189 (1992)
23. Schölkopf, B., Burges, C.J.C., Smola, A.J.: Advances in Kernel Methods: Support Vector Learning. The MIT Press, Cambridge (1999)
24. Urbański, R.: A generalization of the Minkowski-Rådström-Hörmander theorem. Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. **24**, 709–715 (1976)
25. Vapnik, V.: The Nature of the Statistical Learning Theory. Springer, New York (1995)

The Shortest Superstring Problem

Theodoros P. Gevezes and Leonidas S. Pitsoulis

An **alphabet** is a finite non-empty set whose elements are called **letters**. A **string** is a sequence of letters. Given two strings s_i and s_j, the second is a **substring** of the first if s_i contains consecutive letters that match s_j exactly. We say that s_i is a **superstring** of s_j. The **Shortest** (**common**) **Superstring Problem** (SSP) is a combinatorial optimization problem that consists in finding a shortest string which contains as substrings all of the strings in a given set. The strings of the set may be overlapping inside the superstring exploiting their common data.

1 Applications

The SSP has several important applications in various scientific domains and this is the reason why it has attracted the interest of many researchers. In computational molecular biology, the DNA sequencing procedure via fragment assembly can be formulated as SSP. In virology, the SSP models the compression of viral genome. In information technology, the SSP can be used to achieve data compression. In scheduling, SSP solutions can be used to schedule operations in machines with coordinated starting times. In the field of data structures, efficient storage can be achieved in specific cases using the solutions of the SSP.

T.P. Gevezes (✉) • L.S. Pitsoulis
Faculty of Engineering, School of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
e-mail: theogev@gen.auth.gr; pitsouli@gen.auth.gr

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_10,

1.1 DNA Sequencing

The molecule of the DNA encodes the genetic information used in the developing and functioning of living beings. DNA is a double-stranded sequence of four types of nucleotides: adenine, cytosine, guanine and thymine, and thereby it can be viewed as a string over the alphabet $\{a,c,g,t\}$. In the field of molecular biology, the DNA sequencing procedure determines the sequence of a DNA molecule, that is the precise order of the nucleotides within it. DNA sequencing highly accelerates biological and medical research.

Due to laboratory equipment constraints, only parts of DNA up to few hundred nucleotides can be read reliably, while the length of the DNA molecule in many species is quite longer. To recognize a long DNA sequence, many copies of the DNA molecule are made and cut into smaller overlapping pieces, named **fragments**, that can be read at once. Each fragment is chosen from an unknown location of the molecule. To reconstruct the initial DNA molecule, these fragments must be reassembled in their initial order, a procedure known as the **DNA assembly problem**. Due to the huge amount of data generated by the fragment sequencing methods, an automated procedure supported by a computer software is necessary for the assembly process. Intuitively, shortest superstrings of the sequenced fragments preserve important biological structures [33, 46, 53], and in practice they are proved to be good representations of the original DNA molecule [27, 34]. Therefore, the SSP can be considered as an abstraction of the assembly problem, and consequently many researchers developed assembly methods based on it [18, 45, 51]. The most widely used of them, the **shotgun sequencing**, is essentially the natural greedy algorithm for the SSP. Similar assembly problems arise during reconstruction of RNA molecules or proteins from sequenced fragments.

1.2 Data Compression

In the fields of computer science, information technology and data transmission, a crucial issue is the size of the stored or transferred data. Data compression is the process of encoding data using fewer bits than their original representation. According to whether the compressed data is exactly as the original data or not, we distinguish the **lossless compression** and the **lossy compression**, respectively (see [50]).

Considering data as text over an alphabet, an intuitive method of lossless compression is based on the idea of dividing the text into strings and representing it by a superstring of these strings with pointers to their original positions. Based on this principle, several macro schemes concerning the nature of the pointers are taken under consideration in [55, 56], leading in general applications of the SSP in the field of textual substitution. In programming languages, each alphanumerical string in the code may be represented as a pointer to a common string stored in the memory. Therefore, the target of the compiler is to arrange the alphanumerical strings in such a way that they overlap as much as possible [15, 37]. Other general applications of the SSP on data compression are discussed in [14, 54].

1.3 Modelling the Viral Genome Compression

Viruses are forced to reduce their genome size by environmental factors such as the need for quick replication and the small amount of nucleic acid that can be incorporated in them. One way to compress their genome is by overlapping their genes. Genes are the parts of the DNA that specify all proteins in living beings. Between genes there are generally long sequences of nucleotides that do not be coded into proteins. On the other hand, overlapping genes are common in viruses. Therefore, in most virus species, two or more proteins are coded by the same nucleotide sequence, allowing viruses to increase their repertoire of proteins without increasing their genome length, as indicated in [10].

In [24, 25], the SSP is used to model the viral genome compression. The genes are considered as strings and the purpose is to find a shortest superstring that contains them all. The computational results show that the amount of compression achieved by the viruses in the real world is the same or very close to the one obtained by the algorithms in all the examples considered in [24, 25]. Another conclusion from these computations is that the average compression ratio of viruses is remarkably high considering the fact that the DNA molecules are very difficult to compress in general. Finally, by modelling the viral genome compression as SSP, any exact solution or lower bound of the corresponding SSP instance provides a bound on the real size of a viral genome with a given set of genes.

1.4 Scheduling with Coordinated Starting Times

The **Flow Shop Problem** (FSP) and the **Open Shop Problem** (OSP) concern the scheduling of operations in machines and have particular applications in scheduling and planning of experiments. Given a set of k machines $M_1, M_2, \ldots, M_k$ the problem is to schedule a set of jobs on them, where each job consists of k operations and the i-th operation has to be assigned on the machine M_i. A machine can process at most one operation at a time, and any two operations of a job cannot be processed simultaneously. In the FSP, the operation on M_i has to be finished before the operation on M_{i+1} can start for each job, whereas in the OSP there is not such commitment. In the *no-wait* versions of these problems, it is required the operations of a job to be processed directly one after the other. The optional constraint of *coordinated starting times* necessitates an operation starting on one machine only when each of the other machines is either idle or also starts an operation. In all these cases, the task is to find a schedule such that the overall processing time is minimized.

The FSP and the OSP on two machines, and their corresponding no-wait versions are polynomially solvable in general, but this is not always true when the machines have to coordinate the starting times of operations. In [39], this additional constraint is considered. Each instance of the no-wait version of these problems under the additional constraint of the coordinated starting times can be transformed into an SSP instance, where all strings are of a special form. The NP-completeness of these

problem versions is proved using this transformation. Apart from the computational complexity, this transformation can be applied for solving the constrained FSP and OSP. Each exact and heuristic algorithm for the SSP can be applied to these problems too. Also, the special case of the SSP can be used to derive approximation algorithms for the constrained shop problems.

1.5 Data Structure Storage

In [15], a special case of the SSP is considered, where all the strings are of length at most two. It is proved that this version of the SSP is solvable in polynomial time. This SSP case has applications to the storage of data structures, and specifically to the **Huffman trees** [23] that encode pairs of letters, which are used to an entropy encoding algorithm for lossless data compression, and for efficient representation of directed graphs in memory.

2 Definitions and Notations

Let $\mathbb{N}$ be the set of natural numbers including 0. All the numbers in this chapter are natural, unless otherwise stated. For a real x, $\lceil x \rceil$ denotes the smaller integer greater than or equal to x. For a letter l, the notation $l \in \Sigma$ means that l belongs to alphabet Σ, while for a string s, if all letters of s belong to Σ, we say that s is *over* the alphabet Σ. If s is a string, then $|s|$ denotes its length, that is the number of its letters, while if S is a set, then $|S|$ denotes its cardinality. For a string s and $i, j \in \mathbb{N}$ such that $1 \leq i \leq j \leq |s|$, the substring of s from i-th to j-th letter is denoted by $s_{[i,j]}$. Any substring $s_{[1,j]}$ is a **prefix** of s, and if $j < |s|$, then it is called a **proper prefix**. Similarly, any substring $s_{[i,|s|]}$ is a **suffix** of s, and if $i > 1$, then it is called a **proper suffix**.

The placement of two or more strings one next to the other denotes their **concatenation**, e.g. $s_i s_j$ is the concatenation of s_i and s_j. A **coverage string** between strings s_i and s_j, in this specific order, is a string v such that $s_i = uv$ and $s_j = vw$, for some non-empty strings u, w. In other words, v is a string that is a proper suffix of s_i *and* a proper prefix of s_j. The length of the coverage string is called **coverage** between the corresponding strings and is a non-negative integer. A **join string** of s_i and s_j is the concatenation of these two strings with a coverage string appearing only once, that is uvw. We use $J_{\{s_i,s_j\}}$ to denote the set of all join strings of s_i and s_j *regardless* their order.

The **overlap string** between s_i and s_j is their longest coverage string, and is denoted by $o(s_i, s_j)$. Its length $|o(s_i, s_j)|$ is called **overlap**. The overlap of a string with itself is called **self-overlap**, and notice that it is not limited to half the total string length. The **merge string** of s_i and s_j is the concatenation of these two strings with the overlap string appearing only once, that is the shortest join string between

them. It is denoted by $m(s_i, s_j)$. We have $|m(s_i, s_j)| = |s_i| + |s_j| - |o(s_i, s_j)|$. The length of the prefix of s_i before the overlap string with s_j is called **distance** from s_i to s_j and is denoted by $d(s_i, s_j)$.

Example 1. Suppose that we have the strings $s_1 = bbacb$ and $s_2 = bcabbcabb$ over the alphabet $\{a,b,c\}$, so $|s_1| = 5$ and $|s_2| = 9$. The one-letter string b is a proper suffix of s_1 and a proper prefix of s_2. Moreover, it is the longest such string, and thus the overlap string between them, $o(s_1, s_2) = b$, with overlap 1. The coverage strings between s_2 and s_1 are b and bb, and so $o(s_2, s_1) = bb$ with overlap 2. The self-overlap of the first string is $|o(s_1, s_1)| = 1$, while $|o(s_2, s_2)| = 5$. The corresponding merge strings are $m(s_1, s_2) = bbacbcabbcabb$, $m(s_2, s_1) = bcabbcabbacb$, $m(s_1, s_1) = bbacbbacb$, and $m(s_2, s_2) = bcabbcabbcabb$. The distance from s_1 to s_2 is $d(s_1, s_2) = 4$, while $d(s_2, s_1) = 7$, $d(s_1, s_1) = 4$, and $d(s_2, s_2) = 4$. Finally, the set of all join strings is $J_{\{s_1, s_2\}} = \{bbacbbcabbcabb, bcabbcabbbbacb, bbacbcabbcabb, bcabbcabbbacb, bcabbcabbacb\}$. □

Given a finite set S of strings over an alphabet Σ, the sum of lengths of the strings in S is defined as $||S|| = \sum_{s \in S} |s|$. The **orbit size** of a letter $l \in \Sigma$ is the number of its occurrences in the strings of S.

An instance of the SSP is specified by a finite set $S = \{s_1, s_2, \ldots, s_n\}$ of strings. A string s is a superstring of S, if it is a superstring of all $s_i \in S$. A **multiset** is a generalization of the notion of the set where elements are allowed to appear more than once. Without loss of generality, S is defined to be a set since if S is a multiset, then S has exactly the same superstrings as the set $\{s : s \in S\}$. Also, it is assumed that S is a **substring-free set**, i.e., no string $s_i \in S$ is a substring of any *other* string $s_j \in S$. This assumption can be made without loss of generality, since for any set of strings there exists a unique substring-free set that has the same superstrings, obtained by removing any string is a substring of another.

Given a set $S = \{s_1, s_2, \ldots, s_n\}$ of strings over an alphabet Σ, the SSP is the problem of finding a minimum length superstring of S. Note that such a string may not be unique. The length of a shortest superstring of S is denoted by $\mathrm{opt}_l(S)$, while the corresponding achieved **compression** is defined as $\mathrm{opt}_c(S) = ||S|| - \mathrm{opt}_l(S)$. The decision version of the SSP is described as follows. Given a set S of strings and a $k \in \mathbb{N}$, is there a superstring s of S such that $|s| = k$?

Example 2. Suppose that we have the multiset $S' = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ of strings over the alphabet $\{a,b,c\}$, where $s_1 = bababbc$, $s_2 = bbccaac$, $s_3 = bbcaaabb$, $s_4 = acabb$, $s_5 = bcaaab$, and $s_6 = acabb$. The corresponding substring-free set is $S = \{s_1, s_2, s_3, s_4\}$ with $|S| = 4$ and $||S|| = 27$. The orbit size of the letter a in S is 9, of the letter b is 12, and of the letter c is 6. These are the two shortest superstrings of S: $s = bababbccaacabbcaaabb$ and $s' = bababbcaaabbccaacabb$, with $\mathrm{opt}_l(S) = |s| = |s'| = 19$ and $\mathrm{opt}_c(S) = 7$. □

Let I_n be the finite set $\{1, 2, \ldots, n\}$, and Π_n be the set of all permutations of the set I_n. Any solution for the SSP of n strings can be represented as a permutation $p \in \Pi_n$, indicating the order in which strings must be merged to get the superstring.

It is implied that the shortest superstrings are derived only by string merges. If this is not the case, there would be parts of the superstring that do not correspond to any string, or some consecutive strings would not exploit their longest coverage string and could be joined by a larger coverage. In both cases there would be a shorter superstring. The elements of a permutation $p \in \Pi_n$ are denoted by $p(i)$, $i \in \mathrm{I}_n$, where i indicates the order of each element in p such that $p = (p(1), p(2), \ldots, p(n))$.

Given an order of strings $(s_1, s_2, \ldots, s_n)$ the superstring $s = \langle s_1, \ldots, s_n \rangle$ is defined to be the string $m(s_1, m(s_2, \ldots m(s_{n-1}, s_n)\ldots))$. In such an order, the first string s_1 is denoted by first(s) and the last string s_n is denoted by last(s). Notice that s is the shortest string such that $s_1, s_2, \ldots, s_n$ appear in this order as substrings.

For a set $S = \{s_1, s_2, \ldots, s_n\}$ of strings and a permutation $p \in \Pi_n$, the corresponding superstring is defined as $strSp(S,p) = \langle s_{p(1)}, s_{p(2)}, \ldots, s_{p(n)} \rangle$. For any SSP instance $S = \{s_1, s_2, \ldots, s_n\}$, there exists a permutation $p \in \Pi_n$ such that $strSp(S,p)$ is an optimal solution. For any $p \in \Pi_n$, the length of the superstring $strSp(S,p)$ is given by $|strSp(S,p)| = \sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1} |o(s_{p(i)}, s_{p(i+1)})|$. Therefore, the SSP can be formulated as

$$\min_{p \in \Pi_n} \sum_{i=1}^{n} |s_i| - \sum_{i=1}^{n-1} |o(s_{p(i)}, s_{p(i+1)})|. \tag{1}$$

The shortest superstrings that correspond to the permutation p of the optimal solution have length equal to $\mathrm{opt}_l(S)$. A superstring of the minimum length is achieved when the sum of the overlaps between consecutive strings, in the order defined by p, is maximized.

There are two ways to assess the solution quality of a non-exact algorithm for the SSP: the **length measure** and the **overlap** or **compression measure**. According to the first measure, a superstring is better when its length is shorter. In this case, the SSP is described as a minimization problem. According to the second measure, a superstring is better when the achieved compression is greater. In this case, the problem is described as a maximization problem. The two measures are equivalent when applied to exact solutions, but they give different results when they measure the relative preciseness of non-exact solutions obtained by approximation or heuristic algorithms. A good algorithm with respect to one of the above measures is not necessarily a good algorithm with respect to the other measure.

Example 3. For the substring-free set $S = \{s_1, s_2, s_3, s_4\}$ of Example 2 and the two shortest superstrings of it, $s = \langle s_1, s_2, s_4, s_3 \rangle$ and $s' = \langle s_1, s_3, s_2, s_4 \rangle$, we have first$(s)$ = first(s') = s_1, last(s) = s_3, and last(s') = s_4.

Let $s'' = strSp(S,p)$ for the permutation $p = (3,4,2,1)$, which is a superstring of length $|s''| = 24$. According to the length measure the solution s'' is $(|s''| - \mathrm{opt}_l(S))/\mathrm{opt}_l(S) = 26.3\,\%$ far from the optimal length, while according to the compression measure is $(\mathrm{opt}_c(S) - (||S|| - |s''|))/\mathrm{opt}_c(S) = 71.4\,\%$ far from the optimal compression. □

A **directed graph** G is defined by a vertex set $V(G)$ and an arc set $E(G)$ which contains ordered pairs of vertices and is denoted by $G = (V,E)$. For an arc $e = (u,v)$, u is called the **tail** of e, and v the **head** of e. We say that e is **incident** to both vertices,

while for v the arc e is an **incoming arc**, and for u is an **outgoing arc**. An arc with the same tail and head is called a **loop**. For a vertex $v \in V$, the number of incoming arcs of v is denoted by $\deg^-(v)$, and the number of outgoing arcs of v is denoted by $\deg^+(v)$. The overall number of the incident arcs to a vertex v regardless of their direction is the **degree** of v. The degree of a graph is the maximum degree between its vertices. Graph G is **complete** if there is an arc (u,v) for any vertex pair $u,v \in V$, $u \neq v$. For a weight function $w : E \to \mathbb{N}$, we denote by $G = (V,E,w)$ a weighted directed graph. When there is no confusion we denote by w_{ij} the weight of arc $e = (v_i,v_j) \in E$. If the elements of set E have no direction, then they are called **edges** and the corresponding graph is called **undirected**. An undirected graph is called **bipartite** if its vertex set can be partitioned into two subsets, V_1 and V_2, such that every edge is incident to a vertex of V_1 and to a vertex of V_2. If the arc set contains ordered tuples instead of pairs of vertices then we have a **multigraph**.

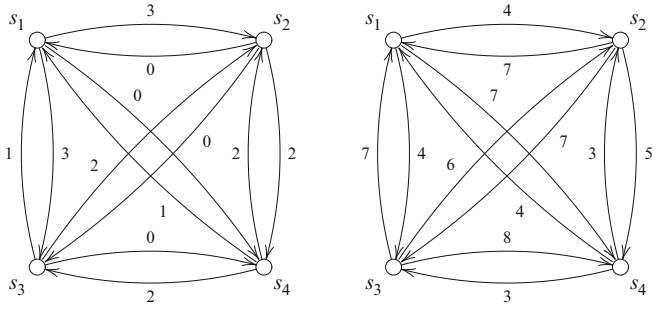
Given a set $S = \{s_1,s_2,\ldots,s_n\}$ of strings, the complete directed weighted graph $G = (V,E,w)$ with

- vertex set $V = \{s_1,s_2,\ldots,s_n\}$,
- arc set $E = \{(s_i,s_j) : s_i,s_j \in V, i \neq j\}$, and
- weight function $w : E \to \mathbb{N}$, with $w_{ij} = |o(s_i,s_j)|$,

is called the **overlap graph** of S and is denoted by $G_o(S)$. If the arc weight function depends on the distance instead of the overlap between the string pairs, that is $w_{ij} = d(s_i,s_j)$, then the corresponding graph is called the **distance graph** of S, and is denoted by $G_d(S)$. Notice that all weights on both graphs are non-negative integers. In the following, it is assumed that the overlap and distance graphs have no loops, unless otherwise stated. For any set $A \subseteq E$ of arcs on both graph, we denote by $o(A)$ the sum of weights of the arcs on $G_o(S)$, that is their total overlap, and by $d(A)$ the sum of weights of the arcs on $G_d(S)$, that is their total distance. For each arc $e = (s,s')$ on both graphs, we have

$$|s| = o(\{e\}) + d(\{e\}). \tag{2}$$

Example 4. For the substring-free set $S = \{s_1,s_2,s_3,s_4\}$ of Example 2 the associated overlap and distance graphs are depicted in the next figure.



$G_o(S)$ $G_d(S)$

For the arc set $A = \{(s_1,s_2),(s_3,s_4)\}$, we have $o(A) = 3$ and $d(A) = 12$. □

A **walk** on a directed graph is a sequence of arcs where the head of each arc except the last one is the tail of the next arc. A walk can be specified either by its vertices or its arcs in the order of appearance in it. A **path** on a directed graph is a walk with no repeating vertices. On an undirected graph, a path is a sequence of consecutive edges that connect no repeating vertices. A walk is called **Eulerian** if it contains all the *arcs* of the graph, while a path is called **Hamiltonian** if it contains all the *vertices* of the graph. A **cycle** is a path where the first and the last vertices are the same. A cycle with k arcs is called a k-cycle. For a string set S and the associated overlap and distance graphs, consider a cycle c on them, a string $s \in S$ corresponds to a vertex of c, and let s' be the unique previous string of s in c. The superstring $\langle s, \ldots, s' \rangle$ where the strings are in the order around c is called the **cycle superstring of** c **with respect to** s and is denoted by $strC(c,s)$. The superstring $\langle s, \ldots, s', s \rangle$ where the strings are in the order around c is called the **extended cycle superstring of** c **with respect to** s and is denoted by $strC^+(c,s)$. Notice that $strC^+(c,s) = m(strC(c,s),s)$.

Example 5. For the substring-free set $S = \{s_1, s_2, s_3, s_4\}$ of Example 2 and the associated overlap and distance graphs presented in Example 4, consider the cycle $c = (s_1, s_2, s_4, s_1)$. We have $strC(c,s_1) = bababbccaacabb$, and $strC^+(c,s_1) = bababbccaacabbababbc$. □

Some combinatorial optimization problems are closely related to the SSP due to their nature and are used in the establishment of many results of the SSP. A **matching** on a directed graph is a set of arcs, no two of which are incident to the same vertex. A **maximum matching** on a weighted graph is a matching with the largest total weight, while the **Matching Problem** (MP) looks for a maximum matching on a weighted directed graph. The MP is defined similarly on undirected weighted graphs. A **directed matching** is a set of arcs, no two of which have the same tail or the same head. In other words, it is a set of disjoint paths and cycles on a graph. The **Directed Matching Problem** (DMP) looks for a maximum directed matching. Both MP and DMP can be solved in polynomial time (see, e.g., [42, 59]). A **cycle cover** on a directed graph is a set of cycles such that each vertex of the graph is in exactly one cycle. The **Cycle Cover Problem** (CCP) on a weighted directed graph consists in finding a cycle cover with maximum total weight. The CCP is solvable in polynomial time by reduction to the MP on bipartite graphs (see, e.g., [42]).

The **Hamiltonian Path Problem** (HPP) on a weighted directed graph consists in finding an optimal Hamiltonian path according to its total weight. If the objective is to minimize the total weight, then the Min-HPP is considered, while if the objective is to maximize the total weight, then the Max-HPP is considered. The decision HPP on a directed graph G asks for the existence of a Hamiltonian path on G. Similarly, we have the maximization and the minimization **Hamiltonian Cycle Problem**, which are also known as **Traveling Salesman Problems** (Min-TSP and Max-TSP). Both HPP and TSP are NP-hard problems [30]. There is a simple relation between these problems. The HPP on a graph G can be transformed to the TSP on a graph G' obtained from G by adding a new vertex u and zero-weighted arcs from u to each vertex of G and from each vertex of G to u.

3 Computational Complexity

The results described in this section concern the computational complexity of the SSP, and justify the fact that there are only few exact algorithms, and on the other hand so many approximation algorithms for it. The SSP cannot be solved efficiently to optimality in polynomial time. It can be approximated within a constant ratio, whereas this ratio has a bound.

3.1 Complexity of Exact Solution

Given a string set S and a string s, there is a polynomial time algorithm for checking if s is a superstring of S, and therefore the decision SSP belongs in class NP.

A string is **primitive** if no letter appears more than once into it. Next theorem establishes the NP-completeness of the decision SSP.

Theorem 1 ([15]). *The decision SSP is NP-complete. Furthermore, this problem is NP-complete even if for any integer $m \geq 3$ the restriction is made that all strings in set S are primitive and of length m.*

The proof is based on a polynomial time transformation from the decision HPP on directed graphs with the following additional restrictions:

- there is a designated start vertex s with $\deg^-(s) = 0$ and a designated end vertex t with $\deg^+(t) = 0$,
- for each $v \neq t$, we have $\deg^+(v) > 1$.

A set S of specific strings of length 3 is constructed, and each string is corresponding to a vertex of a directed graph $G = (V, E)$ that satisfies the above restrictions. Graph G has a Hamiltonian path if and only if set S has a superstring of length $2|E| + 3|V|$. Therefore, there is no efficient algorithm for solving the SSP, unless P = NP.

Due to the nature of the SSP, several parameters can be considered fixed in order to define restricted cases of the problem. Besides the length of the strings and the primitiveness that were mentioned previously, the cardinality of the alphabet, the orbit size of the letters, and the form of the strings were also examined for the conservation or not of the NP-completeness.

The decision SSP remains NP-complete when it is restricted to an alphabet of cardinality 2 as proved in [15]. A restricted version of the SSP concerning both the alphabet cardinality and the string length is also studied and the result is stated in the next theorem. Let $bits(n)$ denote the number of bits that are necessary to represent n in binary, for any $n \in \mathbb{N}$.

Theorem 2 ([15]). *The decision SSP is NP-complete even if for any real $h > 1$, the strings in set S are written over the alphabet $\{0,1\}$ and have length $\lceil h\, bits(||S||) \rceil$.*

The proof is based on Theorem 1 and on the encoding of each letter of the initial alphabet with letters of the alphabet $\{0,1\}$ such that no relative changes yielded to the overlaps between the strings after the new encoding.

In [38, 39], NP-completeness results are proved for some special cases of the decision SSP. For a set S of strings over an alphabet Σ, these complexity results can be briefly presented as follows. The decision SSP is NP-complete even if

- all strings in S are of length 3 and the maximum orbit size of each letter in Σ is 8.
- all strings in S are of length 4 and the maximum orbit size of each letter in Σ is 6.
- $\Sigma = \{0,1\}$ and each string in S is of the form $0^p10^q10^r1$ or $10^p10^q10^r$, where $p,q,r \in \mathbb{N}$.
- $\Sigma = \{0,1\}$ and all strings in S are of the form 10^p10^q, where $p,q \in \mathbb{N}$.
- $\Sigma = \{0,1,2\}$ and each string contains a fixed number of each letter.

3.2 Complexity of Approximation

Since the SSP is a hard problem to be solved to optimality, a huge amount of effort is made to develop approximation algorithms. The theoretical framework for the complexity of this aspect establishes that although the SSP is easy to be approximated within *some* constant ratio, it is hard to be approximated within *any* constant ratio. The **linear reduction** (L-reduction) is necessary for what follows.

Definition 1 ([43]). Let A and B be two optimization problems. Problem A L-reduces to B if there are two polynomial time algorithms F and G and real constants $\alpha, \beta > 0$ such that

- given an instance a of A, algorithm F produces an instance b $=$F(a) of B such that opt(b) is at most $\alpha \times$ opt(a), where opt(a) and opt(b) are the costs of the optimal solution of instances a and b respectively, and
- given any solution of b with cost c', algorithm G produces in polynomial time a solution of a with cost c such that $|c - \text{opt(a)}| \leq \beta |c' - \text{opt(b)}|$.

For two optimization problems A and B and the constants α and β of the Definition 1, the following theorem establishes the basic usage of L-reduction.

Theorem 3 ([43]). *If problem A L-reduces to problem B and there is a polynomial time approximation algorithm for B with worst-case error ε, then there is a polynomial time approximation algorithm for A with worst-case error $\alpha\beta\varepsilon$.*

Therefore, if problem B has a polynomial time approximation scheme (PTAS), then so does problem A.

The class Max-SNP is a class of optimization problems defined syntactically in [43]. Every problem in Max-SNP can be approximated in polynomial time within some constant ratio. A problem is Max-SNP-hard if any other problem in Max-SNP L-reduces to it.

Theorem 4 ([8]). *The SSP is Max-SNP-hard.*

The proof is based on an L-reduction from the Min-HPP, where the degree of the associated directed graph is bounded, and all the weights are either 1 or 2, which is Max-SNP-hard [44]. The reduction from this problem to the SSP is similar to the one used to show the NP-completeness of the decision SSP in Theorem 1, with the extra establishment that it is an L-reduction. The strings that are considered for the above L-reduction have bounded lengths, and so the same reduction can be applied to the maximization version of the superstring problem with respect to the compression measure, and concludes to the same hardness result.

Corollary 1 ([8]). *Maximizing the total compression of a string set is Max-SNP-hard.*

In [5], it is proved that if a Max-SNP-hard problem has a PTAS, then P = NP. Therefore, there is no PTAS for the SSP, unless P = NP, which means that there exists an $\varepsilon > 0$ such that it is NP-hard to approximate the SSP within a ratio of $1+\varepsilon$.

The L-reduction described in [8] for the proof of the Max-SNP-hardness of the SSP produces instances with arbitrarily large alphabets. More precisely, each instance of the special Min-HPP with n vertices is transformed to an SSP instance over an alphabet with $2n+1$ letters. However, the SSP is APX-hard even if the alphabet contains just two letters as stated in the next theorem.

Theorem 5 ([41]). *The SSP is APX-hard both with respect to the length measure and the compression measure, even if the alphabet has cardinality* 2 *and every string is of the form* $10^m1^n01^m0^{n+4}10$ *or* $01^m0^n10^p1^q01^m0^n10^r1^s01$*, where* $m,n,p,q,r,s \geq 2$.

4 Polynomially Solvable Cases

Since the SSP is NP-hard, special cases of the problem that can be solved in polynomial time constitute an interesting aspect. Various additional restrictions on the problem's parameters, similar to these described in Sect. 3 lead to polynomial algorithms revealing the boundaries between hard and easily solvable cases of the problem.

Obviously, if the cardinality of the alphabet is equal to 1 or all the strings in the given set are of length 1, then the SSP is trivial. Also, if the number of the strings in the set is fixed, then the SSP is polynomially solvable by enumerating all the different string orders. However, there are more interesting and complicated polynomial cases of the SSP.

Since Theorem 1 establishes the NP-completeness of the SSP for string lengths greater than 2, the question is what happens in the remaining cases. The answer is given by the next theorem.

Theorem 6 ([37]). *For a string set $S = \{s_1, s_2, \ldots, s_n\}$ and an integer k, if $|s_i| \leq 2, i \in \mathrm{I}_n$, then there is a linear time and space algorithm to decide if S has a superstring of length k.*

A **path decomposition** of a directed graph G is a partition of $E(G)$ into edge-disjoint paths. Such a decomposition is minimum if it contains the minimum number of paths. The linear algorithm in Theorem 6 is based on a minimum path decomposition of a graph associated with the string set S. Besides the algorithm for the decision problem mentioned in Theorem 6, there is also a linear algorithm that finds a shortest superstring for strings of length at most 2.

A fixed maximum orbit size for the letters in the alphabet leads to a special case of the SSP that is also solvable in polynomial time. Assume a set S of strings over alphabet Σ and let $m = \max\{|s| : s \in S\}$.

Theorem 7 ([61]). *If the orbit size of each letter in Σ is at most 2 in S, then a shortest superstring for S is found in polynomial time $O(|\Sigma|^2 m)$.*

Another special case of the SSP concerns the fixed difference between the sum of string lengths and the cardinality of the alphabet as cited in [61]. Given a set S of strings over an alphabet Σ, for a fixed difference $||S|| - |\Sigma|$, the SSP is solvable in polynomial time by a special exhaustive enumeration. The difference $||S|| - |\Sigma|$ is mentioned as a measure of dissimilarity of the strings in S.

In [38], restricted cases of the SSP are studied, and a string form that induces polynomial cases is found.

Theorem 8 ([38]). *The SSP over the alphabet $\{0,1\}$ is polynomial time solvable if each given string contains at most one* 1.

As cited in [61], a particular case of the SSP in which S is the set of *all* three-letter strings over an alphabet Σ is known as the **Code Lock Problem**. In this case, the possible overlaps between the strings are 1 and 2. This problem is reducible to the **Eulerian Walk Problem**, where the existence of a walk that contains all the arcs of a directed graph is sought, and hence, according to [16] it is solvable in polynomial time.

5 Exact Solutions

There are only few exact algorithms in the literature for the SSP. This is due to the computational complexity of the problem, and the lack of necessity for optimal solutions at its main applications in computational molecular biology. In the DNA sequencing practice, the biological properties of a genome molecule can be usually expressed also by a superstring of its fragments that is not the shortest one, but its length is close to the optimum.

5.1 Exhaustive Enumeration

The SSP can be trivially solved by exhaustive enumeration of all possible arrangements of the strings. The merge of the strings in some of these orders would correspond to a shortest superstring. Given a set S of n string, the examination of the superstrings of S that correspond to all permutations in Π_n is enough to find a shortest one. The exhaustive examination of all permutations can be executed in time $O(n!||S||)$, or by a different implementation that also exhaustively enumerates the possible solutions, in time $O(n||S||^{n+1})$ as mentioned in [61]. Optimal solutions for small SSP instances taken by the exhaustive algorithm are used in [24, 25] to compare the compression achieved by the viruses to their genome with the largest possible compression of their genes.

5.2 Integer Programming Formulation

Given an SSP instance specified by a set S of n strings, consider the associated overlap graph $G_o(S)$. An optimal solution to the SSP instance can be obtained by an optimal solution to the Max-HPP on $G_o(S)$, since a maximum Hamiltonian path contains all the vertices (strings) ordered in a single path such that it has the maximum total overlap. Due to the relation between the HPP and the TSP described in Sect. 2, these solutions can be obtained by an optimal solution to the Max-TSP. According to these transformations, optimal solutions for the SSP can be derived by any integer programming formulation for the Max-TSP using branch and bound or cutting plane algorithms. In [17], a benchmark set of instances with known optimal solutions was constructed using the integer program of [40] for the Max-TSP and used to compare the solutions of a heuristic for the SSP with the optimal ones.

6 Approximation Algorithms

The fact that the SSP is Max-SNP motivates many researches to develop approximation algorithms for it. As mentioned in Sect. 2, there are two ways to assess the solution of an approximation algorithm: the length measure considering the SSP as a minimization problem, and the compression measure considering the SSP as a maximization problem.

For a string set S, and any algorithm ALG for the SSP, we use the notation $\text{ALG}_l(S)$ to denote the length of the superstring of S obtained by ALG, and $\text{ALG}_c(S)$ to denote the corresponding achieved compression. An approximation ratio $\varepsilon = \frac{\text{ALG}_l(S)}{\text{opt}_l(S)} \geq 1$ with respect to the length measure means that $\text{ALG}_l(S) \leq \varepsilon \times \text{opt}_l(S)$ for all instances, while an approximation ratio $\varepsilon = \frac{\text{ALG}_c(S)}{\text{opt}_c(S)} \leq 1$ with respect to the compression measure means that $\text{ALG}_c(S) \geq \varepsilon \times \text{opt}_c(S)$ for all instances. Although the two measures

are equivalent regarding the optimal solution, they differ regarding the approximate solutions of the problem. The existence of an algorithm with a constant approximation ratio for the one measure has in general no approximation performance guarantee for the other measure.

In this section, the approximation algorithms for the SSP both with respect to the length and the compression measure are presented, revealing the special features of the superstrings in each case.

6.1 Approximation of Compression

The compression measure counts the number of letters gained in comparison with the simply concatenation of all strings. Algorithms that approximate this gain are presented here.

6.1.1 The Natural Greedy Algorithm

A very well known, simply implemented, and widely used algorithm for the SSP is the natural greedy algorithm. It is routinely used in DNA sequencing practice. It starts with the string set S and repeatedly merges a pair of distinct strings with the maximum possible overlap until only one string remains in S. Next algorithm shows the pseudo-code of the natural greedy for the SSP.

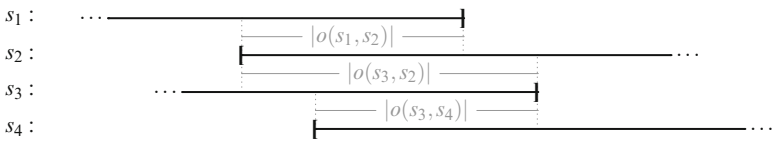
Algorithm: GREEDY
input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. **for** $i = 1$ **to** $n-1$ **do**
2. $\quad L = \{(s_i, s_j) : s_i, s_j \in S, i \neq j\}$
3. $\quad k = \max\{|o(s_i, s_j)| : (s_i, s_j) \in L\}$
4. $\quad$ let $(s'_i, s'_j) \in L$ be a pairs such that $|o(s'_i, s'_j)| = k$
5. $\quad S = (S - \{s'_i, s'_j\}) \cup \{m(s'_i, s'_j)\}$
6. **end**
7. let s be the only string in S
8. **return** s

The operation of the GREEDY algorithm on the string set S is equivalent to the creation of a Hamiltonian path on the overlap graph $G_o(S)$. In general directed weighted graphs, the total weight of the Hamiltonian path obtained by the greedy approach is at least one third the weight of a maximum path [26]. In the case of the overlap graphs, a stronger result can be obtained by exploiting their properties. A basic lemma that concerns the form of these graphs is restated here in terms of strings.

Lemma 1 ([58]). *Let s_1, s_2, s_3, and s_4 be strings, not necessarily distinct, such that $|o(s_3,s_2)| \geq |o(s_1,s_2)|$ and $|o(s_3,s_2)| \geq |o(s_3,s_4)|$. Then $|o(s_1,s_4)| \geq |o(s_1,s_2)| + |o(s_3,s_4)| - |o(s_3,s_2)|$.*

The proof can be derived directly from the next figure, where the alignment of the four strings according to their overlaps is presented.



Notice that, if s_1 and s_4 are not distinct, then the result of Lemma 1 concerns the self-overlap of the string s_1.

The following theorem establishes the approximation performance of the GREEDY algorithm based on the corresponding analysis of the greedy approach for the Max-HPP and on Lemma 1.

Theorem 9 ([58]). *For a string set, the compression achieved by the* GREEDY *algorithm is at least half the compression achieved by a shortest superstring.*

Next example, presented in [58], shows that the result of the Theorem 9 is the best possible.

Example 6. For the string set $\{ab^k, b^{k+1}, b^k a\}$, $k \geq 1$, over the alphabet $\{a,b\}$, GREEDY may produce the superstring $ab^k ab^{k+1}$ or the superstring $b^{k+1}ab^k a$ that achieves compression k, whereas the shortest superstring is $ab^{k+1}a$ and achieves compression $2k$. Notice that GREEDY can also give the shortest superstring depending on how it breaks ties. □

6.1.2 Approximation Based on Matchings

Apart from GREEDY, two other $\frac{1}{2}$-approximation algorithms for the compression of a superstring based on the MP and the DMP are presented in [62]. For an SSP instance S, consider the associated overlap graph $G_o(S)$. In both algorithms, a matching algorithm is repeatedly applied to $G_o(S)$, to produce a Hamiltonian path.

For the description of the first algorithm the notion of the **arc contraction** is necessary. Given a weighted directed graph G and an arc $e = (u,v) \in E(G)$, the contraction of e is denoted by G/e and gives a new graph obtained from G where the vertices u, v and their incident arcs are replaced by a new vertex w which has as incoming arcs the incoming arcs of u and as outgoing arcs the outgoing arcs of v with the same weights as on G. The MATCH algorithm initially finds a maximum matching on $G_o(S)$, and then contracts the arcs of the matching. This process is repeated on the new graph until a graph with no arcs comes up. $G_o(S) = (V,E,w)$ is the initial overlap graph which remains unchanged, whereas G denotes the graph obtained in each iteration after the arc contractions. Initially, $G = G_o(S)$. Let maxm(G) be a maximum matching on graph G.

Algorithm: MATCH

input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. construct the graph $G_o(S)$
2. $P = \emptyset$
3. $G = G_o(S)$
4. **while** $|E(G)| \neq \emptyset$ **do**
5. $\quad M = maxm(G)$
6. $\quad P = P \cup \{$the arcs of $E(G_o(S))$ that correspond to $M\}$
7. $\quad$ **foreach** $(u,v) \in M$ **do**
8. $\quad\quad G = G/(u,v)$
9. $\quad$ **end**
10. **end**
11. let s be the superstring that corresponds to P
12. **return** s

The approximation performance of the MATCH algorithm is based on the observation that any matching on an overlap graph can be extended to a Hamiltonian path on it, since overlap graphs are complete. Moreover, a maximum matching has total weight at least half the weight of a maximum Hamiltonian path. This can easily be shown by considering each Hamiltonian path as two matchings with distinct arcs, constructed by taking alternate arcs from the path. These results imply that the compression achieved by the MATCH algorithm is at least half the optimal compression.

The second algorithm with the same approximation ratio for the SSP is based on the slightly different DMP. Remember that a directed matching on a graph is a set of disjoint paths and cycles. For the description of this algorithm, the notion of the arc contraction is extended naturally to paths. Given a weighted directed graph G and a path $p = (v_1, v_2, \ldots, v_r)$ on it, defined by its vertices, the contraction of p gives a new graph obtained from G where the vertices $v_1, \ldots, v_r$ and their incident arcs are replaced by a new vertex w which has as incoming arcs, the incoming arcs of v_1 and as outgoing arcs the outgoing arcs, of v_r with the same weights as on G. The DIMATCH algorithm described also in [62] operates exactly as MATCH except that it finds a *directed* matching of each step, opens each cycle of it by deleting an arc with the smallest weight, and finally contracts the paths into vertices. The compression achieved by the DIMATCH algorithm is at least half the optimal compression.

6.1.3 Approximation Based on the TSP

Any approximation algorithm for the Max-TSP is also an approximation algorithm for the SSP with respect to the compression measure, or equivalently for the Max-HPP, with the same ratio due to the transformation from the TSP to the HPP. For both problems, it is implied that they are **asymmetric**, which means that they applied on directed graphs, and that the weight of an arc (u,v) is not necessarily equal to the weight of the arc (v,u).

In [7, 31], two approximation algorithms for the Max-TSP are presented. In both cases, procedures with complementary worst cases run on directed graphs with even number of vertices. The best result among them is a Hamiltonian cycle whose weight is at least $\frac{38}{63}$ times the weight of a maximum weight Hamiltonian cycle for the first algorithm, and $\frac{8}{13}$ for the second. Both algorithms achieve their approximation performance without utilizing any special structure of the strings. In both algorithms is required that the complete input graph G has an even number of vertices. In general, for an SSP instance of n strings, the above algorithms achieve approximation ratios $\frac{38}{63}(1-\frac{1}{n})$ and $\frac{8}{13}(1-\frac{1}{n})$, respectively. Finally, an approximation algorithm for the Max-TSP is also designed in [29], achieving the best ratio until now, namely $\frac{2}{3}$. It operates by decomposing a special form of directed multigraphs, where the elements of the arc set are ordered triples of the vertex set.

6.2 Approximation of Length

A plethora of approximation algorithms with respect to the length measure have been developed for the SSP using different variations of the greedy strategy. The best one among them finds a string whose length is at most $2\frac{1}{2}$ times the length of the optimal string.

6.2.1 Naive Approximation Algorithm

A naive algorithm for the SSP is used in [8] for comparison reasons in relative performance of other approximation algorithms. Its approximation performance is not remarkable but the idea is quite simple, showing that it is easy to develop an algorithm for the SSP, but it is not so easy to achieve a good approximation ratio. For a string set S, the algorithm arbitrarily chooses a string from S considering it as the initial current string, and then repeatedly updates the current string by merging it with a remaining string from S that yields the maximum overlap. The performance of this algorithm highly depends on the random choice of the initial point, and it is possible to produce superstrings whose length grows quadratically in the optimal length.

6.2.2 Approximation Algorithm Used in a Learning Process

The first attempt to approximate the shortest superstring of a set was made in [34], where the DNA sequencing procedure is modelled as a string learning process from randomly drawn substrings of it. Under certain restrictions, this may be viewed as a string learning process in Valiant's distribution free learning model [63]. The efficiency of the learning method depends on the solution of an algorithm which approximates the length of a superstring, and seeks in each step for an appropriate join string among the candidate ones.

Given a string set $S = \{s_1, s_2, \ldots, s_n\}$ and a string s, we denote by $subSs(S, s)$ the set of the strings in S that are substrings of s.

Example 7. Suppose that we have the string set $S = \{s_1, s_2, s_3\}$, where $s_1 = caabaa$, $s_2 = abaaca$, and $s_3 = baacaa$ are strings over the alphabet $\{a, b, c\}$. The set of all join strings of s_1 and s_2 regardless their order is $J_{\{s_1,s_2\}} = \{caabaaabaaca, caabaabaaca, caabaaca, abaacacaabaa, abaacaabaa\}$, while the only join string $s \in J_{\{s_1,s_2\}}$ for which $subSs(S, s) = S$ is the string $abaacaabaa$. □

Given a string set S, the GROUP-COMBINE algorithm constructs a superstring of S by an iterative process. The algorithm begins with a string set and combines the strings in groups such that all strings in a group are substrings of a join string of two of them, trying to find as large groups as possible.

Algorithm: GROUP-COMBINE
input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. $T = \emptyset$
2. **while** $|S| > 0$ **do**
3. find $s_i, s_j \in S$ such that $\min_{s \in J_{\{s_i,s_j\}}} \frac{|s|}{||subSs(S,s)||}$ is minimized
4. let $\bar{s}$ be the join string that achieves the minimum in step 3
5. $S = S - subSs(S, \bar{s})$
6. $T = T \cup \{\bar{s}\}$
7. **if** $|S| = 0$ ***and*** $|T| > 1$ **then**
8. $S = T$
9. **end**
10. **end**
11. let s be the only string in T
12. **return** s

Next theorem establishes the approximation ratio of the algorithm.

Theorem 10 ([34]). *Given a string set, if the length of the optimal superstring is m, then* GROUP-COMBINE *produces a superstring of length* $O(m \log m)$.

6.2.3 4-Approximation Algorithms

The first approximation algorithm with a constant ratio for the length of a superstring is described in [8], answering a notorious open problem for the existence of such an algorithm. The algorithm utilizes a minimum cycle cover on the distance graph of a string set to derive a superstring with preferable properties that bound its length. Given an SSP instance S, the CYCLE-CONCATENATION algorithm finds a minimum cycle cover on the graph $G_d(S)$ *with loops* in polynomial time. Then it

opens each cycle of the cover by removing an arc chosen randomly, constructs the superstring that corresponds to the obtained path, and concatenates these strings.

Algorithm: CYCLE-CONCATENATION

input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. construct the graph $G_d(S)$ with loops
2. find a minimum cycle cover $C = \{c_1, c_2, \ldots, c_p\}$ on $G_d(S)$
3. **foreach** $c_i \in C$ **do**
4. choose a vertex $s_i \in c_i$ randomly
5. $s'_i = strC(c_i, s_i)$
6. **end**
7. let s be the concatenation of the strings s'_i
8. **return** s

Next theorem demonstrates the approximation performance of the algorithm establishing the first constant approximation ratio for the SSP.

Theorem 11 ([8]). *For a string set S,* CYCLE-CONCATENATION *produces a superstring of length at most* $4 \times opt_l(S)$.

Another algorithm for the SSP with the same constant approximation ratio is MGREEDY which is presented in [8].

Algorithm: MGREEDY

input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. $T = \emptyset$
2. **while** $|S| > 0$ **do**
3. $k = \max\{|o(s'_i, s'_j)| : s'_i, s'_j \in S\}$
4. let (s_i, s_j) be a string pair such that $|o(s_i, s_j)| = k$
5. **if** $i \neq j$ **then**
6. $S = (S - \{s_i, s_j\}) \cup \{m(s_i, s_j)\}$
7. **else**
8. $S = S - \{s_i\}$
9. $T = T \cup \{s_i\}$
10. **end**
11. **end**
12. let s be the concatenation of the strings in T
13. **return** s

Notice that at line 3, the two strings of each pair are not necessarily distinct allowing in this way the self-overlaps. Since the choices at line 4 are made according to the overlaps in S, MGREEDY can be thought as choosing arcs from the graph $G_o(S)$ *with loops*. The choice of the pair (s_i, s_j) corresponds to the choice of the arc $(\text{last}(s_i), \text{first}(s_j))$ on $G_o(S)$ in each step. Therefore, the algorithm constructs paths, and closes them into cycles when distinctness is not satisfied at line 4. Thus,

MGREEDY ends up with a set of disjoint cycles that cover the vertices of $G_o(S)$, which is a cycle cover. The same cycle cover can be thought on graph $G_d(S)$ *with loops*. For a cycle cover C, by Eq. (2), we have $o(C) + d(C) = ||S||$, and so a cycle cover has minimum total weight on $G_d(S)$ if and only if it has maximum total weight on $G_o(S)$, and in both cases it is called optimal.

Theorem 12 ([8]). *The cover created by* MGREEDY *is an optimal cycle cover.*

Notice that the presence of the loops is a necessary assumption for this result. Since MGREEDY finds an optimal cycle cover, the superstring that is produced by it is no longer than the string produced by algorithm CYCLE-CONCATENATION. Therefore, the approximation ratio with respect to the length measure of MGREEDY for the SSP is also equal to 4. Actually, the superstring of MGREEDY could be shorter than the one obtained by CYCLE-CONCATENATION since MGREEDY simulates the breaking of each cycle in the optimal position, that is between the strings with the minimum overlap in the cycle.

6.2.4 GREEDY Is a $3\frac{1}{2}$-Approximation Algorithm

The GREEDY algorithm has already been presented as an approximation for the compression. A notorious open question is how well GREEDY approximates the length of a shortest superstring, while a common conjecture states that GREEDY produces a superstring of length at most two times the length of the optimum [54, 58, 62]. In fact, GREEDY may give a superstring almost twice as long as the optimal one, as shown in the next example from [8].

Example 8. For the string set $\{c(ab)^k, (ba)^k, (ab)^k c\}$, $k \geq 1$, over the alphabet $\{a, b, c\}$, GREEDY may produce the superstring $c(ab)^k c(ba)^k$ or the superstring $(ba)^k c(ab)^k c$ of length $4k+2$, whereas the shortest superstring is $c(ab)^{k+1}c$ of length $2k+4$. □

In [8], it is proved that GREEDY is a 4-approximation algorithm for the SSP. Next theorem improves this approximation ratio based on a more careful analysis on specially formed strings.

Theorem 13 ([28]). *The* GREEDY *algorithm is a* $3\frac{1}{2}$*-approximation algorithm with respect to the length measure.*

6.2.5 A 3-Approximation Algorithm

The algorithm TGREEDY described in [8] operates in the same way as MGREEDY except that in the last step it merges the strings in set T by running GREEDY on them instead of simply concatenates them. Next theorem establishes its approximation performance.

Theorem 14 ([8]). *For a string set S, algorithm* TGREEDY *produces a superstring of length at most* $3 \times opt_I(S)$.

In [8], a relative performance comparison between GREEDY, MGREEDY, and TGREEDY algorithms is presented. TGREEDY always produces better solutions than MGREEDY since in the last step it greedily merges the strings, whereas MGREEDY just concatenates them. The approximation performance of TGREEDY is better than this of GREEDY, but the superiority of one of these algorithms over the other is not guaranteed as shown in the next example.

Example 9. For the string set $\{c(ab)^k, (ab)^{k+1}a, (ba)^k c\}$, $k \geq 1$, over the alphabet $\{a,b,c\}$, GREEDY produces the shortest superstring $c(ab)^{k+1}ac$ of length $2k+5$, whereas TGREEDY produces the superstring $c(ab)^k ac(ab)^{k+1}a$ or the superstring $(ab)^{k+1}ac(ab)^k ac$ of length $4k+6$, since the initial maximum overlap is the self-overlap of the second string.

On the other hand, for the string set $\{cab^k, ab^k ab^k a, b^k dab^{k-1}\}$, $k \geq 1$, over the alphabet $\{a,b,c,d\}$, TGREEDY produces the shortest superstring $cab^k dab^k ab^k a$ of length $3k+6$, since the initial maximum overlap is the self-overlap of the second string, whereas GREEDY produces the superstring $cab^k ab^k ab^k dab^{k-1}$ or the superstring $b^k dab^{k-1} cab^k ab^k a$ of length $4k+5$. □

6.2.6 Generic Approximation Based on Cycle Covers

Algorithms MGREEDY and TGREEDY implicitly construct optimal cycle covers on the associated overlap and distance graphs of a string set, while CYCLE-CONCATENATION explicitly takes advantage of this construction. A generic algorithm that explains this basic idea is presented in [11].

For a string set S, let $C = \{c_1, c_2, \ldots, c_p\}$ be a cycle cover on the graph $G_d(S)$. Suppose that an arbitrary string r_i is picked from each cycle $c_i \in C$, and these strings form the representative set $R = \{r_1, r_2, \ldots, r_p\}$. Let $r = \langle r_1, r_2, \ldots, r_p \rangle$ be a superstring of R. By replacing each r_i, $i \in \mathrm{I}_p$, in r with the string $strC^+(c_i, r_i)$, we get the string

$$\langle strC^+(c_1,r_1), strC^+(c_2,r_2), \ldots, strC^+(c_p,r_p) \rangle,$$

which is called the **extension string of** r **with respect to** C and is denoted by $\mathrm{ext}(r,C)$. Observe that $\mathrm{ext}(r,C)$ is a superstring of S.

For a string set S, the GENERIC-COVER algorithm constructs a minimum cycle cover C on the graph $G_d(S)$, and chooses a random string from each cycle of this cover to form a set R of representatives. Then, it finds a new minimum cycle cover on $G_d(R)$, opens each cycle of this cover in a random position, and concatenates the resulting cycle superstrings to create a superstring of R. Finally, it returns the extension string of this superstring with respect to the cycle cover C to take a superstring of S.

Algorithm: GENERIC-COVER

input : string set $S = \{s_1, s_2, \ldots, s_n\}$
output: a superstring of S

1. construct the graph $G_d(S)$
2. find a minimum cycle cover C on $G_d(S)$
3. $R = \emptyset$
4. **foreach** $c_i \in C$ **do**
5. choose a string s_i of c_i randomly
6. $R = R \cup \{s_i\}$
7. **end**
8. create the graph $G_d(R)$
9. find a minimum cycle cover C_R on $G_d(R)$
10. **foreach** *cycle* $c_i \in C_R$ **do**
11. let s_i be the head of a randomly chosen arc of c_i
12. $s_i' = strC(c_i, s_i)$
13. **end**
14. let r be the concatenation of the strings s_i'
15. $\overline{r} = ext(r, C)$
16. **return** $\overline{r}$

The GENERIC-COVER algorithm has approximation ratio equal to 3. This algorithm constitutes the base for the design of better approximation algorithms for the SSP as described below.

6.2.7 Handling 2-Cycles and 3-Cycles Separately

For an SSP instance specified by a string set S, $\text{opt}_c(S)$ may grow quadratically in $\text{opt}_l(S)$ in general. Thus, to take advantage of a *compression* approximation to design *length* approximation algorithms with constant ratio based on GENERIC-COVER framework, a key is to construct suitable subproblems for which $\text{opt}_c(S)$ is linear in $\text{opt}_l(S)$. The main difficulty in determining such subproblems and so in improving the length approximation performance of the GENERIC-COVER algorithm appears in handling k-cycles with small k in the cycle cover C_R. In [60], the compression achieved by GREEDY is utilized, to design a length approximation algorithm for the SSP. The algorithm is based on the scheme of GENERIC-COVER handling separately the 2-cycles in the minimum cycle cover C_R. In this way, the algorithm achieves an approximation ratio $2\frac{8}{9}$. In [11], an approximation algorithm that handles separately the 2-cycles and the 3-cycles is developed and gives a superstring of length at most $2\frac{5}{6}$ times the length of a shortest superstring.

6.2.8 Approximation Algorithms Based on the TSP

As cited in [31], a relationship between the SSP and the Max-TSP according to their approximation is given by the following lemma.

Lemma 2 ([8]). *If the Max-TSP has a* $(\frac{1}{2}+\varepsilon)$*-approximation, then the SSP has a* $(3-2\varepsilon)$*-approximation with respect to the length measure.*

Utilizing this relation, the approximation algorithms for the Max-TSP mentioned in Sect. 6.1.3 can be used to derive approximation ratios for the SSP with respect to the length measure. Consider an SSP instance S of n strings and the corresponding Max-TSP instance $G_o(S)$. For even number of strings, the algorithm described in [31] and achieves an approximation ratio of $\frac{38}{63}$ for the Max-TSP, gives a $2\frac{50}{63}$-approximation ratio for the SSP, while the algorithm described in [7] and achieves an approximation ratio of $\frac{8}{13}$ for the Max-TSP, gives a $2\frac{10}{13}$-approximation ratio for the SSP. For odd number of strings the algorithms achieves a ratio of $2(\frac{50}{63}+\frac{1}{n})$ and $2(\frac{10}{13}+\frac{1}{n})$ for the SSP, respectively. The algorithm described in [29] and achieves an approximation ratio of $\frac{2}{3}$ for the Max-TSP, gives a $2\frac{2}{3}$-approximation ratio for the SSP.

6.2.9 Exploiting the Superstring Structures

The approximation algorithms presented above are largely graph-theoretical, meaning that they sufficiently exploit the structure of overlap and distance graphs, but they do not take advantage of the structure inside the strings or in general of the properties not evident in graph representation. In this sense, they solve a more general problem than the one at hand.

An algorithm that captures a great deal of the structure of the SSP instances is presented in [3]. It takes advantage of the structure of strings with large value of overlap, proving several key properties of such strings. It follows the framework of the GENERIC-COVER algorithm using a more sophisticated way to choose the representatives at line 5 and to open each cycle at line 12. After finding a cycle cover on the associated distance graph, the key is to exploit the periodic structure of the cycle superstrings that arise. In this way, the algorithm achieves a bound either to the total overlap of the rejected arcs at line 12 or to the total additional length of extending each cycle at line 15. The result is to construct a superstring whose length is no more than $2\frac{3}{4}$ times the length of an optimal superstring.

This algorithm and the $2\frac{50}{63}$-approximation algorithm for the Max-TSP that is mentioned in Sect. 6.2.8 have complementary worst cases, and so a better ratio can be achieved by their combination. When the worst case of the first algorithm occurs, the Max-TSP algorithm runs as a subroutine on the set of representatives to take a better result. Balancing the two algorithms, an approximation ratio of $2\frac{50}{69}$ for the SSP can be achieved [2].

In [4], the study of the key properties is extended to strings that exhibit a more relaxed form of the periodic structure considered before. Algorithmically,

the new approach is also based on the framework of the GENERIC-COVER and is a generalization of the previous one. On the other hand, the analysis is very different and includes a special structure of 2-cycles. Let c be a 2-cycle in the cycle cover C_R of the GENERIC-COVER algorithm, consisting of the vertices s_i and s_j, which are the representatives of the cycles c_i and c_j in the cycle cover C. Without loss of generality assume that $d(c_i) \geq d(c_j)$. The cycle c is a **g-HO2-cycle** if

$$\min\{|o(s_i,s_j)|,|o(s_j,s_i)|\} \geq g(d(c_i)+d(c_j)).$$

In the new algorithm, during the selection of the representatives a technique is used to anticipate the potential of each string to participate in a $\frac{2}{3}$-HO2-cycle. Such strings have a very specific structure, and if there is a string without such a structure in a cycle, it is chosen as the representative. Otherwise, the knowledge of the structure of the entire cycle can be used to trade the amount of the lost overlap against the additional length of extending the representative to include the rest of the cycle. In this way, a $2\frac{2}{3}$-approximation algorithm for the SSP is designed.

6.2.10 Rotations of Periodic Strings

Two approximation algorithms for the SSP that are based also on the inner structure of the strings and their periodic properties are presented in [9]. They use the same framework of the GENERIC-COVER algorithm, but they make use of new bounds on the overlap between two strings.

Both algorithms pay special attention to the selection of the representatives but without concentrating on k-cycles with small k. Instead of choosing a string obtained by opening each cycle, the new idea is to look for superstrings of the strings in a cycle that are *not* too long and are guaranteed *not* to overlap with each other by too much. Each chosen superstring does not even have to be one of the cycle superstrings obtained by opening the cycle. Given a cycle $c_i = (s_1, s_2, \ldots, s_p, s_1)$ of the cycle cover C of algorithm GENERIC-COVER, a string r_c is a candidate to be a representative of c if for some j

- r_c is a superstring of $strC(s_{j+1})$ and
- r_c is a substring of $strC^+(s_j)$.

A sophisticated procedure is used to choose the representatives such that they satisfy these two conditions and also have an appropriate property to lead to the improved ratio.

After this step, the two approximation algorithms follow different ways. The first algorithm after finding the second cycle cover opens each cycle and concatenates the cycle superstrings, achieving an approximation ratio of $2\frac{2}{3}$. The second algorithm constructs a superstring of the representatives using as subroutine an approximation algorithm with respect to the compression measure for the SSP. As subroutines, we can use the approximation algorithms cited in Sect. 6.1.3. Using

the $\frac{38}{63}$-approximation algorithm described in [31] a length ratio of $2\frac{25}{42}$ is achieved, while using the $\frac{8}{13}$-approximation algorithm described in [7] a length ratio of $2\frac{15}{26}$ is achieved.

6.2.11 $2\frac{1}{2}$-Approximation Algorithms

The best approximation ratio with respect to the length measure for the SSP is the $2\frac{1}{2}$ until now. It can be achieved by two different methods, one from the field of the superstrings and the other from the field of the TSP.

The first algorithm is described in [57]. Given a string set S, the algorithm begins by constructing a minimum cycle cover C on graph $G_d(S)$. Then, instead of choosing representatives, it combines the cycles of C to produce a new cycle cover C', and finally opens each cycle in C' to produce a set of cycle superstrings. The concatenation of these superstrings yields a superstring of S. The algorithm exploits the properties of cycles and cycle covers on a special multigraph to achieve the $2\frac{1}{2}$-approximation ratio.

The second approach that achieves the same length ratio for the SSP is an approximation algorithm for the Max-TSP described in [29]. It finds a Hamiltonian cycle whose weight is at least $\frac{2}{3}$ the weight of a maximum Hamiltonian cycle. Using this procedure as a subroutine in the algorithm cited in Sect. 6.2.10, a length ratio of $2\frac{1}{2}$ for the SSP can be achieved.

7 Parallelizing the Solving Process

In complexity theory, the class NC consists of the decision problems (languages) decidable in polylogarithmic parallel time $O(\log^{O(1)} n)$ on a parallel computer with polynomial number $O(n^{O(1)})$ processors. In this definition, a parallel random access machine (PRAM) is assumed, that is a parallel computer with a central pool of memory, where any processor can access any bit of memory in constant time. The class RNC, which stands for random NC, extends NC with access to randomness. The class RNC consists of the decision problems (languages) that have a randomized algorithm which is solvable in polylogarithmic parallel time on polynomially many processors, and its probability of producing a correct solution is at least $\frac{1}{2}$.

It is conjectured that there are some tractable problems which are inherently sequential and cannot significantly be sped up by using parallelism. For an algorithm, a common method to show that it is hardly parallelizable is to prove that the algorithm is P-complete for the problem it applied to. The Greedy algorithm belongs to this case since the problem of finding a superstring chosen by the Greedy algorithm is P-complete [11]. This means that Greedy is difficult to be parallelized effectively. In the following, parallel approximation algorithms for the SSP are presented.

7.1 NC Algorithm with Logarithmic Length Ratio

Given a ground set X of elements and a family Y of subsets of X, a **set cover of** X **with respect to** Y is a subfamily $Y' \subseteq Y$ of sets whose union equals to X. Assigning a weight $w(x)$ to each element $x \in X$ the total weight of each set and family is naturally defined. The **Set Cover Problem** (SCP) is to find a set cover of the ground set of the minimum weight. The SCP can be approximated within a logarithmic ratio by a parallelizable algorithm [6].

In [11], a similar approach to the one presented in Sect. 6.2.2 for grouping is applied to the SCP. Given a set S of n strings, we define

$$F = \{subSs(S,s) : s \in J_{\{s_i,s_j\}}, s_i, s_j \in S\},$$

that is the family of the sets of substrings of all possible pairwise join strings from S. Considering S as the ground set and F as a family of its subsets, they specify an instance of the SCP. From each set cover $C \subseteq F$ of S, a string s_C can be constructed by merging the join strings that correspond to the sets of C. Observe that s_C is a superstring of S. Let the weight of each set of F be the length of the corresponding join string, and $w(C)$ be the total weight of the set cover C. Because of the merging of the join strings, $|s_C| \leq w(C)$. Also, it is proved that the length of the superstring corresponds to a minimum set cover C^* is at most twice the length of an optimal superstring, that is $|s_{C^*}| \leq 2 \times \mathrm{opt}_l(S)$. These results combined with the parallelization of the SCP imply an NC algorithm with logarithmic approximation for the SSP.

Theorem 15 ([11]). *For a string set S of n strings, there is an NC algorithm that for any $\varepsilon > 0$, finds a superstring whose length is at most $(2+\varepsilon)\log n$ times the length of a shortest superstring.*

Observe that each group of strings selected by the GROUP-COMBINE algorithm is a set of the family F as it was described previously, and so this algorithm constructs implicitly a set cover of S with respect to F. Theorem 15 proves that this result can also be obtained by a parallelizable procedure letting as open problem the design of an NC algorithm with a constant approximation ratio with respect to the length measure for the SSP.

7.2 RNC Algorithm with Constant Length Ratio

An RNC algorithm for the SSP is based on a parallelizable implementation of the sequential $2\frac{5}{6}$-approximation algorithm mentioned in Sect. 6.2.7. The only non-trivially parallelizable steps of this algorithm are the computations of the minimum cycle covers. Remember that, the problem of finding an optimal cycle cover is equivalent to the problem of finding a maximum matching on a bipartite graph. In general, it is not known if it can be done in either NC or in RNC. However, when the weights of the graph are given in unary notation, a condition that can be satisfied in the case

of this algorithm, a maximum matching can be found in RNC (see e.g. [49]), giving the next theorem for the SSP.

Theorem 16 ([11]). *For a string set S, there is an RNC algorithm that finds a superstring of length at most* $2\frac{5}{6} \times opt_I(S)$.

7.3 NC Algorithm with Compression Ratio $\frac{1}{4+\varepsilon}$

Given a weighted directed graph, a natural greedy approach for finding a maximum cycle cover is described as follows. Scan the arcs in non-increasing order of weights, and select an arc that does not have the same head or the same tail with a previously selected arc. Repeat until the selected arcs form a cycle cover. This approach finds a cycle cover of weight at least half the weight of a maximum cycle cover [11]. As mentioned in Sect. 6.2.3, if the graph is an overlap graph *with* loops then this greedy approach always finds a maximum weight cycle cover.

For the development of an NC compression approximation algorithm with a constant ratio for the SSP, a slightly different algorithm from the natural greedy for the CCP is designed. This algorithm achieves a worse approximation ratio, but can be parallelized. It is based on the idea that the natural greedy algorithm could be only a bit worse if in each step it chooses instead of the maximum weight arc, one with a similar weight. The arcs of the graph are partitioned into levels, such that the weights of all arcs in a level are within a constant factor. Given a graph G and a real $c > 1$, an arc $e \in E(G)$ has c-level equal to k if $c^{k-1} < w(e) \leq c^k$, and c-level equal to 0 if $w(e) \leq 1$. The algorithm operates like the natural greedy algorithm assuming that all arcs in each level have the same weight. The usage of this algorithm on overlap graphs for finding superstrings concludes to the next theorem.

Theorem 17 ([11]). *For a set S of n strings, there is an NC algorithm for the SSP that achieves a compression ratio* $\frac{1}{4+\varepsilon}$. *It runs either in time* $O(\log^2 n \log_{1+\varepsilon} ||S||)$ *on a PRAM with* $||S|| + n^4$ *processors or in time* $O(\log^3 n \log_{1+\varepsilon} ||S||)$ *on a PRAM using* $n^2 + ||S||$ *processors.*

8 Inapproximability Bounds

Both minimization and maximization versions of the superstring problem are Max-SNP-hard, which means that there exists an $\varepsilon > 0$ such that it is NP-hard to approximate the SSP within a ratio of $1 + \varepsilon$ with respect to the length measure, or within a ratio of $1 - \varepsilon$ with respect to the compression measure. The practical side of this theoretical result is expressed by explicit bounds to the approximation ratio in both cases.

The first work to this direction appears in [41], where inapproximability bounds are given for a special case of the SSP. Specifically, the result concerns SSP instances

where the alphabet is $\{0,1\}$ and every string is of the form $10^m1^n01^m0^{n+4}10$ or $01^m0^n10^p1^q01^m0^n10^r1^s01$, where $m,n,p,q,r,s \geq 2$. This special case is used also in Theorem 5 that concerns APX-hardness results. Let us refer to this special case as SSP_2 for short. The next two theorems establish the inapproximability results.

Theorem 18 ([41]). *The SSP_2 is not approximable within $1\frac{1}{17245}$ with respect to the length measure, unless P = NP.*

Theorem 19 ([41]). *For every $\varepsilon > 0$, the SSP_2 is not approximable within $1\frac{1}{11216} - \varepsilon$ with respect to the compression measure, unless P = NP.*

In [64], inapproximability bounds for the SSP restricted to instances with equal length strings are given. Moreover, these bounds are extended to instances over alphabets of cardinality 2 improving the previous ones.

Theorem 20 ([64]). *For any $\varepsilon > 0$, unless P = NP, the SSP on instances with equal length strings is not approximable in polynomial time within ratio*

- $1\frac{1}{1216} - \varepsilon$ *with respect to the length measure, and*
- $\frac{1070}{1071} + \varepsilon$ *with respect to the compression measure.*

A very important result about the relation between the inapproximability of the SSP over an alphabet of cardinality 2 and over any alphabet is established in the next theorem. It implies that the alphabet cardinality does not affect the approximability of the SSP.

Theorem 21 ([64]). *Suppose that the SSP can be approximated by a ratio ε on instances over an alphabet of cardinality 2. Then the SSP can be approximated by a ratio ε on instances over any alphabet.*

This result holds for both measures, length and compression. Therefore, the bounds established in Theorem 20 hold also for alphabets of cardinality 2.

The computation of the inapproximability bounds for the SSP reveals the large gap between these and the best known approximation ratios for the problem both for the length measure and the compression measure.

9 Heuristics

The design of the approximation algorithms is oriented to the achievement of the approximation ratio and not to the best possible result. On the other hand, real-world applications usually need practically good results and not theoretically good ratios for the result. A heuristic algorithm can satisfy this requirement by giving solutions to SSP instances that have not approximation performance guarantee but are experimentally close to the optimum. The greedy strategies seem to perform much better than their proved approximation ratios both in average and in real-world cases. In this section, the heuristic algorithms for the SSP are described.

9.1 A Variant of the Natural Greedy

A problem with the GREEDY algorithm is that it makes choices that may forbid good overlaps from future selection. In an attempt to eliminate this behaviour, a heuristic that imitates GREEDY but chooses differently the string pair in each step is described in [58]. Here, the modification is given in terms of strings instead of arcs in the associated overlap graph as made in the original work. The selection criterion in each step is not just the overlap but the overall influence of the choice of each string pair. Given a string set S and two string s_i and s_j in it, let

$$\begin{aligned}\mathrm{oi}(s_i, s_j) = a|o(s_i, s_j)| \\ - \max\{|o(s_{i'}, s_j)|, s_{i'} \in S, i' \neq i\} \\ - \max\{|o(s_i, s_{j'})|, s_{j'} \in S, j' \neq j\}.\end{aligned}$$

where a is a parameter that tunes the method. The idea is to take under consideration also the overlaps that would be eliminated if the pair (s_i, s_j) is selected. The pseudo-code of this heuristic algorithm is exactly as the one of GREEDY except that line 3 changes to $k = \max\{\mathrm{oi}(s'_i, s'_j) : (s'_i, s'_j) \in L\}$. In experiments cited in [58] with this heuristic algorithm, the best results were obtained with parameter a values from 2 to 2.5. In this case, the modified algorithm gives superstrings with average additional length from the optimum about $\frac{1}{5}$ the corresponding average additional length of GREEDY.

9.2 A Heuristic Parametrized by a Learning Process

A three-stage heuristic algorithm for the SSP, named ASSEMBLY, is presented in [20]. It is based on the observation that the set of the remaining strings in the GREEDY algorithm after a number of merges is very possible to contain only string pairs with small overlaps. The ASSEMBLY algorithm, in a try to avoid mistakes, terminates the greedy strategy when false merges are expected to occur, a decision based on the number of remaining strings.

The first stage of the algorithm is similar to the GREEDY algorithm except that it is terminated when the remaining string set has a cardinality c. The second stage of the ASSEMBLY algorithm is also based on greedy choices, although not made among all the possible overlaps, but only among these that pass a certification procedure. Given two strings s_i and s_j in the set of the remaining strings with $|o(s_i, s_j)| > 0$, a third string s_k is a **certificate** if its overlap with both s_i and s_j is greater than 0. It is experimentally determined that for two strings s_i and s_j with $|o(s_i, s_j)| > 0$, the existence of a certificate increases the probability their merge string participates to the shortest superstring. The second stage of the ASSEMBLY algorithm has as input the output string set of the first stage, and utilizes the idea of the certification to boost the greedy choices to string pairs that are also certified. It is terminated when the

cardinality of the remaining string set became equal to the parameter b. The third stage of the ASSEMBLY algorithm is a restricted backtracking procedure. Its input is the output string set of the second stage. It excludes some solutions based on a learning process, and then performs an exhaustive search through the rest solution space. In this way, it tries to balance between time efficiency and accuracy.

The ASSEMBLY algorithm is tested both on domains of random and real-world SSP instances. The first is taken by random string generators over specific distribution specifications, and the second is taken by DNA sequence databases. A number of instances of each domain are used as input to a learning procedure to specify the parameters b, c and the excluded solutions of the third stage, and the rest is used to test the ASSEMBLY algorithm. Every version of the ASSEMBLY algorithm was tested on the domain that was used for its training, but also to the other domain. The results show that ASSEMBLY performs significantly better when trained on the same domain it was tested, whereas the randomly trained version has poor performance on the real-world instances. The version of the algorithm that is trained by a successfully sequenced DNA molecule achieves a very high accuracy and effectiveness to instances of the same domain. This indicates that a successfully sequenced *part* of a DNA molecule can be used to significantly speed up the sequencing of the *whole* DNA molecule. The sequenced part can act as input to the learning procedure to determine the suitable parameter values, and the whole molecule can then be obtained by the ASSEMBLY algorithm with high accuracy and significantly sped up. The running time of the algorithm mainly depends on the running time of its third stage, which may be exponential. The tested instances suggest a sub-exponential growth of search space for this stage, but experiments on larger SSP instances are needed to conjecture a polynomial growth.

9.3 Genetic Algorithm

Some heuristic algorithms are inspired by evolutionary processes in nature. **Genetic algorithms** [22] belong to this class of heuristics. They are search methods that simulates the evolution process of natural selection, and used in many scientific fields to solve optimization problems. In a genetic algorithm for an optimization problem, a **population**, that is a collection of candidate solutions, called **individuals**, is evolved to reach better solutions. The evolution happens in **generations** that reflect the alternations to the population. During each generation the **fitness** of each individual in the population is evaluated proportionally to the suitability of its value for the objective function of the optimization problem. The most suitable individuals are selected to perpetuate their kind by recombining their genomes, i.e., their solutions, in specific points and by possibly randomly mutated. In this way, a new population is formed and the procedure is repeated for the next generation. Commonly, the algorithm terminates when either a maximum number of generations is produced, or a satisfactory fitness level is reached.

A genetic algorithm for the SSP is described in [66]. The input of the algorithm is a set S of strings specifying the SSP instance. The genome of each individual in the population is represented as a collection of strings from S in specific order, such that it is a candidate solution to the SSP instance. A crucial point of the algorithm is that an individual may not contain all the strings from S or may contain duplicate copies of the same string. This choice makes the output not a permutation of the strings in S giving in this way new potentials to the algorithm. The algorithm was tested to SSP instances over an alphabet of cardinality 2 using specific values for the parameters of the population size, the number of generations, and the mutation rates. The input instances were generated randomly following the DNA sequencing procedure. The experimental results show that when the number of the strings is 50 the genetic algorithm is better than GREEDY, while its dominance is lost when the number of the strings becomes 80.

9.4 Coevolutionary Algorithm

Coevolutionary algorithms also belong in the class of the biologically inspired evolutionary procedures. They generalize the idea of the genetic algorithms involving individuals from more than one species. **Coevolution** in nature refers to the simultaneous evolution of two or more species with coupled fitness. There are two different kinds of coevolution: the **competitive** one where the purpose is to obtain exclusivity on a limited resource, and the **cooperative** one where the purpose is to gain access to some hard to attain resource. In cooperative coevolutionary algorithms there is a number of independently evolving species representing components of potential solutions which together form complex structures to solve an optimization problem. Complete solutions are obtained by assembling representative members of each species. The fitness of each individual depends on the quality of the complete solutions it participates in. Therefore, the fitness function measures how well an individual cooperates with individuals from other species to solve the optimization problem.

A cooperative coevolutionary algorithm adjusted to the SSP is presented in [66]. It is based on populations of two species that evolve simultaneously. The first population contains prefixes of candidate solutions of the SSP instance, and the second population contains candidate suffixes. Each species population evolves separately and the only interaction between the two populations is through the fitness function. Computation experiments similar to those for the genetic algorithm show that this algorithm performs at least as good as the genetic algorithm and that requires less computation time since the required involved populations are smaller and the convergence is faster. Compared with GREEDY, it reaches better solutions after a number of generations both in experiments with 50 and 80 input strings.

An attempt to combine the cooperative coevolutionary approach with natural greediness concludes to the design of an improved method, which incorporates both parallelism and greed as described in [66]. The method consists of three

stages. In the first stage, three parallel and independent runs of the cooperative coevolutionary algorithm operate, returning as output the populations of the prefixes and suffixes, instead of the merge string of the best representatives. Also the GREEDY algorithm runs and its solution is split into a prefix and a suffix. In the second stage, two new collections of prefixes and suffixes are generated. The first contains the best $\frac{1}{3}$ individuals of the prefix population of each cooperative coevolutionary run, and the prefix of the greedy solution. The second is constructed similarly by the corresponding suffixes. In the third stage the cooperative coevolutionary algorithm runs with the two collections constructed in the second stage as initial populations, instead of random populations. The experimental results show that this algorithm performs better than the simple cooperative coevolutionary algorithm even if the cardinality of its populations and the number of its generations in each stage are quite smaller.

9.5 Preserving Favoured Subsolutions

An extension of the genetic algorithm motivated by the desire to address the failure of this algorithm in specific domains, is the PUZZLE algorithm described in [67]. It is designed to improve the performance of the genetic algorithm on relative ordering problems, i.e., problems where the order between genes is crucial instead of their global locus in the genome. Corresponding genes to strings and genome to superstring the SSP is exactly a problem of this kind. The main idea behind the PUZZLE algorithm is to preserve good subsolutions found by the genetic algorithm by choosing carefully the combination points between two solutions. In this way, it promotes the assembly of increasingly larger good building blocks from different individuals, a result that explains also the name of this algorithm.

Two different populations are evolved in the PUZZLE algorithm. A population of solutions (s-population) and a population of building subsolutions (b-population). Accordingly, we have the p-individuals and the b-individuals. Notice that this situation is completely different from the one described for the cooperative coevolutionary algorithm, since here the two populations are not complementary components of a complete solution. The interaction between these two populations is performed differently in each way. The fitness of a b-individual depends on the fitness of the s-individuals that contain it, while the choice of the combination points in s-individuals is affected by the b-individuals that contain these points.

The PUZZLE algorithm was compared with the genetic algorithm since it is its extension and with GREEDY. Experimental results with SSP instances over alphabet of cardinality 2 show that the PUZZLE algorithm outperforms both GREEDY and genetic algorithm, producing shorter superstrings in the average. The result is obtained by instances with 50 and 80 strings. Comparing with the cooperative coevolutionary algorithm, PUZZLE is better for instances with 50 strings, whereas it is worse for instances with 80 strings.

In [67], two expansions of the PUZZLE algorithm are discussed. The first one is a direct combination of PUZZLE with cooperative coevolution. The two ideas of the complementary components in different populations and of the solutions and subsolutions also in different populations are combined to derive a new algorithm. During this algorithm four populations are evolved:

1. population of prefixes,
2. population of suffixes,
3. population of building sub-prefixes, and
4. population of building sub-suffixes,

where the interaction between 1 and 2 operates according to cooperative coevolutionary algorithm, and the interaction between 1, 3 and between 2, 4 operates according to algorithm PUZZLE. The second expansion of PUZZLE involves ideas from **messy genetic algorithms** [21]. They are iterative optimization algorithms that use local search techniques, adaptive representation of the genomes, and decision sampling strategies.

9.6 Discrete Neural Network

In computer science, **neural networks** are learning programming structures that simulate the function of biological neural networks as the one constitutes the human brain. They are composed of artificial **neurons** and connections between them called **synapses**. Neural networks are used for solving artificial intelligence problems as well as combinatorial optimization problems.

A discrete neural network used for solving the SSP is described in [35]. Discreteness concerns the values that neurons can handle. In general, it is formed by n neurons, where the state of each neuron $i \in \mathrm{I}_n$ is defined by its output v_i. The vector $V = (v_1, v_2, \ldots, v_n)$ whose components are the corresponding neuron outputs is called the **state vector**. The energy of each state vector is given by the **energy function** of the network. The aim of the network is to minimize the energy function via its learning operation which happens in iterations. The energy function usually coincides with the objective function of the optimization problem to solve, such that a local minimum of the former is also a local, and possibly global, optimum to the latter. In the case of the SSP, and given a string set S, any feasible vector of the neural network represents an order of the strings in S, utilizing the permutation expression of the SSP solutions. So, feasible state vectors are those correspond to permutations, and $v_i = k$ means that string s_k is placed in the i-th place in the superstring. Notice that there is an one-to-one correspondence between neurons and strings in S. In each learning iteration, the neural network searches different solutions using neuron updating schemes. Given a vector $V = (v_1, v_2, \ldots, v_n)$ corresponding to the current state, and two neurons i and j, $1 \leq i < j \leq n$, the network considers updates to the following different states:

- $(v_1, \ldots, v_i, v_{i+1}, \ldots, v_j, v_{j+1}, \ldots, v_n)$,
- $(v_1, \ldots, v_i, v_{j+1}, \ldots, v_n, v_{i+1}, \ldots, v_j)$,

- $(v_{i+1}, \ldots, v_j, v_1, \ldots, v_i, v_{j+1}, \ldots, v_n)$,
- $(v_{i+1}, \ldots, v_j, v_{j+1}, \ldots, v_n, v_1, \ldots, v_i)$,
- $(v_{j+1}, \ldots, v_n, v_1, \ldots, v_i, v_{i+1}, \ldots, v_j)$, and
- $(v_{j+1}, \ldots, v_n, v_{i+1}, \ldots, v_j, v_1, \ldots, v_i)$,

that correspond to the combinations of the three parts that the state vector is separated into according to the specific two neurons. For each of these candidate solutions the one that decrease mostly the energy function value is selected as the next network state. This procedure is repeated until convergence is detected, thus a state vector is found where the updates with all pairs of neurons do not cause any change. Due to the used update scheme, the network remains in a feasible state along all iterations. Once the network converges, the stable state represents a local minimum of the energy function which is equivalent to a local maximum of the total overlap between the strings in S.

Experimental results are performed with SSP instances for strings of fixed and variable lengths. The neural network algorithm runs 100 times for each instance and its results were compared with those of GREEDY. In experiments with fixed string length, neural network outperforms GREEDY in most cases on average, and always on best results. In experiments with variable string lengths, neural network outperforms GREEDY both on average and best results.

9.7 GRASP with Path Relinking

A Greedy Randomized Adaptive Search Procedure (GRASP) is an iterative metaheuristic for combinatorial optimization, which is implemented as a multi-start procedure where each iteration is made up of a construction phase and a local search phase. The first phase constructs a randomized greedy solution, while the second phase starts at this solution and applies repeated improvement until a locally optimal solution is found. The procedure continues until a termination condition is satisfied such as a maximum number of iterations. The best solution over all iterations is kept as the final result. GRASP seems to produce good quality solutions for a wide variety of combinatorial optimization problems. A survey on GRASP can be found in [47] while an annotated bibliography in [12]. Path Relinking (PR) [19] is an approach to integrate intensification and diversification strategies in search for optimal solutions. PR in the context of GRASP is introduced in [32] as a memory mechanism for utilizing information on previously found good solutions.

In [17], an implementation of GRASP with PR for solving the SSP is presented. It solves large scale SSP instances of more than 1,000 strings and outperforms the GREEDY algorithm in the majority of the tested instances. The proposed method is able to provide multiple near-optimum solutions that is of practical importance for the DNA sequencing, and admits a natural parallel implementation. Extended computational experiments on a set of SSP instances with known optimal solutions, produced by using the integer programming formulation presented in Sect. 5.2, indicate that the new method finds the optimum in most of the cases, and its average error relative to the optimum is close to zero.

10 Asymptotic Behaviour

It can be observed a discrepancy between the theoretical results from the worst-case analysis and the experimental observations from the approximation and heuristic algorithms for the SSP. A possible explanation for this fact is given by the average-case analysis for the problem.

The asymptotic behaviour of the compression achieved by an optimal superstring is analysed in [1] under a certain probability model for the lengths of the strings and the letter distribution in them. The average optimal compression of n strings tends to $\frac{n \log n}{H_\mu}$, where $H_\mu = -\sum_{i=1}^{m} p(a_i) \log p(a_i)$ is the Shannon entropy of the choosing law μ for the letters from the alphabet to construct the strings.

The asymptotic behaviour of some algorithms for the SSP is based on the above result and explains the good performance of the greedy strategies. In [13], the algorithms GREEDY, MGREEDY, and NAIVE are analysed in a probabilistic framework and it is proved that they are asymptotically optimal. In [65], the results of the asymptotic behaviour are extended to the TGREEDY and DIMATCH algorithms, after the observation that the performance of TGREEDY is never worse than that of MGREEDY, and that the intermediate result of the maximum directed matching in DIMATCH coincides actually with the result of MGREEDY (see Theorem 12). The steps of DIMATCH up to the construction of the maximum directed matching are analysed in a probabilistic way with the additional assumption that all strings have the same length, and the asymptotic optimality of these algorithms is established.

By the complexity results in Sect. 3.2, we know that there is not PTAS for the SSP for both performance measures unless P = NP. In [48], a *probabilistic* PTAS for the SSP that achieves a $(1+\varepsilon)$-approximation in *expected* polynomial time, for every $\varepsilon > 0$, is presented. This algorithm

1. either returns a possibly non-optimal solution, the solution of GREEDY, in polynomial time,
2. or returns an optimal solution, via a maximum Hamiltonian path on the associated overlap graph, in non-polynomial time.

Under certain conditions in the data of the SSP instance, in the first case GREEDY has asymptotic approximation ratio $1+\varepsilon$ with respect to the length measure, and in the second case the *expected* running time of finding the maximum Hamiltonian path can be polynomial, since it depends on the time spent when it is executed and its execution probability. Analysing these situations, for a random input the algorithm has approximation ratio $1+\varepsilon$ with respect to the length measure and polynomial expected running time.

11 Smoothed Analysis

The classical complexity analysis implies that the SSP is a hard problem in the worst case. The average-case analysis explains the effectiveness of greedy strategies under suitable probability models which are far from reality. In addition to these two

frameworks, the latest developed smoothed analysis explains why greed works so well for the SSP in real-world instances of the DNA sequencing practice. Smoothed analysis is introduced in [52] to demonstrate the fact that some algorithms like the simplex algorithm run in exponential time in the worst case, but in practice they are very efficient.

In [36], the smoothed analysis of the GREEDY algorithm is realized, making the observation that the asymptotic optimal behaviour of the greedy techniques is due to the fact that the random strings do not have large overlaps, and so the concatenation of the strings is not much longer than the shortest common superstring. However, the practical instances arising from DNA assembly are not random and the input strings have significantly large overlaps. By defining small and natural perturbations that represent the mutations of the DNA sequences during evolution, it is proved that for any given instance S of the SSP, the average approximation ratio of the GREEDY algorithm on a small random perturbation of S is $1 + o(1)$. This result points out that the approximation inefficiency of SSP instances indicating by the Max-SNP-hardness result can be destroyed by a very small perturbation. As very handily noted, if there had been a hard instance for the DNA assembly problem in history, the hardness would have likely been destroyed by the random mutations of the DNA sequences during the evolution. This result makes the SSP a characteristic case where the complexity is different in the worst-case analysis and in the smoothed analysis.

Acknowledgements This research has been funded by the European Union (European Social Fund—ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)—Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

References

1. Alexander, K.S.: Shortest common superstrings for strings of random letters. In: Crochemore, M., Gusfield, D. (eds.) Combinatorial Pattern Matching. Lecture Notes in Computer Science, vol. 807, pp. 164–172. Springer, Berlin (1994)
2. Armen, C., Stein, C.: Improved length bounds for the shortest superstring problem. In: Akl, S., Dehne, F., Sack, J.R., Santoro, N. (eds.) Algorithms and Data Structures. Lecture Notes in Computer Science, vol. 955, pp. 494–505. Springer, Berlin (1995)
3. Armen, C., Stein, C.: Short superstrings and the structure of overlapping strings. J. Comput. Biol. **2**(2), 307–332 (1995)
4. Armen, C., Stein, C.: A $2\frac{2}{3}$-approximation algorithm for the shortest superstring problem. In: Hirschberg, D., Myers, G. (eds.) Combinatorial Pattern Matching. Lecture Notes in Computer Science, vol. 1075, pp. 87–101. Springer, Berlin (1996)
5. Arora, S., Lund, C., Motwani, R., Sudan, M., Szegedy, M.: Proof verification and the hardness of approximation problems. J. ACM **45**(3), 501–555 (1998)
6. Berger, B., Rompel, J., Shor, P.W.: Efficient NC algorithms for set cover with applications to learning and geometry. J. Comput. Syst. Sci. **49**(3), 454–477 (1994)
7. Bläser, M.: An 8/13-approximation algorithm for the asymmetric maximum TSP. In: Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '02), pp. 64–73. Society for Industrial and Applied Mathematics, Philadelphia (2002)

8. Blum, A., Jiang, T., Li, M., Tromp, J., Yannakakis, M.: Linear approximation of shortest superstrings. J. ACM **41**, 630–647 (1994)
9. Breslauer, D., Jiang, T., Jiang, Z.: Rotations of periodic strings and short superstrings. J. Algorithm. **24**, 340–353 (1997)
10. Chirico, N., Vianelli, A., Belshaw, R.: Why genes overlap in viruses. Proc. R. Soc. B. Biol. Sci. **277**(1701), 3809–3817 (2010)
11. Czumaj, A., Gąsieniec, L., Piotrów, M., Rytter, W.: Sequential and parallel approximation of shortest superstrings. J. Algorithm. **23**, 74–100 (1997)
12. Festa, P., Resende, M.: GRASP: An annotated bibliography. In: Ribeiro, C., Hansen, P. (eds.) Essays and Surveys in Metaheuristics. Operations Research/Computer Science, pp. 325–367. Kluwer Academic, Dordecht (2002)
13. Frieze, A., Szpankowski, W.: Greedy algorithms for the shortest common superstring that are asymptotically optimal. Algorithmica **21**, 21–36 (1998)
14. Gallant, J.K.: String compression algorithms. Ph.D. thesis, Princeton (1982)
15. Gallant, J., Maier, D., Storer, J.A.: On finding minimal length superstrings. J. Comput. Syst. Sci. **20**(1), 50–58 (1980)
16. Gerver, M.: Three-valued numbers and digraphs. Kvant **1987**(2), 32–35 (1987)
17. Gevezes, T., Pitsoulis, L.: A greedy randomized adaptive search procedure with path relinking for the shortest superstring problem. J. Comb. Optim. (2013) doi: 10.1007/s10878-013-9622-z
18. Gingeras, T., Milazzo, J., Sciaky, D., Roberts, R.: Computer programs for the assembly of DNA sequences. Nucleic Acids Res. **7**(2), 529–543 (1979)
19. Glover, F., Laguna, M.: Tabu Search. Kluwer Academic, Norwell (1997)
20. Goldberg, M.K., Lim, D.T.: A learning algorithm for the shortest superstring problem. In: Proceedings of the Atlantic Symposium on Computational Biology and Genome Information and Technology, pp. 171–175 (2001)
21. Goldberg, D., Deb, K., Korb, B.: Messy genetic algorithms: Motivation, analysis, and first results. Complex Syst. **3**, 493–530 (1989)
22. Holland, J.H.: Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor (1975)
23. Huffman, D.A.: A method for the construction of minimum-redundancy codes. Proc. Inst. Radio Eng. **40**(9), 1098–1101 (1952)
24. Ilie, L., Popescu, C.: The shortest common superstring problem and viral genome compression. Fundam. Inform. **73**, 153–164 (2006)
25. Ilie, L., Tinta, L., Popescu, C., Hill, K.A.: Viral genome compression. In: Mao, C., Yokomori, T. (eds.) DNA Computing. Lecture Notes in Computer Science, vol. 4287, pp. 111–126. Springer, Berlin (2006)
26. Jenkyns, T.A.: The greedy travelling salesman's problem. Networks **9**(4), 363–373 (1979)
27. Jiang, T., Li, M.: Approximating shortest superstrings with constraints. Theor. Comput. Sci. 134(2), 473–491 (1994)
28. Kaplan, H., Shafrir, N.: The greedy algorithm for shortest superstrings. Inf. Process. Lett. **93**, 13–17 (2005)
29. Kaplan, H., Lewenstein, M., Shafrir, N., Sviridenko, M.: Approximation algorithms for asymmetric TSP by decomposing directed regular multigraphs. J. ACM **52**, 602–626 (2005)
30. Karp, R.M.: Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W. (eds.) Complexity of Computer Computations, pp. 85–103. Plenum Press, New York (1972)
31. Kosaraju, S.R., Park, J.K., Stein, C.: Long tours and short superstrings. In: Proceedings of the 35th Annual Symposium on Foundations of Computer Science, pp. 166–177. IEEE Computer Society, Washington, DC (1994)
32. Laguna, M., Martí, R.: GRASP and path relinking for 2-layer straight line crossing minimization. INFORMS J. Comput. **11**, 44–52 (1999)
33. Lesk, A.M.: Computational Molecular Biology. Sources and Methods for Sequence Analysis. Oxford University Press, Oxford (1988)
34. Li, M.: Towards a DNA Sequencing Theory (Learning a String), vol. 1, pp. 125–134. IEEE Computer Society, Los Alamitos (1990)

35. López-Rodríguez, D., Mérida-Casermeiro, E.: Shortest common superstring problem with discrete neural networks. In: Kolehmainen, M., Toivanen, P., Beliczynski, B. (eds.) Adaptive and Natural Computing Algorithms. Lecture Notes in Computer Science, vol. 5495, pp. 62–71. Springer, Berlin (2009)
36. Ma, B.: Why greed works for shortest common superstring problem. In: Ferragina, P., Landau, G. (eds.) Combinatorial Pattern Matching. Lecture Notes in Computer Science, vol. 5029, pp. 244–254. Springer, Berlin (2008)
37. Maier, D., Storer, J.A.: A note on the complexity of the superstring problem. Technical Report 233, Computer Science Laboratory, Princeton University, Princeton (1977)
38. Middendorf, M.: More on the complexity of common superstring and supersequence problems. Theor. Comput. Sci. **125**(2), 205–228 (1994)
39. Middendorf, M.: Shortest common superstrings and scheduling with coordinated starting times. Theor. Comput. Sci. **191**(1–2), 205–214 (1998)
40. Miller, C.E., Tucker, A.W., Zemlin, R.A.: Integer programming formulation of traveling salesman problems. J. ACM **7**, 326–329 (1960)
41. Ott, S.: Lower bounds for approximating shortest superstrings over an alphabet of size 2. In: Widmayer, P., Neyer, G., Eidenbenz, S. (eds.) Graph-Theoretic Concepts in Computer Science. Lecture Notes in Computer Science, vol. 1665, pp. 55–64. Springer, Berlin (1999)
42. Papadimitriou, C.H., Steiglitz, K.: Combinatorial optimization: algorithms and complexity. Prentice-Hall, Englewood Cliffs (1982)
43. Papadimitriou, C.H., Yannakakis, M.: Optimization, approximation, and complexity classes. J. Comput. Syst. Sci. **43**(3), 425–440 (1991)
44. Papadimitriou, C.H., Yannakakis, M.: The traveling salesman problem with distances one and two. Math. Oper. Res. **18**(1), 1–11 (1993)
45. Peltola, H., Söderlund, H., Ukkonen, E.: SEQAID: a DNA sequence assembling program based on a mathematical model. Nucleic Acids Res. **12**(1), 307–321 (1984)
46. Pevzner, P.A., Waterman, M.S.: Open Combinatorial Problems in Computational Molecular Biology, p. 158. IEEE Computer Society, Los Alamitos (1995)
47. Pitsoulis, L., Resende, M.: Greedy randomized adaptive search procedures. In: Pardalos, P., Resende, M. (eds.) Handbook of Applied Optimization, pp. 178–183. Oxford University Press, Oxford (2002)
48. Plociennik, K.: A probabilistic PTAS for shortest common superstring. In: Proceedings of the 34th International Symposium on Mathematical Foundations of Computer Science 2009 (MFCS '09), pp. 624–635. Springer, Berlin (2009)
49. Reif, J.H.: Synthesis of Parallel Algorithms, 1st edn. Morgan Kaufmann, San Francisco (1993)
50. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
51. Shapiro, M.B.: An algorithm for reconstructing protein and RNA sequences. J. ACM **14**, 720–731 (1967)
52. Spielman, D., Teng, S.H.: Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In: Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing (STOC '01), pp. 296–305. ACM, New York (2001)
53. Staden, R.: Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res. **10**(15), 4731–4751 (1982)
54. Storer, J.A.: Data compression: Methods and theory. Computer Science Press, New York (1988)
55. Storer, J.A., Szymanski, T.G.: The macro model for data compression (extended abstract). In: Proceedings of the Tenth Annual ACM Symposium on Theory of Computing (STOC '78), pp. 30–39. ACM, New York (1978)
56. Storer, J.A., Szymanski, T.G.: Data compression via textual substitution. J. ACM **29**, 928–951 (1982)
57. Sweedyk, Z.: A $2\frac{1}{2}$-approximation algorithm for shortest superstring. SIAM J. Comput. **29**, 954–986 (1999)

58. Tarhio, J., Ukkonen, E.: A greedy approximation algorithm for constructing shortest common superstrings. Theor. Comput. Sci. **57**(1), 131–145 (1988)
59. Tarjan, R.E.: Data Structures and Network Algorithms. Society for Industrial and Applied Mathematics, Philadelphia (1983)
60. Teng, S.H., Yao, F.: Approximating Shortest Superstrings, pp. 158–165. IEEE Computer Society, Los Alamitos (1993)
61. Timkovskii, V.G.: Complexity of common subsequence and supersequence problems and related problems. Cybern. Syst. Anal. **25**, 565–580 (1989)
62. Turner, J.S.: Approximation algorithms for the shortest common superstring problem. Inf. Comput. **83**, 1–20 (1989)
63. Valiant, L.G.: A theory of the learnable. Commun. ACM **27**(11), 1134–1142 (1984)
64. Vassilevska, V.: Explicit inapproximability bounds for the shortest superstring problem. In: Jedrzejowicz, J., Szepietowski, A. (eds.) Mathematical Foundations of Computer Science 2005. Lecture Notes in Computer Science, vol. 3618, pp. 793–800. Springer, Berlin (2005)
65. Yang, E., Zhang, Z.: The shortest common superstring problem: Average case analysis for both exact and approximate matching. IEEE Trans. Inf. Theory **45**(6), 1867–1886 (1999)
66. Zaritsky, A., Sipper, M.: Coevolving solutions to the shortest common superstring problem. Biosystems **76**(1–3), 209–216 (2004)
67. Zaritsky, A., Sipper, M.: The preservation of favored building blocks in the struggle for fitness: The puzzle algorithm. Evol. Comput. **8**(5), 443–455 (2004)

Computational Comparison of Convex Underestimators for Use in a Branch-and-Bound Global Optimization Framework

Yannis A. Guzman, M.M. Faruque Hasan, and Christodoulos A. Floudas

1 Introduction

Applications that require the optimization of nonlinear functions involving nonconvex terms include reactor network synthesis, separations design and synthesis, robust process control, batch process design, protein folding, and molecular structure prediction. Deterministic global optimization algorithms can proceed to determine the global minimum of a nonconvex nonlinear optimization model (NLP) through a branch-and-bound framework. Node fathoming occurs through the assignment of lower and upper bounds over each node's subdomain. Lower bounds are generated through convexification to yield a convex NLP at each node. The tightness of the resulting underestimator depends on the method of convexification and how its strengths align with the characteristics of the function over the subdomain. In practice, the performance of the algorithm relies on tight lower bounds to increase the efficiency and frequency of fathoming and pruning for rapid convergence to the global optimum.

There are certain nonconvex functional forms for which explicit convex envelopes are known or can be derived, including bilinear [6, 18], trilinear [19, 20], and fractional terms [17, 27, 28]. In cases where either the convex envelope or an alternative tight relaxation do not exist, or can only be generated with prohibitive computational cost, a general method for the relaxation and convexification of the function can be employed. In [16], a novel convexification method was presented that generates the convex relaxation $\mathcal{L}(\mathbf{x})$ of any $\mathcal{C}^2$-continuous function $f(\mathbf{x})$ through the subtraction of a positive quadratic term with an α parameter that is designed to dominate the nonconvexities of $f(\mathbf{x})$:

$$\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \alpha \sum_i \left(x_i^U - x_i\right)\left(x_i - x_i^L\right). \tag{1}$$

Y.A. Guzman • M.M.F. Hasan • C.A. Floudas (✉)
Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA
e-mail: yannis@titan.princeton.edu; faruque@titan.princeton.edu; floudas@titan.princeton.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_11,

When utilized in a branch-and-bound framework, the method, now called α-branch-and-bound (αBB), can guarantee ε-convergence to the global minimum within a finite number of iterations [1, 7]. This chapter will explore the convexification strength of the nonuniform αBB underestimator (Sect. 2.1), as well as the number of competing methods designed to provide tight, convex underestimators, including:

- piecewise αBB (P-αBB, Sect. 2.2)
- generalized αBB (G-αBB, Sect. 2.3),
- nondiagonal αBB (ND-αBB, Sect. 2.4),
- Brauer αBB (B-αBB, Sect. 2.5),
- Rohn+E αBB (RE-αBB, Sect. 2.6),
- and the moment approach (f_{dk}, Sect. 2.7).

Their performance will be gauged via convexification of 20 multivariate, box-constrained, nonconvex functions whose global minima are known [13]. Section 3 of this chapter outlines implementation details of all the methods, Sect. 4 discusses the results, and Sect. 5 presents our conclusions.

2 Overview of Methods

2.1 Method 1: Nonuniform Diagonal Perturbation I (αBB)

Androulakis et al. [7] presented the form of the αBB underestimator with nonuniform parameters α_i:

$$\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x}) = f(\mathbf{x}) - \sum_i \alpha_i \left(x_i^U - x_i\right)\left(x_i - x_i^L\right), \tag{2}$$

where $f(\mathbf{x})$ is a nonconvex function and $\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x})$ is the resulting αBB underestimator. As previously stated, the structure of the term subtracted from $f(\mathbf{x})$ guarantees that $\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x}) \leq f(\mathbf{x})$ over the entire domain given adequate parameters $\alpha_i \geq 0$. The tightness of $\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x})$ relies on determining small but sufficient α_i parameters that yield a tight but guaranteed convex underestimator, as the maximum separation distance $d_{\max}$ between $f(\mathbf{x})$ and $\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x})$ is directly proportional to the α_i parameters [16]:

$$d_{\max}\left(\mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x})\right) = \max_{\mathbf{x}\in\left[\mathbf{x}^L,\mathbf{x}^U\right]} \left(f(\mathbf{x}) - \mathcal{L}_{\alpha\mathrm{BB}}(\mathbf{x})\right) = \frac{1}{4}\sum_i \alpha_i \left(x_i^U - x_i^L\right)^2. \tag{3}$$

Utilizing the eigenvalues of the Hessian matrix H as a means to guarantee positive semidefiniteness and thus convexity, the subtracted positive quadratic term yields a nonuniform diagonal shift from the original function's Hessian matrix:

$$H_{\mathcal{L}_{\alpha\mathrm{BB}}}(\mathbf{x}) = H_f(\mathbf{x}) + 2\Delta, \tag{4}$$

where the diagonal elements of Δ are α_i and the nondiagonal elements are 0. As a means of estimating H over the entire domain, Androulakis et al. [7] proposed utilizing interval arithmetic in deriving bounds on each Hessian term h_{ij}, yielding an interval Hessian matrix H' of terms $h'_{ij} = \left[\underline{h}_{ij}, \overline{h}_{ij}\right]$ and a relaxed problem for calculating the elements of Δ that yield the tightest convex underestimator.

Of the many methods for calculating α_i that are proposed by Adjiman et al. [3] and explored in [2, 3], computational studies indicated that a scaled method derived from Gershgorin's Circle Theorem [12], referred to as the scaled Gershgorin method, showed consistently strong performance in convergence towards the global optimum with the αBB algorithm:

$$\alpha_i = \max\left\{0, -\frac{1}{2}\left(\underline{h}_{ii} - \sum_{j\neq i}\left|h'_{ij}\right|\frac{d_j}{d_i}\right)\right\} \qquad \forall i, \tag{5}$$

where $\left|h'_{ij}\right| = \max\left\{\left|\underline{h}_{ij}\right|, \left|\overline{h}_{ij}\right|\right\}$. There is a degree of freedom when choosing the scaling factors d_i; here, $d_i = x_i^U - x_i^L$ is chosen as suggested and supported in [2, 3]. For a detailed look at theory and applications of the αBB method, the reader is directed to [11].

2.2 Method 2: Piecewise Diagonal Perturbation (P-αBB)

The α_i parameters of the αBB method dominate the nonconvexities of $f(\mathbf{x})$ over the entire domain. A natural extension, introduced by Meyer and Floudas [21], attempts to produce a tighter underestimator by generating a once-differentiable, piecewise quadratic underestimator after subdividing each variable x_i into N_i subdomains:

$$\mathcal{L}_{\text{P-}\alpha\text{BB}}(\mathbf{x}) = f(\mathbf{x}) - q(\mathbf{x}), \tag{6}$$

where

$$q(\mathbf{x}) = \sum_i (\alpha_i^k (x_i^k - x_i)(x_i - x_i^{k-1}) + \beta_i^k x_i + \gamma_i^k) \qquad \text{for } x_i \in [x_i^{k-1}, x_i^k],$$

the kth interval represents $[x_i^{k-1}, x_i^k]$, and $[x_i^L, x_i^U] = [x_i^0, x_i^{N_i}]$. The system of equations yielded by requiring $q(\mathbf{x})$ to be smooth, continuous, and match $f(\mathbf{x})$ at the vertices of the domain produces the following analytical form for parameters β_i^k and γ_i^k:

$$\beta_i^1 = -\frac{\sum_{k=1}^{N_i-1} s_i^k (x_i^U - x_i^k)}{x_i^U - x_i^L} \qquad \forall i \tag{7}$$

$$\beta_i^k = \beta_i^1 + \sum_{j=1}^{k-1} s_i^j \qquad \forall i, k = 2, \ldots, N_i \tag{8}$$

$$\gamma_i^k = -\beta_i^1 x_i^0 - \sum_{j=1}^{k-1} s_i^j x_i^j \qquad \forall i, k = 1, \ldots, N_i \tag{9}$$

where $s_i^k = -\alpha_i^k (x_i^k - x_i^{k-1}) - \alpha_i^{k+1} (x_i^{k+1} - x_i^k)$. Subdividing the domain reduces the cumulative effects of highly nonconvex regions, thus yielding a spline underestimator $\mathcal{L}_{\text{P-}\alpha\text{BB}}(\mathbf{x})$ that can tighten at each subdomain while meeting continuity and smoothness requirements.

2.3 Method 3: Nonquadratic Diagonal Perturbation (G-αBB)

With the goal of generating a diagonal perturbation matrix whose resulting underestimator is at least as tight as $\mathcal{L}_{\alpha\text{BB}}(\mathbf{x})$, Akrotirianakis and Floudas [4, 5] introduced a generalized separable but nonquadratic form for the relaxation term of the αBB underestimator:

$$\mathcal{L}_{\text{G-}\alpha\text{BB}}(\mathbf{x}) = f(\mathbf{x}) - \sum_i \left(1 - e^{\gamma_i \left(x_i^U - x_i\right)}\right) \left(1 - e^{\gamma_i \left(x_i - x_i^L\right)}\right) \tag{10}$$

where γ_i is selected by solving the system of nonlinear equations

$$\ell_i + \gamma_i^2 + \gamma_i^2 e^{\gamma_i (x_i^U - x_i^L)} = 0, \qquad i = 1, 2, \ldots, n. \tag{11}$$

Here, $\ell_i \leq 0$ and represents a measure of the nonconvexity of $f(\mathbf{x})$, which can be estimated via the scaled Gershgorin method and related to α_i of Method 1 (αBB):

$$\ell_i = -2\overline{\alpha}_i. \tag{12}$$

The α_i parameters from the αBB method provide an upper bound on G-αBB α_i parameters and are hence denoted with an overbar. Parameters γ_i and α_i are related by the equation

$$\gamma_i = \frac{2 \log\left(1 + \sqrt{\alpha_i}\left(x_i^U - x_i^L\right)/2\right)}{x_i^U - x_i^L}. \tag{13}$$

The parameters obtained from (11) represent lower bounds on the eventual parameters of the convex underestimator. While $\mathcal{L}_{\text{G-}\alpha\text{BB}}(\mathbf{x}; \underline{\alpha}) \geq \mathcal{L}_{\alpha\text{BB}}(\mathbf{x}; \overline{\alpha})$, the method utilizes a heuristic algorithm that attempts to prove the convexity of a given $\mathcal{L}_{\text{G-}\alpha\text{BB}}(\mathbf{x})$ while updating γ_i (and corresponding α_i) towards $\overline{\gamma}_i$ ($\overline{\alpha}_i$), whose corresponding underestimator ($\mathcal{L}_{\alpha\text{BB}}(\mathbf{x})$) is guaranteed convex. The algorithm is represented graphically in Fig. 1; if convexity is not proven before the maximum separation distance of $\mathcal{L}_{\text{G-}\alpha\text{BB}}(\mathbf{x})$ becomes greater or equal to that of $\mathcal{L}_{\alpha\text{BB}}(\mathbf{x})$, then the algorithm defaults to $\mathcal{L}_{\alpha\text{BB}}(\mathbf{x})$.

The text and pseudocode of the convexification algorithm presented in [4] stated to update all γ_i^K at every Kth outer-loop iteration. However, this step would not require comparing the maximum separation distance with the classical αBB

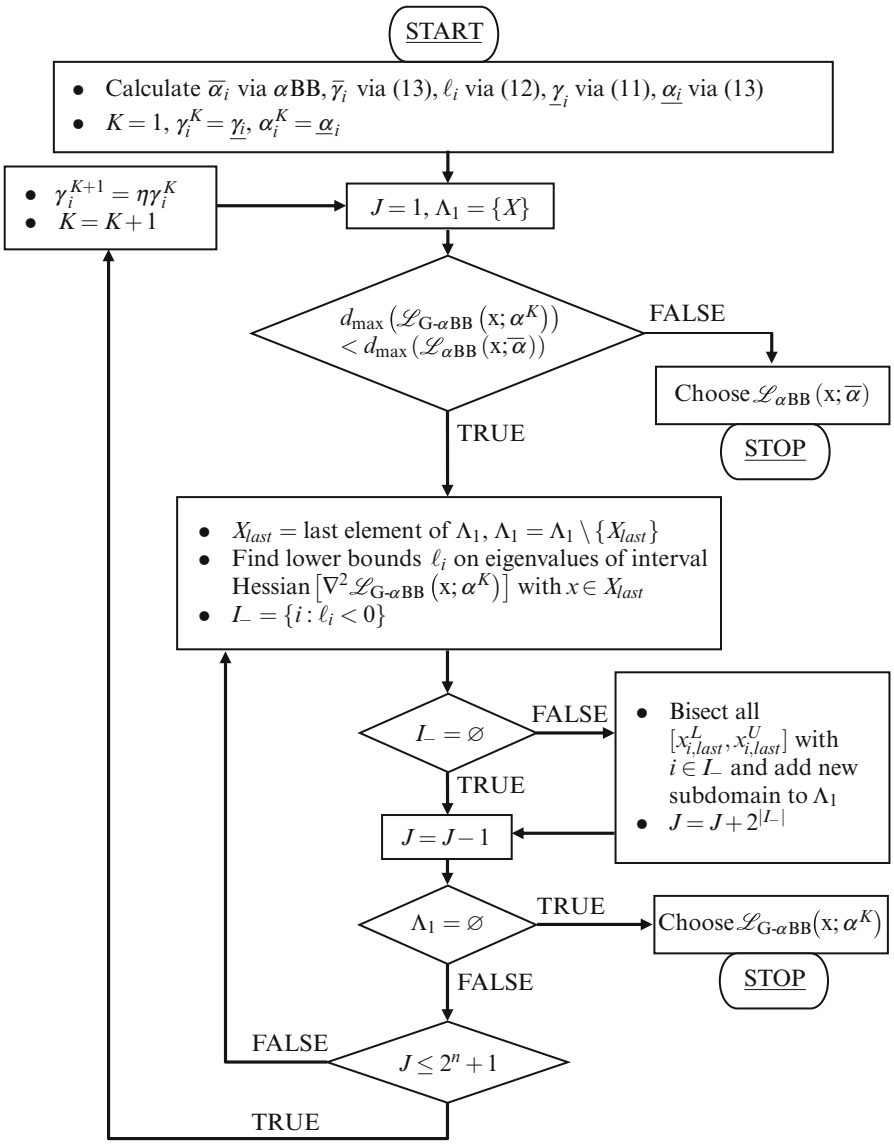


Fig. 1 Flowchart representing the original G-αBB algorithm ("v1") for verifying convexity and parameter selection, where Λ_1 is a set of subdomains, $\{X\} = \left\{[x_i^L, x_i^U], \forall i\right\}$, and $\eta > 1$ is an updating parameter. In the proposed alternative version ("v2"), the updating step would become $\gamma_i^{K+1} = \eta\gamma_i^K$, $\forall i \in I_-$

parameters from Method 1, as all α_i^K will progress towards $\overline{\alpha}_i$ at the same relative rate and the maximum separation distance criterion will be breached when $\alpha_i^K \geq \overline{\alpha}_i$. In one of the examples Akrotirianakis and Floudas [4] present, it is inferred that

only some γ_i^K needed to be updated, as only particular dimensions displayed negative lower bounds on the corresponding eigenvalues of the underestimator and were thus preventing the underestimator from being declared convex. This alternative updating scheme, applied whenever the inner loop breaks and according to the most recently obtained set I_-, would necessitate the maximum separation distance check, and represents an attempt to increase particular α_i's beyond their corresponding $\overline{\alpha}_i$ so as to obtain a validated convex underestimator with a lower maximum separation distance than $\mathcal{L}_{\alpha\text{BB}}(\mathbf{x};\overline{\alpha})$. Both of these strategies will be explored in this work, delineated as G-αBB[v1] and G-αBB[v2] for the explicit and inferred strategies, respectively.

2.4 Method 4: Nondiagonal Perturbation Elements I (ND-αBB)

The underestimator methods presented in Sects. 2.1 and 2.2 use a separable quadratic term that yields a guaranteed convex underestimator by shifting the Hessian matrix with diagonal perturbations. A natural extension to this idea is to search for a tighter underestimator by applying a perturbation matrix that contains nondiagonal elements. Skjäl et al. [25] presented criteria for when nondiagonal terms would represent a possible improvement over diagonal perturbations alone, as well as two methods for obtaining a Hessian perturbation matrix with nondiagonal terms. The perturbation matrix H^P replaces Δ, and the form of the underestimator is given as

$$\mathcal{L}_{\text{ND-}\alpha\text{BB}}(\mathbf{x}) = f(\mathbf{x}) - \sum_i \alpha_i \left(x_i^U - x_i\right)\left(x_i - x_i^L\right) + \sum_i \sum_{j>i} \left(\beta_{ij} x_i x_j + \left|\beta_{ij}\right| z_{ij}\right), \quad (14)$$

where

$$z_{ij} = \begin{cases} \max\left\{ \begin{array}{l} x_i x_j^L + x_i^L x_j - x_i^L x_j^L, \\ x_i x_j^U + x_i^U x_j - x_i^U x_j^U \end{array} \right\} & \text{if } \beta_{ij} < 0 \\ \max\left\{ \begin{array}{l} -x_i x_j^L - x_i^U x_j + x_i^U x_j^L, \\ -x_i x_j^U - x_i^L x_j + x_i^L x_j^U \end{array} \right\} & \text{if } \beta_{ij} > 0, \end{cases}$$

which can be modeled by following inequality constraints [25]:

$$\left.\begin{array}{l} z_{ij} \geq x_i x_j^L + x_i^L x_j - x_i^L x_j^L \\ z_{ij} \geq x_i x_j^U + x_i^U x_j - x_i^U x_j^U \end{array}\right\} \quad \forall i, j : j > i, \beta_{ij} < 0 \quad (15)$$

$$\left.\begin{array}{l} z_{ij} \geq -x_i x_j^L - x_i^U x_j + x_i^U x_j^L \\ z_{ij} \geq -x_i x_j^U - x_i^L x_j + x_i^L x_j^U \end{array}\right\} \quad \forall i, j : j > i, \beta_{ij} > 0. \quad (16)$$

The symmetric perturbation matrix

$$H^P = \begin{bmatrix} 2\alpha_1 & \beta_{1,2} & \cdots & \beta_{1,n} \\ \beta_{1,2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta_{n-1,n} \\ \beta_{1,n} & \cdots & \beta_{n-1,n} & 2\alpha_n \end{bmatrix}$$

is chosen such that the convexity of the resulting underestimator is guaranteed. Using this requirement as a constraint, the maximum underestimation error is minimized as per the following NLP:

$$\begin{aligned} \min_{\alpha,\beta} \quad & \sum_i \frac{\alpha_i}{4}(x_i^U - x_i^L)^2 + \sum_i \sum_{j>i} \frac{|\beta_{ij}|}{4}(x_i^U - x_i^L)(x_j^U - x_j^L) \\ \text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i - \sum_{j\neq i} |h'_{ij} + \beta_{ij}| \geq 0 \qquad \forall i \\ & \alpha_i \geq 0 \qquad \forall i \\ & \beta_{ij} = \beta_{ji} \qquad \forall i, j : j > i. \end{aligned} \tag{17}$$

Skjäl et al. [25] provided a decision tree by which H^P can be calculated without solving problem (17). First, if all off-diagonal elements of H' are centered on 0, that is, if

$$\text{mid}(h'_{ij}) = 0 \qquad \forall i, j : j > i,$$

where $\text{mid}(h'_{ij}) = (\overline{h}_{ij} + \underline{h}_{ij})/2$, then the classical nonuniform diagonal perturbation is a unique optimal solution to problem (17) and should be chosen (i.e., Method 1—αBB). Second, if the condition

$$\underline{h}_{ii} - \sum_{j\neq i} \text{rad}(h'_{ij}) \leq 0 \qquad \forall i,$$

where $\text{rad}(h'_{ij}) = (\overline{h}_{ij} - \underline{h}_{ij})/2$, holds, then an optimal solution to problem (17) is given by

$$\alpha_i = -\frac{1}{2}\left(\underline{h}_{ii} - \sum_{j\neq i} \text{rad}(h'_{ij})\right) \qquad \forall i \tag{18}$$

$$\beta_{ij} = -\text{mid}(h'_{ij}) \qquad \forall i, j. \tag{19}$$

As a last resort, problem (17) has the same optimal solutions as the following linear program (LP):

$$\begin{aligned}
\min_{\alpha,\beta} \quad & \sum_i \alpha_i \left(x_i^U - x_i^L\right)^2 / 4 - \sum_i \sum_{\substack{j>i \\ j\in J_i^+}} \beta_{ij} \left(x_i^U - x_i^L\right)\left(x_j^U - x_j^L\right) / 4 \\
& + \sum_i \sum_{\substack{j>i \\ j\in J_i^-}} \beta_{ij} \left(x_i^U - x_i^L\right)\left(x_j^U - x_j^L\right) / 4 \\
\text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i - \sum_{\substack{j\neq i \\ j\in J_i^+}} \left(\overline{h}_{ij} + \beta_{ij}\right) + \sum_{\substack{j\neq i \\ j\in J_i^-}} \left(\underline{h}_{ij} + \beta_{ij}\right) \geq 0 \qquad \forall i \\
& \alpha_i \geq 0 \qquad \forall i \\
& \beta_{ij} = \beta_{ji} \qquad \forall i, j : j > i \\
& \min\left\{0, -\text{mid}(h'_{ij})\right\} \leq \beta_{ij} \leq \max\left\{0, -\text{mid}(h'_{ij})\right\} \qquad \forall i, j,
\end{aligned} \tag{20}$$

where $J_i^+ = \left\{j : j \neq i, \text{mid}(h'_{ij}) \geq 0\right\}$ and $J_i^- = \left\{j : j \neq i, \text{mid}(h'_{ij}) < 0\right\}$.

2.5 Method 5: Nonuniform Diagonal Perturbation II (*B*-α*BB*)

The novel methods presented by Skjäl and Westerlund [24] utilize alternatives to the scaled Gershgorin method as a means for guaranteeing the convexity of the resulting underestimator via determining lower bounds on eigenvalues of H' and thus proving positive semidefiniteness. The first of two new methods utilizes an eigenvalue inclusion set developed by Brauer [9] similar to Gershgorin's Circle Theorem. By using an extension of Brauer's method towards interval matrices and minimizing the maximum underestimation error, the following convex NLP gives an alternative diagonal perturbation matrix:

$$\begin{aligned}
\min_{\alpha} \quad & \sum_i \alpha_i \left(x_i^U - x_i^L\right)^2 / 4 \\
\text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i \geq 0 \qquad \forall i \\
& \alpha_i \geq 0 \qquad \forall i \\
& \frac{R_i R_j}{\left(\underline{h}_{ii} + 2\alpha_i\right)\left(\underline{h}_{jj} + 2\alpha_j\right)} \leq 1 \qquad \forall i, j : j > i, R_i > 0, R_j > 0,
\end{aligned} \tag{21}$$

where $R_i = \sum_{j\neq i} \left|h'_{ij}\right|$. The resulting parameters α_i are used with the quadratic αBB underestimator form in Eq. (2).

2.6 Method 6: Nondiagonal Perturbation Elements II (RE-αBB)

The second method applied by Skjäl and Westerlund [24] for bounding the eigenvalues of a matrix was specifically developed for interval matrices by Rohn [23]. The application of Rohn's method yields another symmetric nondiagonal perturbation matrix H^P with the same form as in Sect. 2.4. By minimizing the separation error of the resulting underestimator and modeling the constraints given by Rohn's method as a semidefinite programming constraint, elements α_i and β_{ij} of the underestimator form given by (14) are obtained by solving the following semidefinite program (SDP):

$$\begin{aligned}
\min_{\alpha,\beta,b} \quad & \sum_i \alpha_i \left(x_i^U - x_i^L\right)^2 / c_1 + \sum_i \sum_{j>i} b_{ij} \left(x_i^U - x_i^L\right)\left(x_j^U - x_j^L\right) / c_2 \\
\text{s.t.} \quad & \mathrm{mid}(H_0') + E + H^P \succeq \rho\left(\mathrm{rad}(H_0') + |E|\right) \\
& \alpha_i \geq 0 && \forall i \\
& \beta_{ij} = \beta_{ji} && \forall i, j : j > i \\
& b_{ij} \geq \beta_{ij} && \forall i, j : j > i \\
& b_{ij} \geq -\beta_{ij} && \forall i, j : j > i,
\end{aligned} \tag{22}$$

where $\rho(A)$ denotes the spectral radius of matrix A and terms $h'_{0,ij}$ of H_0' are defined as

$$h'_{0,ij} = \begin{cases} h'_{ij} & \text{if } i \neq j \\ \left[\underline{h}_{ij}, \underline{h}_{ij}\right] & \text{if } i = j. \end{cases}$$

There is a degree of freedom in choosing c_1 and c_2. By choosing $c = (6, 12)$, the objective minimizes the average separation error. Here we use $c = (4, 4)$, which minimizes the maximum separation error as per the original presentation of the underestimator for ND-αBB (which itself could be similarly modified to minimize the average separation error). There is also a degree of freedom in choosing an appropriate matrix E for (22), with options $E = 0$ and $E = \mathrm{diag}(\mathrm{rad}(H'))$ provided by Adjiman et al. [3]. Both options are explored here, denoted as RE-αBB[0] and RE-αBB[1], respectively.

2.7 Method 7: Using Putinar's Positivstellansatz (f_{dk})

The only method to diverge from utilization of the features of H' as a method to guarantee convexity of the underestimator is that of Lasserre and Thanh [15]. The moment approach is only applicable to polynomial functions, and searches for a polynomial function $f_d(\mathbf{x})$, i.e., constrained to degree d, that is both guaranteed convex and meets the criterion $f_d(\mathbf{x}) \leq f(\mathbf{x})$. It should be emphasized that $f_d(\mathbf{x})$ fully represents the underestimator and is not subtracted from $f(\mathbf{x})$ as was seen in

the other methods. This method is also the only one that does not constrain the underestimator to match the endpoints of $f(\mathbf{x})$. Key to the method is utilization of Putinar's Positivstellansatz [22] to guarantee convexity of $f_d(\mathbf{x})$ and underestimation of $f(\mathbf{x})$. For Archimedian quadratic modules

$$Q_{\mathbf{B}} = \left\{ \sum_{j=0}^{n} \sigma_j(\mathbf{x}) g_j(\mathbf{x}) : \sigma_j(\mathbf{x}) \in \Sigma[\mathbf{x}] \quad j = 1,2,\ldots,n \right\}, \tag{23}$$

where $\mathbf{B} = [0,1]^n$, $g_j(\mathbf{x}) = \left(x_j^U - x_j\right)\left(x_j - x_j^L\right), j = 1,2,\ldots,n$, $g_0 = 1$, and $\Sigma[\mathbf{x}]$ represents the cone of sum of squares, Putinar states that every strictly positive polynomial on $\mathbf{B}$ belongs to $Q_{\mathbf{B}}$. The constraint for convexity is given as

$$\mathbf{y}^{\mathsf{T}} \nabla^2 f_d(\mathbf{x}) \mathbf{y} \geq 0 \qquad \forall \mathbf{x} \in [\mathbf{x}^L, \mathbf{x}^U], \mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\| \leq 1. \tag{24}$$

The underlying mechanism behind Method 7 is to subtract from $f(\mathbf{x})$ a positive polynomial over $\mathbf{B}$ (guaranteeing underestimation) while constraining the result to be convex. The subtracted polynomial $\sum_{j=0}^{n} \sigma_j(\mathbf{x}) g_j(\mathbf{x})$ is constructed from positive quadratics $g_j(\mathbf{x})$ multiplied by polynomials selected from $\Sigma[\mathbf{x}]$, thus guaranteeing its positivity. A hierarchy of SDPs can be constructed, parametrized by integer $k \geq \max\{\lceil d/2 \rceil, \lceil (\deg f)/2 \rceil\}$, to construct the underestimator $f_{dk}(\mathbf{x})$, i.e., from program k and constrained to degree d. Higher values of k represent a higher complexity in guaranteeing positivity, and as $k \to \infty$ the tightest possible $f_d(\mathbf{x})$ is obtained. Alongside the general formulation, Lasserre and Thanh [15] present a considerably simplified SDP formulation for obtaining the underestimator by restricting $f_{dk}(\mathbf{x})$ to be quadratic ($d = 2$):

$$\begin{aligned} \max_{b,\mathbf{a},\mathbf{A}} \quad & b + \mathbf{a}^{\mathsf{T}}\gamma + \langle \mathbf{A}, \Lambda \rangle \\ \text{s.t.} \quad & f(\mathbf{x}) = b + \mathbf{a}^{\mathsf{T}}\mathbf{x} + \mathbf{x}^{\mathsf{T}}\mathbf{A} + \sum_{j=0}^{n} \sigma_j(\mathbf{x}) g_j(\mathbf{x}) \qquad \forall \mathbf{x} \\ & \mathbf{A} \succeq 0 \\ & \sigma_0(\mathbf{x}) \in \Sigma[\mathbf{x}]_k \\ & \sigma_j(\mathbf{x}) \in \Sigma[\mathbf{x}]_{k-1} \qquad\qquad j = 1,2,\ldots,n, \end{aligned} \tag{25}$$

where $\Sigma[\mathbf{x}]_k$ denotes the cone of sum of squares of degree at most $2k$. Parameters γ_i and Λ_{ij} are members of the moment matrix of order 1 of the *normalized* Lebesgue measure λ on $\mathbf{B}$,

$$\mathbf{M}_\lambda = \begin{bmatrix} 1 & \gamma^{\mathsf{T}} \\ \gamma & \Lambda \end{bmatrix},$$

and evaluate to

$$\gamma_i = \int_{\mathbf{B}} x_i \, d\lambda \qquad \forall i \tag{26}$$

$$\Lambda_{ij} = \int_{\mathbf{B}} x_i x_j \, d\lambda \qquad \forall i,j = 1,2,\ldots,n. \tag{27}$$

The parameter k is chosen here to be its minimum value, i.e.,

$$\max\{\lceil d/2\rceil, \lceil(\deg f)/2\rceil\},$$

as the complexity of the SDP problem explodes with large k. Solving the program yields $f_{dk}(\mathbf{x}) = b + \mathbf{a}^{\mathsf{T}}\mathbf{x} + \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$.

Problem (25) is still difficult to implement as written due to symbolic representation of the sum of squares search space. Thus, the first constraint is rewritten on the monomial basis:

$$f_\alpha = h_\alpha + \sum_{j=0}^{n} \langle \mathbf{Z}^j, \mathbf{C}_\alpha^j \rangle \qquad \forall \alpha \in \mathbb{N}^n_{2k}, \tag{28}$$

where $\alpha \in \mathbb{N}^n_{2k}$ represents the monomial basis of the ring of polynomials $\mathbb{R}[\mathbf{x}]_{2k}$ such that $\sum_i \alpha_i \leq 2k$, that is, the degree of each monomial is at most $2k$. With the selection of k as stated above, k is no smaller than $\deg f$, and thus each f_α represents the coefficient of $f(\mathbf{x})$ of the corresponding monomial $\mathbf{x}^\alpha : \alpha \in \mathbb{N}^n_{\deg f}$. Each variable h_α is the corresponding variable coefficient of $\mathbf{x}^\alpha$ from $f_{dk}(\mathbf{x})$; for example, with variable coefficients b, $\mathbf{a}$, and $\mathbf{A}$ of problem (25) and $n = 2$, vector $(h_\alpha) = \left[b, \mathbf{a}^{\mathsf{T}}, \mathrm{A}_{11}, \mathrm{A}_{12} + \mathrm{A}_{21}, \mathrm{A}_{22}\right]^{\mathsf{T}}$. Each variable matrix $\mathbf{Z}^j$ is constrained to be symmetric and positive semidefinite:

$$\mathbf{Z}^j \succeq 0 \qquad j = 0, 1, \ldots, n, \tag{29}$$

while each coefficient matrix $\mathbf{C}_\alpha^j$ is defined as

$$g_0(\mathbf{x})\mathbf{v}_k(\mathbf{x})\mathbf{v}_k(\mathbf{x})^{\mathsf{T}} = \sum_{\alpha \in \mathbb{N}^n_{2k}} \mathbf{C}_\alpha^0 \mathbf{x}^\alpha \tag{30}$$

$$g_j(\mathbf{x})\mathbf{v}_{k-1}\mathbf{v}_{k-1}(\mathbf{x})^{\mathsf{T}} = \sum_{\alpha \in \mathbb{N}^n_{2k}} \mathbf{C}_\alpha^0 \mathbf{x}^\alpha \qquad j = 1, 2, \ldots, n \tag{31}$$

where the vector $\mathbf{v}_k = \left(\mathbf{x}^\alpha : \alpha \in \mathbb{N}^n_k\right)$. Solving the equalities for each $\mathbf{C}_\alpha^j$ yields $n \cdot \binom{n+2k}{n}$ many coefficient matrices of size $\binom{n+k-1}{n} \times \binom{n+k-1}{n}$ for $\mathbf{C}_\alpha^j, j = 1, 2, \ldots, n$, and $\binom{n+2k}{n}$ many matrices of size $\binom{n+k}{n} \times \binom{n+k}{n}$ for $\mathbf{C}_\alpha^0$.

3 Implementation Details

3.1 General Implementation Details

We compared the methods under a general implementation that would be utilized in an automated workflow to construct the underestimator for any function. Automatic differentiation was employed to construct all Hessian matrices. Second-order automatic differentiation was performed via reverse accumulation and then forwards

accumulation of operators using C++ template headers provided by Bendtsen and Stauning [8] and overloaded for interval objects. Interval arithmetic was implemented using header files from the Boost C++ Interval Arithmetic Library (http://www.boost.org, version 1.53), although it should be noted that the final version of the interval object was heavily modified from the original implementation due to difficulties with transcendental functions and rational exponents. All convex NLPs that resulted from obtaining a convex underestimator were solved via external calls to local NLP solver CONOPT [10] through GAMS. Recorded times include all aspects of creating and minimizing the underestimator, except in the case of f_{dk} as noted below, and exclude the overhead of any external calls.

3.2 Method-Specific Implementation Details

To make the implementation of P-αBB general for any selected number of subdomains, a recursive algorithm was implemented for variable domain division. Both 2 and 4 intervals per each x_i are tested here, denoted as P-αBB[2] and P-αBB[4], respectively. The resulting piecewise function from P-αBB was minimized with CONOPT by directly interacting with the solver using a precompiled external module with a GAMS front-end.

To initiate the G-αBB method, the nonlinear system of equations in (11) was solved via an external call to MATLAB and by using $\overline{\gamma}_i$ as an initial guess to the solution. The convexification algorithm was prevented from proceeding when the bounds of a dimension of a subdomain in Λ_1 (see Fig. 1) spanned less than 5×10^{-6}. Updating parameter η was chosen to be 1.1, as suggested by Akrotirianakis and Floudas [4].

The LP from ND-αBB in (20) and the NLP from B-αBB in (21) were solved via external calls to GAMS with CPLEX (ILOG 2012, v12.4) and CONOPT, respectively. It should be noted that the allowable domain in problem (21) does not prevent the final constraint from having a zero denominator, a fact which can cause premature termination in NLP solvers such as CONOPT and required special attention and constraint rearrangements in a few cases.

The SDP in problem (22) for the RE-αBB method could not be solved through GAMS and was solved via MATLAB toolbox CVX [14, v1.22] calling the SeDuMi SDP solver [26, v1.1R3].

The moment approach (f_{dk}) was by far the most difficult to implement. Due to the enormous amount of matrices being generated in the formulation of the problem, for which some scheme that would take advantage of their sparsity could be developed, all times reported only include the SDP solve time. Formulated SDPs were solved via CVX calling SeDuMi.

4 Results and Discussion

The optimal values for every method applied to all 20 cases are shown in Table 1. It should be noted that the coefficient of the third term of function 5 in [13] should be corrected to 1/3 from 1/6 and was implemented as such here. There are some differences between the baseline αBB results here and those presented by Gounaris and Floudas [13], as they derived each Hessian matrix symbolically instead of through automatic differentiation. The tightness of αBB-based methods is sensitive to the bounds of the interval Hessian matrices; a few of the cases displayed improved αBB results using automatic differentiation to derive H', but cases 8a and 8b show markedly looser underestimators due to the inability of automatic differentiation to reduce the number of accumulated operations by taking advantage of cancellation of terms along the diagonal, which is exploited by symbolic differentiation.

The αBB method was by far the easiest method to implement, and it compares favorably to the other methods which employed much higher levels of sophistication in derivation and implementation. This indicates that, as suggested by Adjiman et al. [2, 3], the diagonal perturbation method driven by the scaled Gershgorin method is already a high performing method that is difficult to improve upon. It is also apparent that there is a high degree of similarity between many of the methods regarding their reliance on the tightness of H' and Gershgorin's Circle Theorem for proving convexity. The αBB method can be easily employed in a branch-and-bound algorithm with little computational cost.

Among the six quadratic subtraction methods (Methods 1–6), P-αBB showed by far the most consistent results with the test cases, and was in general the best performing method. It was the only method that took advantage of the diagonal structure of functions 7 and 14 to provide tighter underestimators than the other five methods (dramatically so in cases 14a–c). Utilization of 4 intervals per x_i often resulted in substantial gains in tightness, most dramatically with functions 12–14 which contain sharp differences in character across different subdomains. The P-αBB underestimators were rapidly generated here, but would not scale particularly well with very high dimensionality, as dividing n variables into N intervals would yield N^n subdomains. A more careful implementation of the method could subdivide only those dimensions with highly nonconvex characteristics.

The G-αBB algorithm in general produced similar results to αBB; in fact, the explicit algorithm (G-αBB[v1]) produced identical results in all cases except 5, 10, and 11. The suggested alteration to G-αBB, denoted as G-αBB[v2], either obtained identical results or outperformed G-αBB[v1] in all cases except 11, where it obtained a slightly looser underestimator. In particular, G-αBB[v2] was the only algorithm that was able to improve upon the very loose lower bounds found by the other methods with cases 8a and 8b, suggesting it took advantage of differences in nonconvexities between the dimensions. The α_i parameters obtained by G-αBB in cases 8a and 8b showed large differences in the order of magnitude between α_1 and $\alpha_i, i \neq 1$, reflecting the dimensionality of the nonconvexities of those problems. As expected, G-αBB[v2] was able to exploit this in cases 8a and 8b by further reducing α_1 while increasing all $\alpha_i, i \neq 1$, and producing an underestimator with a

Table 1 Minima of the convex underestimators generated by each method

$f(\mathbf{x})$	$[x_i^L, x_i^U]$	n	GO	αBB	P-αBB[2]	P-αBB[4]	G-αBB[v1]	G-αBB[v2]	ND-αBB	B-αBB	RE-αBB[0]	RE-αBB[1]	f_{dk}
1	[−1,1]	3	−2	−4	−4	−4	−4	−4	−4	−4	−4	−4	**−2**
2	[0,1]	4	−1	−1.54	−1.54	−1.54	−1.54	−1.54	−1.39	−1.54	−1.53	−1.53	**−1.22**
3a	[−1,1]	5	−6	−15	−15	−15	−15	−15	−15	−15	−15.7	−15.7	**−6**
3b	[1,3]	5	−66	−73.5	−73.5	−73.5	−73.5	−73.5	−69.3	−73.5	−72.0	−72.0	**−66**
4	[−1,1]	6	−3	−70.1	−70.1	−70.1	−70.1	−70.1	−70.0	−70.1	−74.1	−74.1	**−3.1**
5	[−2,2]	3	−1.26	−411.2	−411.2	−145.9	−311.6	−311.6	−411.2	−411.2	−411.2	−864.0	**−12.1**
6a	[−1,1]	3	−0.3	**−34.3**	**−34.3**	**−34.3**	**−34.3**	**−34.3**	**−34.3**	**−34.3**	**−34.3**	**−34.3**	–
6b	[−1,1]	4	−0.4	**−45.7**	**−45.7**	**−45.7**	**−45.7**	**−45.7**	**−45.7**	**−45.7**	**−45.7**	**−45.7**	–
7a	[1,3]	3	0.396	−108.3	−107.1	**−106.7**	−108.3	−108.3	−108.3	−108.3	−108.3	−108.3	–
7b	[1,3]	4	0.396	−108.3	−107.1	**−106.7**	−108.2	−108.2	−108.3	−108.3	−108.3	−108.3	–
8a	[−10,10]	3	0	−5.3 E+6	−5.3 E+6	−5.3 E+6	−5.3 E+6	**−4.8 E+6**	−5.3 E+6	−5.3 E+6	−5.3 E+6	−7.7 E+6	–
8b	[−10,10]	4	0	−6.0 E+6	−6.0 E+6	−5.9 E+6	−6.0 E+6	**−5.1 E+6**	−6.0 E+6	−6.0 E+6	−6.0 E+6	−7.8 E+6	–
9	[−10,10]	3	−4 E+2	**−2.0 E+6**	**−2.0 E+6**	**−2.0 E+6**	**−2.0 E+6**	**−2.0 E+6**	**−2.0 E+6**	**−2.0 E+6**	−2.1 E+6	−2.3 E+6	–
10	[0,1]	4	0	−202.4	−142.5	−112.1	−177.0	−135.5	−202.4	−202.4	−205.1	−338.4	**−1.0**
11	[0,1]	4	0	−43.2	−37.8	−21.3	−32.4	−33.6	−43.2	−43.2	−36.9	−32.9	**−0.6**
12	[1,10]	4	−38	−1.4 E+6	−8.7 E+5	−4.0 E+5	−1.4 E+6	−1.4 E+6	−1.4 E+6	−1.4 E+6	−1.5 E+6	−2.3 E+6	–
13	[0.1,10]	4	−9.2	−779	−452	**−186**	−779	−629	−779	−779	−987	−17,993	–
14a	[−5,2]	3	−300	−2,409	−1,340	**−647**	−2,409	−2,409	−2,409	−2,409	−2,409	−2,409	**−300**
14b	[−5,2]	4	−400	−3,212	−1,787	−863	−3,212	−3,212	−3,212	−3,212	−3,212	−3,212	**−400**
14c	[−5,2]	5	−500	−4,015	−2,234	−1,079	−4,015	−4,015	−4,015	−4,015	−4,015	−4,015	**−500**

Function numbers in the first column refer to Table 2 of [13], while the third and fourth columns denote the number of variables and global minimum, respectively, of each case. The tightest underestimator for each problem is in boldface

lower maximum separation distance and higher minimum. It should be noted that the G-αBB method often displayed long execution times (see Table 2) due to requiring an algorithm for proving convexity; to execute either version at each node of a branch-and-bound algorithm for problems of large dimensionality would likely be impractical.

The ND-αBB method, which was the earliest method in this report that attempted to introduce nondiagonal terms, largely reproduced the results of αBB. Although the decision tree did include the option to utilize the classical αBB parameters under certain conditions, it is interesting to note that the ND-αBB method recreated very similar results while only using classical αBB for functions 6, 8, 9, and 13. The most common outcome of the decision tree was utilization of explicit Eqs. (18) and (19) for calculating diagonal and nondiagonal perturbation terms. The LP model (20) was utilized only in cases 10 and 11, and the optimal parameters were identical to the diagonal shift matrix of αBB with all $\beta_{ij} = 0$. As is observed and suggested in [25], the ND-αBB method likely results in more substantial improvements over αBB with greater disparity between variable bounds.

The B-αBB and RE-αBB methods use alternatives to the scaled Gershgorin theorem as a convexity constraint with identical to mixed results. The results of the B-αBB underestimators, where parameters were obtained by solving an NLP, were identical to that of αBB. Skjäl and Westerlund [24] state that the union of Brauer ovals is a subset of the union of Gershgorin disks; in these test cases, the Brauer subsets seemed not to be a proper subset of the Gershgorin unions. RE-αBB[0] produced a slightly tighter underestimator than αBB in cases 2, 3b, and 11. However, RE-αBB[0] otherwise performed identically to αBB or worse as in cases 3a, 4, and 9–13. Other than in case 11, where RE-αBB[1] outperformed RE-αBB[0], the RE-αBB[1] method either matched the results of RE-αBB[0] or produced looser underestimators, sometimes dramatically so. The underestimators of RE-αBB[1] in cases 5, 10, 12, and 13 would have resulted in an extremely loose underestimator and would have delayed convergence by a large degree; it can be concluded that the performance of RE-αBB[1] did not justify the computational expense of solving an SDP for obtaining the underestimator.

It is difficult to judge the utility of the f_{dk} method, which was applied to all polynomial cases. As seen in Table 1, all of the f_{dk} underestimators produced were the tightest of all seven methods. However, both the formulation of the SDP and determination of its solution were computationally expensive. Lasserre and Thanh [15] report that the SDP of a typical example of $\deg f = 4$ and $n = 5$ took less than a second to be solved; for problems of similar size, this was observed here as well. It would be nontrivial for an algorithm traversing the branch-and-bound tree of a model of appreciable size to repeatedly solve the SDP. There also remain concerns about implementation of the SDP formulation stage; the number and size of matrices that needs to be generated is staggering and would be impractical for larger problems, regardless of whether or not sparse objects are used to represent the various two-dimensional parameter matrices. A problem of degree 6 with 50 variables, still a small optimization problem, would require the generation of 1.6 billion matrices of size $1{,}326 \times 1{,}326$ and 32.5 million matrices of size $23{,}426 \times 23{,}426$. Lasserre

Table 2 Execution times of the different convexification methods

$f(\mathbf{x})$	$[x_i^L, x_i^U]$	n	αBB	P-αBB[2]	P-αBB[4]	G-αBB[v1]	G-αBB[v2]	ND-αBB	B-αBB	RE-αBB[0]	RE-αBB[1]	f_{dk}
1	[−1,1]	3	0.002	0.002	0.003	0.349	0.352	0.002	0.064	0.386	0.173	0.271
2	[0,1]	4	0.002	0.003	0.015	0.360	0.376	0.002	0.069	0.440	0.191	0.284
3a	[−1,1]	5	0.002	0.005	0.093	0.372	0.373	0.002	0.066	0.412	0.201	0.345
3b	[1,3]	5	0.002	0.007	0.098	0.357	0.357	0.003	0.065	0.433	0.197	0.357
4	[−1,1]	6	0.002	0.030	1.502	0.379	0.371	0.004	0.067	0.412	0.202	6.942
5	[−2,2]	3	0.002	0.002	0.009	0.369	0.356	0.003	0.066	0.438	0.191	0.322
6a	[−1,1]	3	0.001	0.002	0.006	0.348	0.353	0.002	0.067	0.352	0.166	–
6b	[−1,1]	4	0.001	0.002	0.021	0.360	0.352	0.001	0.064	0.355	0.168	–
7a	[1,3]	3	0.002	0.002	0.008	0.352	0.350	0.002	0.067	0.386	0.177	–
7b	[1,3]	4	0.002	0.003	0.031	0.363	0.361	0.003	0.073	0.391	0.178	–
8a	[−10,10]	3	0.002	0.003	0.008	0.358	0.382	0.002	0.071	0.360	0.166	–
8b	[−10,10]	4	0.002	0.005	0.035	0.367	0.631	0.002	0.070	0.359	0.168	–
9	[−10,10]	3	0.002	0.005	0.021	0.364	0.364	0.003	0.068	0.361	0.168	–
10	[0,1]	4	0.002	0.005	0.024	0.367	0.361	0.071	0.069	0.410	0.186	0.318
11	[0,1]	4	0.002	0.016	0.016	0.379	0.377	0.067	0.068	0.431	0.217	0.311
12	[1,10]	4	0.002	0.003	0.022	0.410	0.507	0.003	0.076	0.452	0.204	–
13	[0.1,10]	4	0.003	0.003	0.018	0.411	0.603	0.003	0.068	0.366	0.220	–
14a	[−5,2]	3	0.001	0.003	0.006	0.384	0.442	0.001	0.068	0.355	0.168	0.289
14b	[−5,2]	4	0.002	0.004	0.023	0.372	0.597	0.002	0.066	0.353	0.167	0.334
14c	[−5,2]	5	0.002	0.006	0.107	0.383	0.828	0.002	0.065	0.355	0.169	0.396

The methods were implemented and executed with no parallelization on a Linux workstation containing one Intel Core 2 Quad processor with four 2.83 GHz cores. The times for all methods include the minimization of the resulting convex NLPs through CONOPT but do not include the overhead of the external call. It should be noted that not all of these times are directly comparable due to the differences in implementation as noted in Sect. 3

and Thanh [15] recommended applying the method to each individual nonconvex term, though a loss of tightness would result and a SDP would need to be solved (with nonnegligible computational cost) for each polynomial nonconvex term at each node. This method also did not constrain the endpoints of the underestimator to match the original function; it is unclear what consequences this would have on convergence of a branch-and-bound algorithm, especially as subdomains become increasingly tight. It remains to be seen if the method can be efficiently applied to problems of any appreciable dimensionality and degree.

5 Conclusions

The determination of tight, convex lower bounds in a branch-and-bound algorithm is crucial for the global optimization of models spanning numerous applications and fields. We explored the performance of a variety of competing methods across a diverse test suite of nonconvex functions. The moment approach (f_{dk}) generated very tight underestimators for polynomial functions at high computational cost, and the P-αBB method greatly improved upon the αBB underestimator in cases where sharp differences in subdomain characteristics could be exploited. The results also confirm the excellent performance of the classical nonuniform αBB formulation and the strength and low computational cost of the scaled Gershgorin method, as other methods of increasing sophistication and computational complexity often did not appreciably improve upon the results of αBB. Methods similar to αBB that moved away from the scaled Gershgorin method did not display superior performance and often produced inferior results. Furthermore, the P-αBB, G-αBB, RE-αBB, and f_{dk} methods face serious challenges in efficient implementation and application towards problems of high dimensionality, with f_{dk} also facing rising computational costs in problems of high polynomial degree. The most intractable constraint of the underestimator generation problem is the requirement of convexity; the relative utility of new convexification methods for general terms will likely hinge on the performance and computational cost of their treatment of this constraint with respect to the scaled Gershgorin method.

References

1. Adjiman, C.S., Floudas, C.A.: Rigorous convex underestimators for general twice-differentiable problems. J. Glob. Optim. **9**(1), 23–40 (1996)
2. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: A global optimization method, αBB, for general twice-differentiable constrained NLPs–II. Implementation and computational results. Comput. Chem. Eng. **22**(9), 1159–1179 (1998)
3. Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, αBB, for general twice-differentiable constrained NLPs—I. Theoretical advances. Comput. Chem. Eng. **22**(9), 1137–1158 (1998)

4. Akrotirianakis, I.G., Floudas, C.A.: A new class of improved convex underestimators for twice continuously differentiable constrained NLPs. J. Glob. Optim. **30**(4), 367–390 (2004)
5. Akrotirianakis, I.G., Floudas, C.A.: Computational experience with a new class of convex underestimators: box-constrained NLP problems. J. Glob. Optim. **29**(3), 249–264 (2004)
6. Al-Khayyal, F.A., Falk, J.E.: Jointly constrained biconvex programming. Math. Oper. Res. **8**(2), 273–286 (1983)
7. Androulakis, I.P., Maranas, C.D., Floudas, C.A.: αBB: a global optimization method for general constrained nonconvex problems. J. Glob. Optim. **7**(4), 337–363 (1995)
8. Bendtsen, C., Stauning, O: Fadbad, a flexible C++ package for automatic differentiation. Department of Mathematical Modelling, Technical University of Denmark, Kongens Lyngby (1996)
9. Brauer, A.: Limits for the characteristic roots of a matrix. II. Duke Math. J. **14**(1), 21–26 (1947)
10. Drud, A.S.: CONOPT - a large-scale GRG code. ORSA J. Comput. **6**(2), 207–216 (1994)
11. Floudas, C.A.: Deterministic Global Optimization: Theory, Methods and Applications, vol. 37. Springer, New York (2000)
12. Gershgorin, S.A.: Über die abgrenzung der eigenwerte einer matrix. Izv. Akad. Nauk SSSR Ser. Fiz.-Mat. **6**, 749–754 (1931)
13. Gounaris, C.E., Floudas, C.A.: Tight convex underestimators for $\mathcal{C}^2$-continuous problems: II. Multivariate functions. J. Glob. Optim. **42**(1), 69–89 (2008)
14. Grant, M., Boyd, S.: CVX: MATLAB Software for Disciplined Convex Programming, Version 1.22. http://cvxr.com/cvx (September 2012)
15. Lasserre, J.B., Thanh, T.P.: Convex underestimators of polynomials. J. Glob. Optim. **56**(1), 1–25 (2013)
16. Maranas, C.D., Floudas, C.A.: Global minimum potential energy conformations of small molecules. J. Glob. Optim. **4**(2), 135–170 (1994)
17. Maranas, C.D., Floudas, C.A.: Finding all solutions of nonlinearly constrained systems of equations. J. Glob. Optim. **7**(2), 143–182 (1995)
18. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: part I - convex underestimating problems. Math. Program. **10**(1):147–175 (1976)
19. Meyer, C.A., Floudas, C.A.: Trilinear monomials with positive or negative domains: facets of the convex and concave envelopes. Nonconvex Optim. Appl. **74**, 327–352 (2003)
20. Meyer, C.A., Floudas, C.A.: Trilinear monomials with mixed sign domains: facets of the convex and concave envelopes. J. Glob. Optim. **29**(2), 125–155 (2004)
21. Meyer, C.A., Floudas, C.A.: Convex underestimation of twice continuously differentiable functions by piecewise quadratic perturbation: spline αBB underestimators. J. Glob. Optim. **32**(2), 221–258 (2005)
22. Putinar, M.: Positive polynomials on compact semi-algebraic sets. Indiana Univ. Math. J. **42**(3), 969–984 (1993)
23. Rohn, J.: Bounds on eigenvalues of interval matrices. Zeitschrift fAr Angewandte Mathematik und Mechanik **78**(S3), 1049–1050 (1998)
24. Skjäl, A., Westerlund, T.: New methods for calculating αBB-type underestimators. J. Glob. Optim. **58**(3), 411–427 (2014)
25. Skjäl, A., Westerlund, T., Misener, R., Floudas, C.A.: A generalization of the classical αBB convex underestimation via diagonal and nondiagonal quadratic terms. J. Optim. Theory Appl. **154**(2), 462–490 (2012)
26. Sturm, J.F.: Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. Optim. Methods Softw. **11**(1–4), 625–653 (1999)
27. Tawarmalani, M., Sahinidis, N.V.: Semidefinite relaxations of fractional programs via novel convexification techniques. J. Glob. Optim. **20**(2), 133–154 (2001)
28. Tawarmalani, M., Sahinidis, N.V.: Convex extensions and envelopes of lower semi-continuous functions. Math. Program. **93**(2), 247–263 (2002)

A Quasi Exact Solution Approach for Scheduling Enhanced Coal Bed Methane Production Through CO_2 Injection

Yuping Huang, Anees Rahil, and Qipeng P. Zheng

1 Introduction

For unminable coals, enhanced coal bed methane (ECBM) production via CO_2 injection (CO_2-ECBM) is a promising way to further extend the economic value. With the advancement of this technology and wide presence of unminable coals, CO_2-ECBM technology has been implemented in various coal mines and locations (e.g., [11]). In addition to the profits made from the extracted methane (the major component of natural gas), this also allows the natural gas production company to get CO_2 credits by storing CO_2 in the coal bed seam. With increasing demand of natural gas (due to the growing presence of natural gas fired electricity generators) and environmental concerns over excessive CO_2 emission, CO_2-ECBM is becoming more profitable and applicable.

CO_2 level in our atmosphere has been in a nonstop increasing trend since the industrial revolution and reached a record level. Our human society has taken many actions to combat this trend. One of the major efforts is to deal with the ever-growing CO_2 emissions from the electrical power sector. The actions to this respect include introduction of renewable resources, more environmentally friendly generation technologies (e.g., combined cycle gas turbine CCCT, close cycle gas turbine), and CO_2 storage and sequestration, etc. Due to these new features and requirements, many research on scheduling in power systems and CO_2 storage and sequestration have been conducted in recent years. These include generation scheduling with wind power (e.g., [14]), unit commitment with both traditional and quick-start

Y. Huang • Q.P. Zheng (✉)
Department of Industrial Engineering and Management Systems,
University of Central Florida, Orlando, FL, USA
e-mail: yuping.huang@knights.ucf.edu; Qipeng.Zheng@ucf.edu

A. Rahil
Department of Industrial and Management Systems Engineering, West Virginia University,
Morgantown, WV, USA
e-mail: arahil@mix.wvu.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_12,

generators (e.g., [18]), and investment, development, and strategies on carbon storage and sequestration (e.g., [6, 17]). In addition, as an important resource of energy, natural gas has been receiving a lot of research attention, including its production, transportation, trading, and usage (e.g., [15, 16]). To our best knowledge, this is the first paper, which is sitting in the crossing point between research efforts on carbon storage and sequestration, and natural gas production and trading.

Compared to CO_2-ECBM, enhanced oil recovery with CO_2 (CO_2-EOR) is more mature technology and has been demonstrated to increase medium and light oil production effectively [1, 2, 8]. Although the mechanisms of CO_2 on oil recovery enhancement and methane recovery enhancement are different, the operations of CO_2-ECBM are analogous to CO_2-EOR but rely less on pure CO_2 [5]. The economic evaluations of CO_2-EOR indicate that the project feasibility depends largely on oil prices and carbon credit prices, where oil price is the main driver for CO_2 storage investment [3, 9]. Learning from the experience of CO_2-EOR, natural gas (NG) prices and carbon credit prices are two major factors to affect the CO_2-ECBM implementation. However, since the NG prices are more volatile in historical prices, which makes the project a more risky investment, the only revenue from NG sales hardly offsets the routine O&M costs and CO_2 storage costs. Therefore, benefits of environmental policies, e.g. CO_2 credits and allowances, are also considered to promote CO_2-ECBM projects smoothly. Additionally, coal bed methane recovery is subject to the physical environment and reactions and further impacts the gas production. We thus have a strong motivation to explore the CO_2-ECBM profit-maximization scheduling for project's economic analysis and long-term operation management.

The model proposed in [7] is a nonlinear multi-stage optimization problem. It is a computationally demanding nonlinear program due to the large number of variables and constraints. In this paper, we are proposing a quasi exact solution approach, where the nonlinear terms are discretized and linearized as in [12]. This is in contrast to using other global optimization methods (e.g., [4]). Using this quasi exact approach, the original nonlinear program is transformed to a mixed integer linear programming (MILP) problem. Due to the way the fractional number (small number, and usually less than 1) is represented in computer, the MILP problem is equivalent to the original problem when we have enough binary variables to represent the fractional number. Because there are many easy-to-use and advanced MILP solvers (e.g., CPLEX, GUROBI, XPRESS, etc.), the new problem can be solved more conveniently.

In the following, we will first present the model and its descriptions in Sect. 2; then we will discuss the quasi exact method used to solve the nonlinear program and show the equivalent MILP formulation after discretization and linearization in Sect. 3; based on the proposed method, we will show our numerical experiments in Sect. 4; we will conclude the paper and discuss future research in Sect. 5.

2 Problem Description

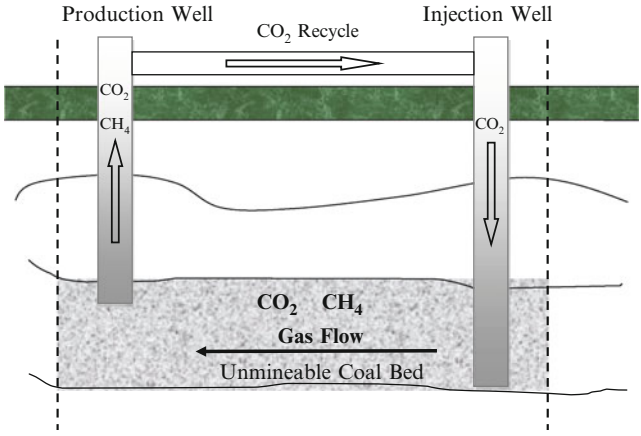


Fig. 1 Schematic diagram of single well CO_2 injection as shown in [7]

The main purpose of this paper is to provide a convenient and efficient method to solve the nonlinear programming models proposed in the paper [7]. The model is from the original paper, and hence we will not focus on the discussion of model itself. But the following will give a brief introduction of the model and discuss the difficulty to solve it.

Enhance Coal Bed Methane production through CO_2 injection is an effective technology to get additional value from the unminable coal mines. The whole process starts with injecting CO_2 into the unminable coal seams through the injection well. The absorption rate of CO_2 in coal seam is usually about twice the rate of CH_4, mainly depending on the type of the coal bed. With the new absorption of CO_2, the coal seam will release CH_4 (the major component of natural gas) which is adsorbed to the surface of coal. On the other side, the mixture of CO_2 and CH_4 will be drawn from the production well. A separation process follows to separate the extracted mixture to CO_2 and CH_4. The CH_4 will be sent to generate more profits (either by selling to the spot market or power generation plants); and the CO_2 will be sent to the injection well again. The general picture of the whole process is shown in Fig. 1 which is from [7]. The main variables used to model the whole process is linked to the process as shown in Fig. 2 also from [7], where one period of process are presented. To facilitate the description of the model, indices, parameters, variables, and the deterministic model are presented in the next following two subsections.

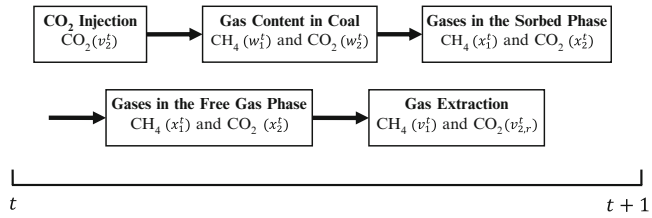


Fig. 2 Production operations for CO_2-ECBM recovery as shown in [7]

2.1 General Nomenclature

This paper is focusing on the solution method. We only list the indices, parameters and variables used in the model in the following. For detailed description and explanation of the variables, parameters, indices, and rationale of the model, please refer to the paper [7].

Indices

t Time period(days, months or years)
i Time period, $i = t$
1^* CH_4
2^* CO_2

* For convenience, variables with subscript 1 usually are referring to a value related to CH_4, and 2 for CO_2.

Variables

v_1^t The amount of CH_4 extracted at time t (MMcf /day)
v_2^t The amount of CO_2 injected at time t (MMcf/day)
$v_{2,r}^t$ The amount of CO_2 extracted at time t (MMcf/day)
x_1^t The molar fraction of component CH_4 in sorbed phase at time t
x_2^t The molar fraction of component CO_2 in sorbed phase at time t
y_1^t The molar fraction of component CH_4 in gas phase at time t
y_2^t The molar fraction of component CO_2 in gas phase at time t
w_1^t The gas content of component CH_4 on the coal at time t (Mcf/mton)
w_2^t The gas content of component CO_2 on the coal at time t (scf/ton)

Parameters

P_1^t Wellhead price for CH_4 sold at time t (US$/MMcf)
P_2^t Unit price for CO_2 credits trading at time t (US$/MMcf)
C_1^t Gas production cost at time t (US$/MMcf)
C_2^t CO_2 operation cost at time t (US$/MMcf)
$C_{2,r}^t$ CO_2 removal cost at time t (US$/MMcf)
NS_2^t The CO_2 supply amount at time t (MMcf/Month)
Q^t The actual flow rate at time t (MMcf/Month)
GIP The total amount of reserve for a CBM well (MMcf)
M_c The coal mass(mmton)
τ The percentage of CO_2 reinjection amount
θ The separation rate of methane
δ Time interval
γ The minimum methane molar fraction for allowable production

2.2 Optimization Model

In this paper, we are trying to use a quasi exact method to solve the most basic case of the ECBM scheduling problem, the general deterministic programming model. The model is shown as follows:

$$\textbf{[P]:}\ \max \sum_{t=0}^{T} P_1^t v_1^t + \sum_{t=0}^{T} P_2^t v_2^t - \sum_{t=0}^{T} C_1^t (v_1^t + v_{2,r}^t) - \sum_{t=0}^{T} C_2^t v_2^t - \sum_{t=0}^{T} C_{2,r}^t v_{2,r}^t \quad (1a)$$

$$\text{s.t.}\ \ v_2^t \le NS_2^t + \tau v_{2,r}^{t-1},\ \ t = 1,\ldots,T \quad (1b)$$

$$\sum_{t=0}^{T} Q^t \delta y_1^t \le \text{GIP}, \quad (1c)$$

$$x_1^t = \frac{w_1^{t-1}}{w_1^{t-1} + w_2^{t-1}},\ \ t = 1,\ldots,T \quad (1d)$$

$$x_2^t = 1 - x_1^t,\ \ t = 0,1,\ldots,T \quad (1e)$$

$$y_1^t = \frac{x_1^t}{\theta + (1-\theta)x_1^t},\ \ t = 0,1,\ldots,T \quad (1f)$$

$$y_2^t = 1 - y_1^t,\ \ t = 0,1,\ldots,T \quad (1g)$$

$$M_c w_1^t = M_c w_1^{t-1} - Q^t \delta y_1^t,\ \ t = 1,\ldots,T \quad (1h)$$

$$M_c w_2^t = M_c w_2^{t-1} + v_2^t - Q^t \delta y_2^t,\ \ t = 1,\ldots,T \quad (1i)$$

$$v_1^t \le \sum_{i=0}^{t} Q^i \delta y_1^i - \sum_{i=0}^{t-1} v_1^i,\ \ t = 1,\ldots,T \quad (1j)$$

$$v_{2,r}^t = \frac{v_1^t y_2^t}{y_1^t},\ \ t = 1,\ldots,T \quad (1k)$$

$$y_1^t \ge \gamma,\ \ t = 1,\ldots,T \quad (1l)$$

$$v_1^t,\ v_2^t,\ v_{2,r}^t,\ x_1^t,\ x_2^t,\ y_1^t,\ y_2^t,\ w_1^t,\ w_2^t \ge 0,\ t = 0,1,\ldots,T \quad (1m)$$

The objective is to maximize the total profit, which is composed of five parts: the profit from CH_4, the profit from CO_2, gas extraction cost, CO_2 injection cost, and CO_2 removal cost respectively shown in (1a). Constraint (1b) is the constraint on total CO_2 supply. Constraint (1c) is defining the CH_4 supply limit. Constraints (1d) and (1e) are used to model the composition of sorbed-phase gases. Constraints (1f) and (1g) are used to model the composition of gas-phase CO_2 and CH_4. Constraints on the variations of gas contents are defined by (1h) and (1i). Constraints (1j) and (1k) model the CH_4 and CO_2 extraction limits and extracted gas composition. A lower bound of CH_4 gas molar fraction is presented in (1l). All variables are continuous and nonnegative as shown in (1m). The major difficulty of this optimization problem comes from three nonlinear constraints, (1d), (1f), and (1k). These three constraints can be easily reformed to have bilinear terms instead of fractional nonlinear function. Computational times are relatively long and sometimes it is even

hard to converge, as is reported in [7]. In this paper we will take advantage of the specific characteristics of this formulation and transform it to mixed integer linear programs, which generally have easy-to-use and advanced solvers.

3 The MILP Solution Method

By constraints (1d) and (1e), it is clear that x_1 and x_2 are fractional numbers which only can chose between 0 and 1 because w_1 and w_2 are nonnegative variables. Together with constraints (1f) and (1g), it can be shown that y_1 and y_2 are nonnegative fractional numbers less or equal to 1. If we divide x_1 on both the denominator and numerator of the right-hand side of constraint (1f), the numerator becomes 1 and denominator becomes $\frac{\theta}{x_1} + (1-\theta)$, which is always greater or equal to 1 because $\frac{\theta}{x_1} \geq \theta$. In modern computer systems, a fractional number is represented by a series of binary numbers (bits). Hence, we can utilize this fact to treat a fractional variable by a series of binary variables.

Note also that (1d), (1f), and (1k) are equivalent to constraints with bilinear terms by multiplying the denominators of their right sides. The corresponding resulting bilinear constraints are shown as follows:

$$x_1^t w_1^{t-1} + x_1^t w_2^{t-1} = w_1^{t-1}, \quad t = 1, \ldots, T \tag{2a}$$

$$\theta y_1^t + (1-\theta) y_1^t x_1^t = x_1^t, \quad t = 0, 1, \ldots, T \tag{2b}$$

$$v_{2,r}^t y_1^t = v_1^t y_2^t, \quad t = 1, \ldots, T \tag{2c}$$

As in the above equations, all bilinear terms involve the fractional variables. Also, we know that the fractional variable can be represented by binary variables. Hence if we replace them, the resulting bilinear terms will involve one binary variable and one continuous variable. It is well-known that such kind of bilinear terms can be linearized by introducing additional constraints and a big number. Hence, we can transform the original nonlinear optimization problem to a MILP problem. This method is also used in [12]. In the following subsection, we will explain in details how the problem is converted to an MILP.

3.1 Discretization-Linearization Procedure to Eliminate Nonlinear Terms

Three nonlinear terms appear in the current deterministic model of ECBM production as in constraints (1d), (1f), and (1k). Firstly, we can transform them to bilinear constraints as shown in (2a)–(2c). Secondly, replace the fractional variables by combination of binary variables. Thirdly, we linearize the bilinear term with exactly one binary variable and one continuous variable. Then we get a MILP optimization problem. We refer this procedure to Discretization-Linearization procedure as discussed

in [12]. The validity and accuracy to the original model is mainly controlled by the number of binary variables introduced to replace each fractional variable. For example, we would like to replace variable x by a series of binary variables (i.e., z_l, $l = 1, \ldots$), and the formulation is shown as follows:

$$x = \sum_{l=0}^{L} 2^{-l} z_l \tag{3}$$

Say, if we require a degree of accuracy $\varepsilon = 10^{-p}$ where $p \geq 0$ and $z_l \in \{0, 1\}$. The number of binary variables needed, L, for the binary representation is

$$L = \left\lceil p \frac{\log 10}{\log 2} \right\rceil \tag{4}$$

In the next three subsubsections, we will discuss how each of the three nonlinear constraints are linearized by the discretization-linearization procedure.

3.1.1 Linearization of Composition of Sorbed Phase Constraints

The first nonlinear term appears in the expression of composition of sorbed phase as follows:

$$x_1^t = \frac{w_1^{t-1}}{w_1^{t-1} + w_2^{t-1}}, \quad t = 1, \ldots, T.$$

which can be converted to an equivalent bilinear constraints as in (2a). First, we use binary representation to replace x_1^t as follows:

$$x_1^t = \sum_{l=0}^{L} 2^{-l} z_l^t, \quad t = 1, \ldots, T. \tag{5}$$

where $z_l \in \{0, 1\}$, $l = 0, \ldots, L$. After applying the new replacement in the bilinear constraint (2a), we have the following new bilinear constraints

$$\left[\sum_{l=0}^{L} 2^{-l} z_l^t w_1^{t-1} \right] + \left[\sum_{l=0}^{L} 2^{-l} z_l^t w_2^{t-1} \right] - w_1^{t-1} = 0, \, t = 1, \ldots, T. \tag{6}$$

The new bilinear terms include exactly one binary variable and one continuous non-negative variable. We introduce λ_l^t and φ_l^t to assist linearize the bilinear terms $z_l^t w_1^{t-1}$ and $z_l^t w_2^{t-1}$ above, respectively. Let

$$\lambda_l^t = w_1^{t-1} z_l^t, \quad t = 1, \ldots, T, \;\; l = 0, \ldots, L \tag{7a}$$

$$\varphi_l^t = w_2^{t-1} z_l^t, \quad t = 1, \ldots, T, \;\; l = 0, \ldots, L \tag{7b}$$

Then we need to introduce the following constraints to have equivalent transformation (the standard linearization technique as in [10])

$$0 \le \lambda_l^t \le w_1^{t-1}, \ \ t = 1, \ldots, T, \ l = 0, \ldots, L \tag{8a}$$

$$w_1^{t-1} - R_l(1 - z_l^t) \le \lambda_l^t \le R_l z_l^t, \ \ t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{8b}$$

$$0 \le \varphi_l^t \le w_2^{t-1}, \ \ t = 1, \ldots, T, \ l = 0, \ldots, L \tag{8c}$$

$$w_2^{t-1} - R_l(1 - z_l^t) \le \varphi_l \le R_l z_l^t, \ \ t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{8d}$$

where R_l is a large number to bound the variables. Instead, the bilinear constraint (6) is then replaced as follows:

$$w_1^{t-1} = \sum_{l=1}^{L} 2^{-l}(\lambda_l^t + \varphi_l^t), \ \ t = 1, \ldots, T, \tag{9}$$

Then the final linear formulation only includes constraints (5), (8a)–(8d), and (9).

3.1.2 Linearization of Composition of Gas Phase Constraints

By using the extended Langmuir isotherm, the composition of free gas y_i can be obtained. The fraction of CO_2/CH_4 in the gas phase is represented as follows,

The gas molar fraction of CH_4:

$$y_1^t = \frac{x_1^t}{\theta + (1 - \theta)x_1^t}$$

The gas molar fraction of CO_2:

$$y_2^t = 1 - y_1^t$$

The key is to make the first constraint linearized. Note that the bilinear terms for the reformulated first constraint (2b) involve also x_1^t. Since we have already discretized x_1^t while linearizing composition of sorbed phase constraints, we keep using the binary representation of (5). The new bilinear constraints are shown as follows:

$$x_1^t = \theta y_1^t + \sum_{l=0}^{L} (1 - \theta) 2^{-l} z_l^t y_l^t, \ t = 0, 1, \ldots, T \tag{10}$$

To linearize the above constraint, we introduce a new variable ζ_l^t to replace the bilinear term $z_l^t y_l^t$. Using the same technique as in the last subsubsection, we need to introduce the following new constraints:

$$0 \le \zeta_l^t \le y_1^t, \ \ t = 0, \ldots, T, \ \ l = 0, \ldots, L \tag{11a}$$

$$y_1^t - R_l(1 - z_l^t) \le \zeta_l^t \le R_l z_l^t, \ \ t = 0, \ldots, T, \ \ l = 0, \ldots, L \tag{11b}$$

where R_l is a large number to bound the variables. Instead, the bilinear constraint (10) is then replaced by the following equation:

$$x_1^t = \theta y_1^t + \sum_{l=0}^{L} (1-\theta) 2^{-l} \zeta_l^t,\ t = 0, 1, \ldots, T \tag{12}$$

Then the final linear formulation only includes constraints (5), (11a), (11b), and (12).

3.1.3 Linearization of CO_2 Extraction Constraints

The third nonlinear term appears as the expression for the amount of CO_2 extraction. According to Dalton's Law of partial pressure, we have the following constraints:

$$v_{2,r}^t = \frac{v_1^t y_2^t}{y_1^t},\ t = 1, \ldots, T$$

The equivalent bilinear constraints (2c) include two bilinear terms, which include two sets of completely different variables. However, we know $y_2^t = 1 - y_2^t$ and can then reduce the number of variables involved in bilinear terms. After make the substitution, we get the following:

$$v_1^t = y_1^t v_{2,r}^t + y_1^t v_1^t,\ \ t = 1, \ldots, T \tag{13}$$

We already know that y_1^t is a fractional number as well. In addition, it appears in both of the bilinear terms of (13). Hence it is convenient if we use binary representation to replace it as follows:

$$y_1^t = \sum_{l=0}^{L} 2^{-l} \eta_l^t,\ \ t = 1, \ldots, T \tag{14}$$

After introducing the above binary representation, we obtain the following new bilinear constraints:

$$v_1^t = \left[\sum_{l=0}^{L} 2^{-l} \eta_l^t v_{2,r}^t \right] + \left[\sum_{l=0}^{L} 2^{-l} \eta_l^t v_1^t \right],\ \ t = 1, \ldots, T \tag{15}$$

Then we introduce two new nonnegative continuous variables to replace the bilinear terms in (15). Let

$$\psi_{1l}^t = v_1^t \eta_l^t,\ \ t = 1, \ldots, T,\ \ l = 0, \ldots, L \tag{16a}$$

$$\psi_{2l}^t = v_{2,r}^t \eta_l^t,\ \ t = 1, \ldots, T,\ \ l = 0, \ldots, L \tag{16b}$$

Using the same standard linearization technique as in the previous two subsubsections, we need to introduce the following new constraints:

$$0 \le \psi_{1l}^t \le v_1^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{17a}$$
$$0 \le \psi_{2l}^t \le v_{2,r}^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{17b}$$
$$v_1^t - R_l(1-\eta_l^t) \le \psi_{1l}^t \le R_l\eta_l^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{17c}$$
$$v_{2,r}^t - R_l(1-\eta_l^t) \le \psi_{2l} \le R_l\eta_l^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{17d}$$

where R_l is a large number to bound the variables. Instead, the bilinear constraint (15) is then replaced by the following equation:

$$v_1^t = \sum_{l=0}^{L} 2^{-l}(\psi_{1l}^t + \psi_{2l}^t) \quad t = 1,\ldots,T \tag{18}$$

Then the final linear formulation only includes constraints (14), (17a)–(17d), and (18).

3.2 Linear Optimization Model

After the nonlinear/bilinear terms are discretized and therefore linearized, the new linear deterministic model is obtained. In order to clarify the mixed integer linear program, we show all constraints and variables of the model as follows:

[MILP]:

$$\max \quad \sum_{t=0}^{T} P_1^t v_1^t + \sum_{t=0}^{T} P_2^t v_2^t - \sum_{t=0}^{T} C_1^t(v_1^t + v_{2,r}^t) - \sum_{t=0}^{T} C_2^t v_2^t - \sum_{t=0}^{T} C_{2,r}^t v_{2,r}^t \tag{19a}$$
$$\text{s.t.} \quad v_2^t \le NS_2^t + \tau v_{2,r}^{t-1}, \quad t = 1,\ldots,T \tag{19b}$$
$$\sum_{t=0}^{T} Q^t \delta y_1^t \le \text{GIP}, \tag{19c}$$
$$x_1^t = \sum_{l=0}^{L} 2^{-l} z_l^t, \quad t = 1,\ldots,T. \tag{19d}$$
$$0 \le \lambda_l^t \le w_1^{t-1}, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{19e}$$
$$w_1^{t-1} - R_l(1-z_l^t) \le \lambda_l^t \le R_l z_l^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{19f}$$
$$0 \le \varphi_l^t \le w_2^{t-1}, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{19g}$$
$$w_2^{t-1} - R_l(1-z_l^t) \le \varphi_l \le R_l z_l^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{19h}$$
$$w_1^{t-1} = \sum_{l=1}^{L} 2^{-l}(\lambda_l^t + \varphi_l^t), \quad t = 1,\ldots,T, \tag{19i}$$
$$x_2^t = 1 - x_1^t, \quad t = 0,1,\ldots,T \tag{19j}$$
$$0 \le \zeta_l^t \le y_1^t, \quad t = 1,\ldots,T, \quad l = 0,\ldots,L \tag{19k}$$
$$y_1^t - R_l(1-z_l^t) \le \zeta_l^t \le R_l z_l^t, \quad t = 0,\ldots,T, \quad l = 0,\ldots,L \tag{19l}$$

$$x_1^t = \theta y_1^t + \sum_{l=0}^{L} (1-\theta) 2^{-l} \zeta_l^t, \quad t = 0, 1, \ldots, T \tag{19m}$$

$$y_2^t = 1 - y_1^t, \ \ t = 1, \ldots, T \tag{19n}$$

$$M_c w_1^t = M_c w_1^{t-1} - Q^t \delta y_1^t, \ \ t = 1, \ldots, T \tag{19o}$$

$$M_c w_2^t = M_c w_2^{t-1} + v_2^t - Q^t \delta y_2^t, \ \ t = 1, \ldots, T \tag{19p}$$

$$v_1^t \le \sum_{i=0}^{t} Q^i \delta y_1^i - \sum_{i=0}^{t-1} v_1^i, \ \ t = 1, \ldots, T \tag{19q}$$

$$y_1^t = \sum_{l=0}^{L} 2^{-l} \eta_l^t, \quad t = 1, \ldots, T \tag{19r}$$

$$0 \le \psi_{1l}^t \le v_1^t, \quad t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{19s}$$

$$0 \le \psi_{2l}^t \le v_{2,r}^t, \quad t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{19t}$$

$$v_1^t - R_l(1 - \eta_l^t) \le \psi_{1l}^t \le R_l \eta_l^t, \quad t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{19u}$$

$$v_{2,r}^t - R_l(1 - \eta_l^t) \le \psi_{2l} \le R_l \eta_l^t, \quad t = 1, \ldots, T, \ \ l = 0, \ldots, L \tag{19v}$$

$$v_1^t = \sum_{l=0}^{L} 2^{-l} (\psi_{1l}^t + \psi_{2l}^t), \quad t = 1, \ldots, T \tag{19w}$$

$$y_1^t \ge \gamma, \quad t = 0, \ldots, T \tag{19x}$$

$$v_1^t, v_2^t, v_{2,r}^t, x_1^t, x_2^t, y_1^t, y_2^t, w_1^t, w_2^t, \lambda_l^t, \varphi_l^t, \zeta_l^t, \psi_{1l}^t, \psi_{2l}^t \ge 0, \quad t = 0, 1, \ldots, T, \ l = 0, \ldots, L \tag{19y}$$

$$z_l^t, \eta_l^t \in \{0, 1\}, \quad t = 1, \ldots, T, \ l = 0, 1, \ldots, L \tag{19z}$$

where z_l^t, η_l^t are the two sets of binary variables. As in the previous discussion, the degree of accuracy of this program is largely dependent on the number of binary variables introduced to represent the two sets of continuous variables x_1^t and y_1^t. However, the more the binary variable introduced, the more the computational demanding becomes the new MILP problem. In the next section, we will report our computational experiments based on the MILP formulation.

4 Numerical Experiments

The linearized CO_2-ECBM model is programmed in C++ and solved by CPLEX 12.2. All experiments are implemented on a PC Dell Vostro with Intel Pentium CPU at 2.80 GHz and 3 GB memory. The experiment results are compared with the original results that were gained through solving the original model in GAMS with the commercial solver BARON.

The linearized model is applied to a 12-period case studies. This case selected is to illustrate the impacts of economical factors such as prices and operational costs on the production scheduling when the technical parameters are given and the maximum extraction rate for a single well is fixed, shown in Table 1. This case is based

Table 1 Technical parameters

Parameter	Value
GIP	1,200,000 MMcf
θ	1.2
γ	0.1
τ	0.55
M_c	4,000 million tons
Max. extr. rate	30 MMcf/month

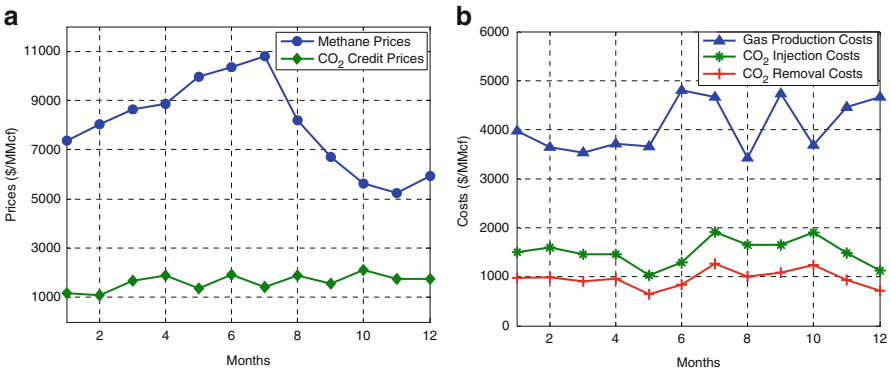


Fig. 3 The trends of prices and costs ($T = 12$). (**a**) Methane wellhead prices and CO_2 trading prices. (**b**) Gas production costs, CO_2 injection costs and CO_2 removal costs

on a 1-year horizon, where the methane wellhead prices are selected from US EIA, starting in January 2008 and ending in December 2008. CO_2 credit price for each period is generated within the range of historical prices [13]. Figure 3 shows two types of prices (Fig. 3a) and three types of costs (Fig. 3b), respectively. The total CO_2 supply amount in each period is forecasted and shown in Fig. 4. Due to the continuity of production planning, it is assumed 30 work days in a month.

Table 2 lists the computing times of four different period projects given to a series of numbers of required binary variables (L) for each discretization and a fixed large number (R_l). We specify all results under the situation of $R_l = 1{,}000$. For above four cases, they have the same starting period, but different ending periods which are employed to describe the relationship between a project period and the computation time. As the project period increased, under the condition with same L, the computation time also grows significantly. For example, the computation time for annual planning is at least ten times of time for half-year planning, yet the computation time is still within a feasible range. Moreover, we tested many instances with

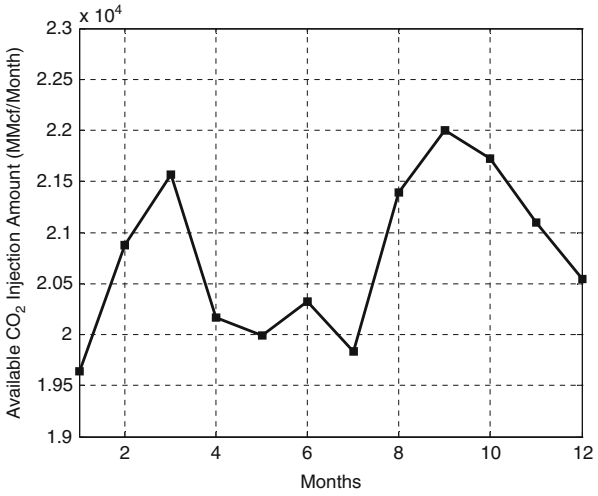


Fig. 4 CO_2 supply amount ($T = 12$)

more than 12 periods, whereas no feasible solution is obtained in less than 3 h, and then generally the solving process was terminated.

As can be seen in Table 2, there exists the minimum number of binary variables (L) required for each discretization based on the length of a project. In general, a larger number of binary variables (L) are required when the longer schedule is made. If the number of integer variable is reduced below a threshold, solving the corresponding case only yields an infeasible solution. We observed the four cases individually and found that the computation times fluctuate within the same order of magnitude in spite of number of binary variables increased. However, the computation time for 12-period case even has a larger variance, which means that the computation could be costly in some instances due to the number of binary variables needed.

In addition, for the solution deviation, we take an example of a half-year case with $L = 50$ to compare the solutions solved by BARON using the original model and by CPLEX using the linearized MILP model, respectively. Table 3 shows that the percentage of deviation ($\%\Delta$) on gas production rate v_1, CO_2 injection rate v_2, and CO_2 removal amount v_{2r}. Most percentage deviations are lower than 1 %, which indicates that the discretization-linearization technique not only helps solve the nonlinear CO_2-ECBM problems successfully but also is applicable to small-size and median-size cases.

Table 2 Number of binary variables and computation time for four cases

		L										
Case	Period	15	20	25	30	40	50	60	70	80	90	100
I	$T = 3$	inf.	65.99	115.83	3.93	1.26	2.88	8.63	10.03	8.16	5.71	11.86
II	$T = 6$	inf.	inf.	inf.	19.48	85.63	53.96	74.68	50.09	48.39	41.03	55.57
III	$T = 9$	inf.	inf.	inf.	137.5	239.95	339.97	702.07	179.96	1,071.99	110.4	703.77
IV	$T = 12$	inf.	inf.	inf.	inf.	610.71	1891.25	900.63	490.59	16,187.54	588.50	273.41

Table 3 Comparison of solutions from original model and linearized MILP model

	v_1			v_2			v_{2r}		
Period	Org.	Lnz.	$\%\Delta$	Org.	Lnz.	$\%\Delta$	Org.	Lnz.	$\%\Delta$
1	29.8805	29.8790	0.005	0	0	0.000	0.1195	0.1210	1.251
2	29.8805	29.8726	0.026	0	0	0.000	0.1195	0.1274	6.594
3	29.8805	29.8757	0.016	21,565.07	21,565.1	0.000	0.1195	0.1243	4.003
4	29.2521	29.2527	0.002	20,170.07	20,170.1	0.000	0.7479	0.7473	0.086
5	28.6878	28.6812	0.023	19,993.41	19,993.4	0.000	1.3122	1.3188	0.504
6	28.1495	28.1382	0.040	20,327.72	20,327.7	0.000	1.8505	1.8618	0.608

5 Conclusion

This paper discusses an ECBM scheduling problem through CO_2 injection. The original model is proposed in Huang et al. [7], which takes into account the profits from both natural gas sales and CO_2 credits, and chemical/physical reaction details of the process. It is a management problem including a great amount of real technical details in practice. However, the model is a nonlinear optimization problem and is computationally very challenging. The main contribution of this paper is the use of a quasi exact reformulation, which is a mixed integer linear program, to solve the original model. Both discretization and linearization techniques are used to construct the MILP reformulation. Accuracy of the reformulation is dependent on the number binary variables used to discretize the fractional variables in the original model. Computational experiments show that the results (obtained in reasonable computing times) from the reformulation are almost as same as the exact solutions. With the popularity and advancement of integer/MILP software packages (e.g., CPLEX, EXPRESS, GUROBI, etc.), the reformulation approach will be more accessible to general users and provide efficient and effective solutions. This paper uses the reformulation to solve the deterministic models from [7]. Future research in this direction would be solving the multi-stage stochastic models, which will be very difficult, because the number of introduced binary variables will grow exponentially. Advanced and specifically devised decomposition algorithms will be required to handle these cases. In addition, including transportation constraints (given multiple locations of resources) on both natural gas and CO_2 will be more interesting to higher level decision makers.

References

1. Alvarado, V., Manrique, E.: Enhanced oil recovery: an update review. Energies **3**(9), 1529–1575 (2010)
2. Firoozabadi, A., Cheng, P.: Prospects for subsurface CO_2 sequestration. Am. Inst. Chem. Eng. **56**(6), 1398–1405 (2010)
3. Fleten, S.-E., Lien, K., Ljønes, K., Pagés-Bernaus, A., Aaberg, M.: Value chains for carbon storage and enhanced oil recovery: optimal investment under uncertainty. Energy Syst. **37**, 457–470 (2010)
4. Floudas, C.A., Pardalos, P.M.: State of the Art in Global Optimization: Computational Methods and Applications. Kluwer, The Netherlands (1996)
5. Gunter, B.: Alberta Research Council Enhanced Coalbed Methane Recovery Project in Alberta, Canada. In: COAL-SEQ I, Houston, TX, 14–15 Mar 2002
6. Huang, Y. Rebennack, S., Zheng, Q.P.: Techno-economic analysis and optimization models for carbon capture and storage - a survey. Energy Syst. **4**(4), 315–353 (2013)
7. Huang, Y., Zheng, Q.P., Fan, N., Aminian, K.: Optimal scheduling for enhanced coal bed methane production through CO_2 injection. Appl. Energy **113**, 1475–1483 (2014)
8. NETL: 2010 Carbon Sequestration Atlas of the United States and Canada. Technical Report, U.S. Department of Energy (November 2010)
9. Phares, L.: Storing CO_2 with Enhanced Oil Recovery. Technical Report, U.S. Department of Energy (February 2008)
10. Prokopyev, O.A., Meneses, C., Oliveira, C.A.S., Pardalos, P.M.: On multiple-ratio hyperbolic 0-1 programming problems. Pac. J. Optim. **1**(2), 327–345 (2005)
11. Robertson, E.P.: Enhanced Coal Bed Methane Recovery and CO_2 Sequestration in the Powder River Basin. Technical Report INL/EXT-10-18941, Idaho National Laboratory, 08 (2010)
12. Temiz, N.A., Trapp, A., Prokopyev, O.A., Camacho, C.J.: Optimization of minimum set of protein.dna interactions: a quasi exact solution with minimum over-fitting. Bioinformatics **26**(3), 319–325 (2010)
13. U.S. EIA: Annual energy outlook 2012. DOE/EIA-0383(2012). http://www.eia.gov/forecasts/aeo (2012)
14. Wang, J., Shahidehpour, M., Li, Z.: Security-constrained unit commitment with volatile wind power generation. IEEE Trans. Power Syst. **23**(3), 1319–1327 (2008)
15. Zheng, Q.P., Pardalos, P.M.: Stochastic and risk management models and solution algorithm for natural gas transmission network expansion and lng terminal location planning. J. Optim. Theory Appl. **147**, 337–357 (2010)
16. Zheng, Q.P., Rebennack, S., Iliadis, N.A., Pardalos, P.M.: Optimization models in the natural gas industry. In: Handbook of Power Systems, pp. 121–148. Springer, Berlin (2010)
17. Zheng, Q.P., Rebennack, S., Pardalos, P.M., Pereira, M.V.F., Iliadis, N.A.: Handbook of CO_2 in Power Systems. Springer, New York (2012)
18. Zheng, Q.P., Wang, J., Pardalos, P.M., Guan, Y.: A decomposition approach to the two-stage stochastic unit commitment problem. Ann. Oper. Res. **210**(4), 387–410 (2013)

A Stochastic Model of Oligopolistic Market Equilibrium Problems

Baasansuren Jadamba and Fabio Raciti

1 Introduction

We provide a stochastic formulation of the classical deterministic oligopolistic market equilibrium, *à la Cournot* [2] in this short note. Equilibria of this kind are particular cases of Nash equilibria, and it is well known (see, e.g., [1] for the general Hilbert space case, and [4] for a finite-dimensional framework close to operations research problems) that under standard hypotheses solutions can be obtained by solving a variational inequality. Thus, we can apply the theory of random (or stochastic) variational inequalities in Lebesgue spaces to our model. This approach has been proposed quite recently to study many stochastic equilibrium problems arising from applied sciences and operations research [5–8, 10]. Other approaches to stochastic variational inequalities have been proposed by other authors. Here we cite only the very recent paper [13] which also contains applications to Nash equilibrium problems.

The paper is structured in four sections. In the remainder of this introduction we briefly recall the connection between Nash equilibrium problems and variational inequalities in the deterministic, finite-dimensional setting; in Sect. 2 we introduce random data in the deterministic oligopolistic market model; in Sect. 3 we present the Lebesgue-space formulation of the stochastic model; in Sect. 4 we study a particular class of utility functions, and use them to illustrate our model by means of a numerical example.

B. Jadamba
Center for Applied and Computational Mathematics, Rochester Institute of Technology,
85 Lomb Memorial Drive, Rochester, NY 14623, USA
e-mail: bxjsma@rit.edu

F. Raciti (✉)
Dipartimento di Matematica e Informatica, Università di Catania,
Viale A. Doria 6-I, 95125 Catania, Italy
e-mail: fraciti@dmi.unict.it

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_13,

Consider m players each acting in a selfish manner in order to maximize their individual welfare. Each player i has a strategy vector $q_i = (q_{i1}, \ldots, q_{in}) \in X_i$, where $X_i \subset \mathbb{R}^n$ is a convex and closed set, and a utility (or welfare) function $w_i : X_1 \times X_2 \times \cdots \times X_m \to \mathbb{R}$. He/she chooses his/her strategy vector q_i so as to maximize w_i, given the moves $(q_j)_{j \neq i}$ of the other players. We will use the notation $q_{-i} = (q_1, \ldots, q_{i-1}, q_{i+1}, \ldots, q_m)$ and $q = (q_i, q_{-i})$.

Definition 1. A Nash equilibrium is a vector $q^* = (q_1^*, \ldots, q_m^*) \in X$, such that:

$$w_i(q_i^*, q_{-i}^*) \geq w_i(q_i, q_{-i}^*),\ \forall q_i \in X_i, \forall i \in \{1, \ldots, m\}.$$

The following theorem (see e.g. [12, Chap. 6]) relates Nash equilibrium problems and variational inequalities.

Theorem 1. *Let $w_i \in C^1(X), \forall i$, and concave with respect to q_i. Let $F : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$ be the mapping built with the partial gradients of the utility functions as follows:*

$$F(q) = (-D_{q_1} w_1(q), \ldots, -D_{q_m} w_m(q)).$$

Then, $q^* \in X$ is a Nash equilibrium if and only if it satisfies the variational inequality:

$$\sum_{r=1}^{mn} F_r(q^*) \cdot (q_r - q_r^*) \geq 0, \forall q \in X$$

2 The Stochastic Oligopoly Model

We consider here the model in which m players are the producers of the same commodity. The quantity produced by firm i is denoted by q_i so that $q \in \mathbb{R}^m$ denotes the global production vector. Let (Ω, P) be a probability space and for every $i \in \{1, \ldots, m\}$ consider functions $f_i : \Omega \times \mathbb{R} \to \mathbb{R}$ and $p : \Omega \times \mathbb{R}^m \to \mathbb{R}$.

More precisely, for almost every $\omega \in \Omega$, (i.e. P-almost surely in probabilistic language), $f_i(\omega, q_i)$ represents the cost of producing the commodity by firm i, and is assumed to be nonnegative, increasing, concave, and C^1, while $p(\omega, q_1 + \cdots + q_m)$ represents the demand price associated with the commodity. For almost every $\omega \in \Omega$, p is assumed nonnegative, increasing, convex w.r.t. q_i, and C^1. We also assume that all these functions are random variables w.r.t. ω, i.e. they are measurable with respect to the probability measure P on Ω. In this way, we have introduced the possibility that both the production cost and the demand price are affected by a certain degree of uncertainty or randomness.

Thus, the welfare (or utility) function of player i is given by:

$$w_i(\omega, q_1, \ldots, q_m) = p(\omega, q_1 + \cdots + q_m) q_i - f_i(\omega, q_i). \tag{1}$$

Although many authors assume no bounds on the production, in a more realistic model the production capability is bounded from above and we allow also for the upper bound being a random variable: $0 \leq q_i \leq \overline{q}_i(\omega)$.

Thus, the specific Nash equilibrium problem associated with this model takes the following form. For a.e. $\omega \in \Omega$, find $q^*(\omega) = (q_1^*(\omega), \ldots, q_m^*(\omega))$:

$$w_i(q^*(\omega)) = \max_{0 \leq q_i \leq \overline{q}_i(\omega)} \left\{ p(\omega, q_i + \sum_{j \neq i} q_j^*(\omega)) q_i - f_i(\omega, q_i) \right\}, \forall i. \tag{2}$$

In order to write the equivalent variational inequality, consider the closed and convex subset of $\mathbb{R}^m$:

$$K(\omega) = \{(q_1, \ldots, q_m) : 0 \leq q_i \leq \overline{q}_i(\omega), \forall i\}$$

for each ω and define the functions

$$F_i(\omega, q) := \frac{\partial f_i(\omega, q_i)}{\partial q_i} - \frac{\partial p(\omega, \sum_{j=1}^m q_j)}{\partial q_i} q_i - p\left(\omega, \sum_{j=1}^m q_j\right). \tag{3}$$

The Nash problem is then equivalent to the following variational inequality: for a.e. $\omega \in \Omega$, find $q^*(\omega) \in K(\omega)$ such that

$$\sum_{j=1}^m F_j[\omega, q^*(\omega)](q_j - q_j^*(\omega)) \geq 0, \ \forall q \in K(\omega). \tag{4}$$

Since $F(\omega, \cdot)$ is continuous, and $K(\omega)$ is convex and compact, problem (4) is solvable for almost every $\omega \in \Omega$, due to the Stampacchia's theorem. Moreover, we assume that $F(\omega, \cdot)$ is monotone, i.e.:

$$\sum_{i=1}^m (F_i(\omega, q) - F_i(\omega, q'))(q_i - q_i') \geq 0 \ \ \forall \omega \in \Omega, \forall q, q' \in \mathbb{R}^m.$$

F is said to be strictly monotone if the equality holds only for $q = q'$ and in this case (4) has a unique solution. In the sequel the following uniform strong monotonicity property will be useful:

$$\exists \alpha > 0 : \sum_{i=1}^m (F_i(\omega, q) - F_i(\omega, q'))(q_i - q_i') \geq \alpha \|q - q'\|^2 \ \ \forall \omega \in \Omega, \forall q, q' \in \mathbb{R}^m. \tag{5}$$

Although the uniform strong monotonicity property is quite demanding, nonetheless it is verified by some classes of utility functions frequently used in the literature (see, e.g., Sect. 4).

3 The Lebesgue Space Formulation

Now we are interested in computing statistical quantities associated with the solution $q^*(\omega)$, in particular its mean value. For this purpose we introduce a Lebesgue space formulation of problems (2) and (4). Moreover, in view of the numerical approximation of the solution, from now on, we assume that the random and the deterministic part of the operator can be separated. Thus, let:

$$w_i(\omega,q) = p\left(\sum_{j=1}^{m} q_j\right) + \beta(\omega) - \alpha(\omega) f_i(q_i) - g_i(q_i)$$

where α, β are real random variables, with $0 < \underline{\alpha} \le \alpha(\omega) \le \overline{\alpha}$, and the part of the cost which is affected by uncertainty is denoted now by f_i (with an abuse of notation). As a consequence, the operator F takes the form:

$$F_i(\omega,q) = \alpha(\omega)\frac{\partial f_i(q_i)}{\partial q_i} + \frac{\partial g_i(q_i)}{\partial q_i} - p\left(\sum_{j=1}^{m} q_j\right) - \beta(\omega) - \frac{\partial p\left(\sum_{j=1}^{m} q_j\right)}{\partial q_i} q_i.$$

The separation of variables allows us to use the approximation procedure developed in [6]. Furthermore, we assume that F is uniformly strongly monotone according to (5) and satisfies the following growth condition:

$$|F_i(\omega,q)| \le c(1+|q|), \forall q \in \mathbb{R}^m, \forall \omega \in \Omega, \forall i \tag{6}$$

and $w_i(\omega,0) \in L^1(\Omega)$. Moreover, we shall assume that $\alpha \in L^\infty(\Omega)$, while $\beta, \overline{q}_i \in L^2(\Omega)$. Under these assumptions the following Nash equilibrium problem can be derived (see [9] or [3] for a similar derivation which can be easily extended to our functional setting):

Find $u^* \in L^2(\Omega, P, \mathbb{R}^m)$ such that, $\forall i$

$$\int_\Omega w_i(\omega, u^*(\omega)) dP_\omega = \max_{0 \le u_i \le \overline{q}_i} \int_\Omega w_i(\omega, (u_i(\omega), u^*_{-i}(\omega)) dP_\omega, \tag{7}$$

where we used the notation: $(u_i, u^*_{-i}) := (u^*_1, \ldots, u^*_{i-1}, u_i, u^*_{i+1}, \ldots, u^*_m)$. Then, we define a closed and convex set K_P by

$$K_P = \{u \in L^2(\Omega, P, \mathbb{R}^m) : 0 \le u_i(\omega) \le \overline{q}_i(\omega), P\text{-a.s.}, \forall i\}$$

and consider the variational inequality formulation of (7): Find $u^* \in K_P$ such that

$$\int_\Omega \sum_{j=1}^{m} F_j(\omega, u^*(\omega))(u_j(\omega) - u^*(\omega)) \ge 0, \forall u \in K_P. \tag{8}$$

The relation between problems (7) and (8) is clarified by the following theorem.

Theorem 2. *u^* is a solution of (7) if and only if it is a solution of (8).*

Proof. The proof can be obtained along the same lines as in [3], with minor modifications. □

Since the stochastic oligopolistic market problem will be studied through (8), we ensure its solvability by the following:

Theorem 3. *Let $f_i(\cdot,q_i), p(\cdot,\sum_{j=1}^m q_j)$ be measurable, and $f_i(\omega,\cdot), d_i(\omega,\cdot)$ are of class C^1. Let F be uniformly strongly monotone and satisfy the growth condition (6). Then (8) admits a unique solution.*

Proof. Under our assumption $F : \Omega \times \mathbb{R}^m \to \mathbb{R}^m$ is a Carathéodory function and it is well known that for each measurable function $u(\omega)$, the function $F(\omega,u(\omega))$ is also measurable. Under the growth condition (6) the superposition operator $N_F : u(\omega) \to F(\omega,u(\omega))$ maps $L^2(\Omega,P,\mathbb{R}^m)$ in $L^2(\Omega,P,\mathbb{R}^m)$ and is continuous, being P a probability measure. Moreover the uniform strong monotonicity of F implies the strong monotonicity of N_F. The set K_P is convex, closed, and (norm) bounded, hence weakly compact. Then, monotone operator theory applies (see, e.g., [11] for a recent survey on existence theorems) and (8) admits a unique solution. □

Remark 1. The Lebesgue formulation is the natural one for our stochastic problem, in that the solution of (8) is a function which, by definition, admits finite mean value and variance. If the unique solution of (4) is square integrable, then it also satisfies (8) (see also Proposition 1 in [8]).

Let us note that we worked with the abstract probability space (Ω,P) up to this point, and this was sufficient in providing the general formulation of our problem in Lebesgue spaces in a concise manner. However, in concrete applications the sample space Ω is not known. On the other hand, one can measure the distributions of the real valued random variables that are involved in the model. Hence, it is natural to work with the probability distributions induced on the images of the functions: $A = \alpha(\omega), B = \beta(\omega), Q_i = \overline{q}_i(\omega)$. Thus, let $y = (A,B,Q)$ and consider the probability space $(\mathbb{R}^d,\mathbb{P})$ with $d = 2+m$. In order to formulate the problem (8) in the image space we introduce the closed convex set $K_{\mathbb{P}}$ by:

$$K_{\mathbb{P}} = \{u \in L^2(\mathbb{R}^d,\mathbb{P},\mathbb{R}^m) : 0 \le u_i(A,B,Q) \le Q_i, \forall i, \mathbb{P}\text{-a.s.}\}$$

and consider the following problem: Find $u^* \in K_{\mathbb{P}}$ such that $\forall u \in K_{\mathbb{P}}$

$$\int_{\mathbb{R}^d} \sum_{i=1}^m \left[A\frac{\partial f_i(u_i^*(y))}{\partial q_i} + \frac{\partial g_i(u_i^*(y))}{\partial q_i} - p\left(\sum_{j=1}^m u_j^*(y)\right) - B - \frac{\partial p\left(\sum_{j=1}^m u_j^*(y)\right)}{\partial q_i} u_i^* \right] (u_i(y) - u_i^*(y)) dP(y) \ge 0. \quad (9)$$

We assume that all the random variables are independent. Moreover, as it is verified in most applications, we assume that each probability distribution

is characterized by its density φ. Thus, we have $\mathbb{P} = \mathbb{P}_A \otimes \mathbb{P}_B \otimes \mathbb{P}_Q$, $dP_\alpha(A) = \varphi_\alpha(A)dA$, $dP_\beta(B) = \varphi_\beta(B)dB$, $dP_{\overline{q}}(Q) = \varphi_{\overline{q}}(Q)dQ$, where we used the compact notation $\varphi_x(X) = \prod_{i=1}^{n} \varphi_{x_i}(X_i)$. Hence, we can write (8) using the Lebesgue measure:

$$\int_{\underline{\alpha}}^{\overline{\alpha}} \int_{\mathbb{R}} \int_{\mathbb{R}^n_+} \sum_{i=1}^{m} \left[A \frac{\partial f_i(u_i^*(A,B,Q))}{\partial q_i} + \frac{\partial g_i(u_i^*(A,B,Q))}{\partial q_i} - p\left(\sum_{j=1}^{m} u_j^*(A,B,Q) \right) - B \right.$$
$$\left. - \frac{\partial p\left(\sum_{j=1}^{m} u_j^*(A,B,Q) \right) u_i^*}{\partial q_i} \right] (u_i(y) - u_i^*(y)) \varphi_\alpha(A) \varphi_\beta(B) \varphi_{\overline{q}}(Q) dA\, dB\, dQ \geq 0 \tag{10}$$

for all $u \in K_{\mathbb{P}}$. The advantage of this formulation is that it is suitable for an approximation procedure based on discretization and truncation. The approximation method is applied to the example presented in Sect. 4.1, for the details of the method we refer the interested reader to [6, 8]. The outcome of the above mentioned procedure is a sequence of simple functions $(u_k^*)_k$ which converges in L^2 to the exact solution u^* when $k \to \infty$ (see [8, Theorem 4.2]). We can then use this sequence to approximate the mean value of the solution, which is defined in the standard way as

$$\langle u^* \rangle := \int_{\mathbb{R}^d} u^*(y) dP(y).$$

4 A Class of Utility Functions

In this section we consider a random version of a class of utility functions widely used in the literature (see, e.g., [12, Chap. 6]) and show that these functions satisfy the theoretical requirements stated in the preceding section.
Thus, let

$$f_i(\omega, q_i) = a(\omega)\, a_i q_i^2 + b_i q_i + c_i$$
$$p\left(\omega, \sum_{i=1}^{m} q_i \right) = -d \sum_{i=1}^{m} q_i + e(\omega)$$

where $0 < \underline{a} \leq a(\omega) \leq \overline{a}$, $a \in L^\infty(\Omega)$, $e \in L^2(\Omega)$, and a_i, b_i, d, c_i are positive real numbers. Thus, $w_i(\omega, q) = -[a(\omega)\, a_i q_i^2 + b_i q_i + c_i] + (-d \sum_{i=1}^{m} q_i + e(\omega)) q_i$, and

$$F_i(\omega, q) = [2a(\omega)\, a_i + 2d] q_i + d \sum_{j \neq i} q_j + b_i - e(\omega) \tag{11}$$

For each ω the operator F consists of a linear part and a constant vector. The following theorem shows that $F(\omega, q)$ satisfies the monotonicity requirement mentioned in the previous section.

Theorem 4. *Let $F : \Omega \times \mathbb{R}^m \to \mathbb{R}^m$ defined as in (11). Then F is strongly monotone, uniformly with respect to ω.*

Proof. Let T be the matrix associated to the linear part of F. A straightforward computation gives that the diagonal elements of T are $2a(\omega)a_i + 2d$ while its off diagonal elements are all equal to d. Now let us decompose T as the sum of three matrices:

$$T = 2a(\omega)\, diag(a_1, a_2, \ldots, a_m) + d\, I_m + \mathbf{d} \tag{12}$$

The first matrix is a diagonal matrix which as $a(\omega)\min_i\{a_i\}$ as its minimum eigenvalue. Given that $0 < \underline{a} \leq a(\omega)$ this matrix is positive definite, uniformly with respect to ω. The second matrix is a scalar matrix, and because d is strictly positive this matrix is positive definite. The third matrix, $\mathbf{d}$, has each entry equal to d, hence it is positive semidefinite. Hence, T is positive definite, uniformly with respect to ω, and as a consequence, F is strongly monotone, uniformly with respect to ω. □

4.1 Numerical Example

We consider the random version of a classical oligopoly problem presented in [12] where three producers are involved in the production of a homogeneous commodity. The cost f_i of producing the commodity by firm i and the demand function p are given by

$$\begin{aligned} f_1(\omega, q_1) &= a(\omega)q_1^2 + q_1 + 1 \\ f_2(\omega, q_2) &= 0.5a(\omega)q_2^2 + 4q_2 + 2 \\ f_3(\omega, q_3) &= a(\omega)q_3^2 + 0.5q_3 + 5 \\ p\left(\omega, \sum_{i=1}^{3}\right) &= -\sum_{i=1}^{3} q_i + e(\omega) \end{aligned}$$

where $a(\omega)$ and $e(\omega)$ are random parameters that follow truncated normal distributions:

$$\begin{aligned} a &\sim 0.5 \leq N(1, 0.25) \leq 1.5 \\ e &\sim 4.5 \leq N(5, 0.25) \leq 5.5 \end{aligned}$$

Although we do not put upper bounds on the production capabilities, the existence of the solution is ensured because of the coercivity of the operator generated by f and p. Solution of the nonrandom problem $(q_1, q_2, q_3) = (23/30, 0, 14/15)$ where $a(\omega) \equiv 1$, $e(\omega) \equiv 5$ is given in [12]. We use the following approximation procedure to evaluate mean value of q (see [6] for a detailed description of the method). First, we choose a discretization of the parameter domain $[0.5, 1.5] \times [4.5, 5.5]$ using

$N_1 \times N_2$ points and solve the problem for each pair $(a(i), e(j))$ using an extragradient method. Then we evaluate the mean value of q by using appropriate probability distribution functions. Approximate mean values of q_1, q_2, and q_3 are shown in Table 1.

Table 1 Mean value of $q = (q_1, q_2, q_3)$

	$N_1 = 100, N_2 = 100$	$N_1 = 200, N_2 = 200$	$N_1 = 400, N_2 = 400$
$\langle q_1 \rangle$	0.76935	0.77154	0.77262
$\langle q_2 \rangle$	$2.903E{-}08$	$2.9109E{-}08$	$2.9185E{-}08$
$\langle q_3 \rangle$	0.94103	0.9436	0.94487

5 Conclusions and Future Developments

We used the theory of random variational inequalities to incorporate uncertain data in an oligopolistic market model. The model presented makes use of quadratic cost functions and a linear demand price, which yields to a linear random variational inequality. In future work we plan to treat other classes of functions which yield to nonlinear variational inequalities and to perform more extended numerical experiments.

References

1. Baiocchi, C., Capelo, A.: Variational and Quasivariational Inequalities: Applications to Free Boundary Problems. Wiley, Chichester (1984)
2. Cournot, A.A.: Researches into the Mathematical Principles of the Theory of Wealth, 1838 (English Translation). MacMillan, London (1897)
3. Faraci, F., Raciti, F.: On generalized Nash equilibrium in infinite dimension: the Lagrange multipliers approach. Optimization (published online first, December 2012). doi:10.1080/02331934.2012.747090
4. Gabay, D., Moulin, H.: On the uniqueness and stability of Nash Equilibria in noncooperative games. In: Bensoussan, A., Kleindorfer, P., Tapiero, C.S. (eds.) Applied Stochastic Control in Econometrics and Management Sciences, pp. 271–294. North Holland, Amsterdam (1980)
5. Gwinner, J., Raciti, F.: Random equilibrium problems on networks. Math. Comput. Model. **43**, 880–891 (2006)
6. Gwinner, J., Raciti, F.: On a class of random variational inequalities on random sets. Numer. Funct. Anal. Optim. **27**(5–6), 619–636 (2006)
7. Gwinner, J., Raciti, F.: On monotone variational inequalities with random data. J. Math. Inequalities **3**(3), 443–453 (2009)
8. Gwinner, J., Raciti, F.: Some equilibrium problems under uncertainty and random variational inequalities. Ann. Oper. Res. **200**, 299–319 (2012). doi:10.1007/s10479-012-1109-2
9. Jadamba, B., Raciti, F.: On the modelling of some environmental games with uncertain data. J. Optim. Theory Appl. (published online first). doi:10.1007/s10957-013-0389-2

10. Jadamba, B., Khan, A.A., Raciti, F.: Regularization of stochastic variational inequalities and a comparison of an L_p and a sample-path approach. Nonlinear Anal. **94**, 65–83 (2014). http://dx.doi.org/10.1016/j.na.2013.08.009
11. Maugeri. A., Raciti, F.: On existence theorems for monotone and nonmonotone variational inequalities. J. Convex Anal. **16**(3 and 4), 899–911 (2009)
12. Nagurney, A.: Network Economics: A Variational Inequality Approach, 2nd and revised edn. Kluwer, Dordrecht (1999)
13. Ravat, U., Shanbhag, U.V.: On the existence of solutions to stochastic variational inequality and complementarity problems. arXiv:1306.0586v1 [math.OC] (3 Jun 2013)

Computing Area-Tight Piecewise Linear Overestimators, Underestimators and Tubes for Univariate Functions

Josef Kallrath and Steffen Rebennack

1 Introduction

The motivation for this publication is to follow-up on a previous work by Rebennack and Kallrath [11] to construct over- and underestimators for one-dimensional functions. These over- and underestimators are used to replace non-linear expressions by piecewise linear ones with the idea to approximate a non-linear (and non-convex) core and to place it into a large mixed-integer linear programming (MILP) problem. If the approximations of the feasible region and/or the objective function are constructed carefully, then the resulting MILP problem yields a lower bound (for minimization problems). In some applications, it is important to detect infeasibility of the original non-convex mixed-integer non-linear programming problem (MINLP). Again, careful use of over- and underestimators allows for the safe conclusion of infeasibility of the original MINLP from the infeasibility of the approximate MILP problem; cf. [11, Sect. 3.3].

The concept of approximating non-linear functions by piecewise linear ones has been around for some time. However, new developments in efficient representation of the resulting breakpoint systems [15] have lead to more interest in piecewise linear approximators. Recently, Misener and Floudas [8, 9] utilize such approximators for relaxations (underestimators) when solving mixed-integer quadratically-constrained quadratic programs.

The automatic computation of optimal breakpoint systems, however, received very little treatment in the literature. The seminal work by Rosen and Pardalos [13] and Pardalos and Rosen [10, Chap. 8] uses a system of equidistant breakpoints to

J. Kallrath
Department of Astronomy, University of Florida, Gainesville, FL 32611, USA
e-mail: josef.kallrath@web.de

S. Rebennack (✉)
Colorado School of Mines, Division of Economics and Business, Golden, CO 80401, USA
e-mail: srebenna@mines.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_14,

achieve a predefined maximal deviation between a concave quadratic function and the piecewise linear approximator. Geißler [1] and Geißler et al. [2] can compute piecewise linear approximators (over- and underestimators) automatically when certain assumptions on the functions are satisfied. For more than one dimension, Misener and Floudas [7] utilize piecewise linear formulations via simplices; Rebennack and Kallrath [12] use triangulations.

In Rebennack and Kallrath [11], we minimize the number of breakpoints used to achieve a maximal deviation of δ between the piecewise linear approximator and the original function. Furthermore, we constructed tight approximators by minimizing the maximal vertical distance between the approximator and the original function, for a given number of breakpoints. In this paper, we utilize an area-based tightness definition: allowing a maximal deviation of $\delta > 0$ and $B \in \mathbb{N} \geq 2$ breakpoints, we seek a piecewise linear, continuous approximator which minimizes the area between the approximator and the original function. Minimizing the error between the approximator and the original function through an area-based measure is expected to produce better results (e.g., tighter bounds) when replacing non-linear functions by piecewise linear ones, compared to approaches which ignore any tightness measure.

The idea of minimizing the area between a function is briefly mentioned in Geyer et al. [3]. However, their paper does not further follow this idea but rather prefers a curvature-based approach pointing out that this is of similar quality than using vertical distances or an area-based approach. Different to our approach, they cannot guarantee the computation of an optimal breakpoint system.

The contributions of this article are as follows. For univariate functions, we develop methodologies to compute over- and underestimators as well as tubes which are (1) continuous, (2) do not deviate more than a given tolerance $\delta > 0$ from the original function, (3) stay above (for overestimators), below (for underestimators) or a combination of both (for tubes) and are (4) area-minimizing. Thus, it is the first paper to describe a framework to automatically compute (optimal) area-minimizing breakpoint systems for univariate functions.

The remainder of the paper is organized as follows: in Sect. 2, we provide various definitions in the context of piecewise linear approximators. We treat over- and underestimators in Sect. 3, tubes in Sect. 4 and approximators in Sect. 5. Section 6 contains our computational results. We conclude with Sect. 7.

2 Definitions

The original (non-linear, non-convex, continuous, and real) function to be approximated is $f(x)$ over the compactum $[X^-, X^+] \subset \mathbb{R}$. We denote by $\ell(x) : [X^-, X^+] \to \mathbb{R}$ a function approximating $f(x)$.

We start with the definition of a δ-approximator for univariate functions.

Definition 1 (δ-Approximator, [11]). Let $f(x) : [X^-, X^+] \to \mathbb{R}$ be a univariate function and let scalar $\delta > 0$. A piecewise linear, continuous function $\ell(x) : [X^-, X^+] \to \mathbb{R}$ is called a δ-*approximator* for $f(x)$, if the following property holds

$$\max_{x\in[X^-,X^+]} |\ell(x) - f(x)| \leq \delta. \tag{1}$$

We require for the piecewise linearity property of a function that the function is non-differentiable at a finite number of points. δ-Over- and δ-underestimators are δ-approximators with the additional requirement to stay above or below function $f(x)$ in the domain $[X^-,X^+]$. This is formalized in

Definition 2 (δ-Overestimator/δ-Underestimator, [11]). We call a piecewise linear, continuous function $\ell^+(x) : [X^-,X^+] \to \mathbb{R}$ a *δ-overestimator* for function $f(x) : [X^-,X^+] \to \mathbb{R}$, if condition (1) is satisfied along with

$$\ell^+(x) \geq f(x), \quad \forall\, x \in [X^-,X^+]. \tag{2}$$

We call a piecewise linear, continuous function $\ell^-(x)$ a *δ-underestimator* of function $f(x)$, if $-\ell^-(x)$ is a δ-overestimator of $-f(x)$.

We continue with the definition of a δ-tube.

Definition 3 (δ-Tube). We call any combination of a piecewise linear, continuous δ-overestimator $\ell^+(x)$ for function $f(x)$ and a piecewise linear, continuous δ-underestimator $\ell^-(x)$ for $f(x)$ a *δ-tube* for $f(x)$.

The definitions of δ-approximators, δ-overestimators, δ-underestimators, and δ-tubes require piecewise linearity and continuity. Thus, we will no longer mention these function properties explicitly in the remainder of the paper, except if we want to emphasis these two properties.

Given univariate function $f(x)$ over a compactum and the δ-tolerance, we have two desires on an automatic procedure: (1) it computes δ-approximators, δ-overestimators and/or δ-underestimators and (2) the number of required breakpoints (i.e., discontinuities) is minimal. This has been achieved already [11]. Their approach can easily be extended to compute δ-tubes which require the minimal number of breakpoints; in most cases, such optimal δ-tubes exhibit the property that the breakpoint systems of the δ-overestimator and δ-underestimator are identical, i.e., both the δ-overestimator and δ-underestimator share the same discontinuities.

Vice-versa, one can provide the number of breakpoints and ask for the "tightest" possible δ-approximator, δ-overestimator, δ-underestimator, and δ-tube. In [11], the authors use an absolute function deviation error tolerance criterion as a tightness definition:

Definition 4 (Absolute-Error-Tolerance-Tightness (AETT), [11]). A δ-approximator, δ-overestimator, δ-underestimator, or δ-tube with B breakpoints for function $f(x)$ is called *tighter* (in the absolute-error-tolerance sense) than a ϑ-approximator, ϑ-overestimator , ϑ-underestimator, or ϑ-tube, respectively, with B breakpoints for function $f(x)$, if $\delta < \vartheta$. A δ-approximator, δ-overestimator, δ-underestimator or δ-tube with B breakpoints is called *tight* (in the absolute-error-tolerance sense) for $f(x)$, if there is no *tighter* ϑ-approximator, ϑ-overestimator, ϑ-underestimator, or ϑ-tube for $f(x)$.

In this paper, we utilize an area-based tightness definition:

Definition 5 (Area-Tightness (AT)). Let $\ell(x)$ be a δ-approximator, δ-overestimator, δ-underestimator, or δ-tube with B breakpoints for function $f(x)$. Further, let A_1 be the area between $\ell(x)$ and $f(x)$ over the compactum $[X^-, X^+]$. Another δ-approximator, δ-overestimator, δ-underestimator, or δ-tube with B breakpoints for function $f(x)$ and area A_2 is called *tighter* (in the area sense) than $\ell(x)$ for function $f(x)$, if $A_2 < A_1$. $\ell(x)$ is called *tight* (in the area sense) for $f(x)$, if there is no *tighter* δ-approximator, δ-overestimator, δ-underestimator, or δ-tube with B breakpoints for function $f(x)$.

To compute an area-tight δ-approximator, δ-overestimator, δ-underestimator, or δ-tube, we treat the error-tolerance, δ, and the number of breakpoints, B, as input parameters. Thus, we more precisely call them (δ,B)-approximator, (δ,B)-overestimator, (δ,B)-underestimator, or (δ,B)-tube.

Interestingly, AETT is preserved when shifting an absolute-error-tolerance-tight (δ,B)-approximator to obtain a (δ,B)-overestimator or (δ,B)-underestimator.

Corollary 1 ([11]). *Let $\ell(x) : [X^-, X^+] \to \mathbb{R}$ be an absolute-error-tolerance-tight (δ,B)-approximator for $f(x)$ and let $\varepsilon = 2\delta$. Then $\ell^+(x) := \ell(x) + \delta$ and $\ell^-(x) := \ell(x) - \delta$ define an absolute-error-tolerance-tight (ε,B)-underestimator and an absolute-error-tolerance-tight (ε,B)-overestimator, respectively, for $f(x)$ with the same number of breakpoints B.*

For AETT, it therefore suffice to develop one single algorithm to compute optimal (δ,B)-approximators, (δ,B)-overestimators and/or (δ,B)-underestimators; a different procedure is required for absolute-error-tolerance-tight (δ,B)-tubes. Unfortunately, AT is not preserved through (careful) shifting.

We present algorithms to compute area-tight (δ,B)-overestimators and (δ,B)-underestimators in Sect. 3, area-tight (δ,B)-tubes in Sect. 4 and area-tight (δ,B)-approximators in Sect. 5. However, before we proceed with the methodology, we discuss how to choose the two parameters: the absolute-error tolerance, δ, and the number of breakpoints, B. Dependent on the application, we might want to follow one of the following two paths.

If we desire to compute an approximate solution to the original MINLP problem with a specific tolerance guarantee in mind (e.g., a safe gap of $\varepsilon > 0$) via piecewise linear approximations, one needs to compute δ-approximators, δ-overestimators, δ-underestimators or δ-tubes with a certain absolute tolerance δ and apply them appropriately; cf. [11, Sect. 3.3]. In this case, we might want to proceed as follows:

1. first, compute the minimum number of breakpoints, B^*, needed to obtain a given δ-approximation (as discussed in [11]),
2. second, compute an absolute-error-tolerance-tight approximator—(ϑ,B^*)-approximator, (ϑ,B^*)-overestimator, (ϑ,B^*)-underestimator, or (ϑ,B^*)-tube—using B^* breakpoints ($\vartheta \leq \delta$; as discussed in [11]), and
3. third, compute an area-tight approximator—(ϑ,B^*)-approximator, (ϑ,B^*)-overestimator, (ϑ,B^*)-underestimator, or (ϑ,B^*)-tube.

Instead of pre-defining the tolerance (δ dependent on ε) to achieve a good lower bound for minimization problems, we might provide the number of breakpoints, B, to be spend on the piecewise linear approximators. The number of breakpoints directly affect the model size in the MILP framework. Thus, we might want to choose the number of breakpoints in such a way that the resulting MILP problem remains efficiently solvable with (standard) solvers. Another reason for pre-defining the number of breakpoints are the use of logarithmic representations in the number of breakpoints (both in the number of binary variables and constraints involved) of the resulting breakpoint system; it is efficient to choose B as a power of 2. Given B, we would skip the first step above and compute an absolute-error-tolerance-tight approximator yielding the tolerance δ. This allows for the computation of an area-tight approximator using δ and B.

3 Computing Area-Tight (δ, B)-Overestimators and (δ, B)-Underestimators

We are given the absolute-error tolerance δ (i.e., maximal vertical absolute difference between the function $f(x)$ and the approximator $\ell(x)$) and the number of breakpoints, B, for the univariate function $f(x)$ along with the closed interval $[X^-, X^+]$. We seek to automatically compute area-tight (δ, B)-overestimators. The case of area-tight (δ, B)-underestimators follows the same logic; we discuss it in brief at the end of the section as well.

For the following discussions, we require:

- $f(x) - \delta \geq 0$ for all $x \in [X^-, X^+]$, and
- $X^- \geq 0$.

Both requirements can be achieved through a shift in either the function value direction ($f(x)$ attains a minimum in $[X^-, X^+]$, cf. Extreme Value Theorem) or the x-axis direction.

For our derivations, we assume that the primitive of $f(x)$ exists and we denote it by $F(x)$, for $x \in [X^-, X^+]$. We do not require its existence for our computations, though. We are interested in minimizing the area between function $f(x)$ and the piecewise linear function $\ell^+(x)$; let $L^+(x)$ denote the primitive of $\ell^+(x)$. Therefore, we need to compute the area between the two functions. Let $x_b \in [X^-, X^+]$ denote the x-value (i.e., footpoint) of the bth breakpoint and let $\ell^+(x_b)$ be its corresponding function value. Then, the area between $f(x)$ and $\ell^+(x)$ can be calculated as

$$\int_{X^-}^{X^+} (\ell^+(x) - f(x))dx = \sum_{b=1}^{B-1} [L^+(x) - F(x)]_{x_b}^{x_{b+1}}$$

$$= \sum_{b=1}^{B-1} (L^+(x_{b+1}) - L^+(x_b)) + F(x_1) - F(x_B)$$
$$= \frac{1}{2} \sum_{b=1}^{B-1} (\ell^+(x_{b+1}) + \ell^+(x_b))(x_{b+1} - x_b) + F(x_1) - F(x_B).$$

Note that the first identity is true because the approximator, $\ell^+(x)$, never crosses the function $f(x)$, cf. requirement (2).

We define $x_1 := X^-$ and $x_B := X^+$ implying that both $F(x_1)$ and $F(x_B)$ are fixed, i.e., they are constants. Thus, we are interested in minimizing the non-linear expression

$$\sum_{b=1}^{B-1} (\ell^+(x_{b+1}) + \ell^+(x_b))(x_{b+1} - x_b).$$

Notice that we do not require the primitive (or its existence) of function $f(x)$ anymore; the numerical value of $\int_{X^-}^{X^+} f(x)dx$ suffices.

Next, we need to model the decisions on the placement of the B breakpoints, via decision variables x_b ($x_b \in [X^-, X^+]$, $x_{b+1} > x_b$, $b = 2, \ldots, B-1$), and the function values of $\ell^+(x)$ at the breakpoints, via the shit variables s_b ($s_b \in [-\delta, \delta]$, $b = 1, \ldots, B$) with respect to $f(x)$. In this context, we define

$$\phi(x_b) := f(x_b) + s_b, \quad \forall\, b = 1, \ldots, B \tag{3}$$

which equals $\ell^+(x_b)$. The approximator $\ell^+(x)$ is then the corresponding interpolator between the values of $\phi(x_b)$.

Further, we need to ensure conditions (1) and (2). Both requirements lead to semi-infinite programming problems because an infinite number of (non-linear, non-convex) constraints need to hold; cf. Hettich and Kortanek [4] or Lopez and Still [5]. We follow the idea of formulation OBSD as described in [11] and discretize each interval (x_{b-1}, x_b) into I equidistant grid points. Conditions (1) and (2) need then to hold on this finite grid; we increase the number of grid points dynamically until a pre-defined tolerance has been reached.

This leads us to the following (non-convex) non-linear programming (NLP) problem, computing an area-tight (δ, B)-overestimator for the continuous function $f(x)$ on the interval $[X^-, X^+]$:

$$\tilde{A}^+(\delta, B, I, M) :=$$
$$\min \sum_{b=1}^{B-1} \Big(\phi(x_{b+1}) + \phi(x_b)\Big)\Big(x_{b+1} - x_b\Big) \tag{4}$$
$$\text{s.t.} \quad x_b - x_{b-1} \geq \frac{1}{M}, \quad \forall\, b = 2, \ldots, B \tag{5}$$
$$x_{bi} = x_{b-1} + \frac{i}{I+1}(x_b - x_{b-1}), \quad \forall\, b = 2, \ldots, B, \quad i = 1, \ldots, I \tag{6}$$

$$l_{bi} = \phi(x_{b-1}) + \frac{\phi(x_b) - \phi(x_{b-1})}{x_b - x_{b-1}} (x_{bi} - x_{b-1}),$$
$$\forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{7}$$
$$l_{bi} - f(x_{bi}) \leq \delta, \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{8}$$
$$l_{bi} \geq f(x_{bi}), \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{9}$$
$$x_1 = X^-, \quad x_B = X^+ \tag{10}$$
$$x_b \in [X^-, X^+], \quad \forall\, b = 2,\ldots,B-1 \tag{11}$$
$$x_{bi} \in [X^-, X^+], \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{12}$$
$$l_{bi} \text{ free}, \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{13}$$
$$s_b \in [0, \delta], \quad \forall\, b = 1,\ldots,B. \tag{14}$$

The logic of the constraint set (5)–(14) is as follows. Constraints (5) ensure the sorting of the breakpoints and that no two breakpoints can be identical. This becomes numerically important to avoid a division by zero when calculating the slope of the approximator $\ell^+(x)$. The value of the constant M needs to be chosen carefully in order to avoid exclusion of an optimal distribution of the breakpoints. Actually, it is non-trivial to mathematical (and computational) safely conclude what a sufficiently large value for M is. Constraints (6) model the I grid points, x_{bi}, for the interval (x_{b-1}, x_b). These grid points are the discretization introduced in order to ensure that (I) the maximal vertical distance between function $f(x)$ and the approximator $\ell^+(x)$ is at most δ, as required in (1) and modeled via (7) and (8), and that (II) approximator $\ell^+(x)$ stays above function $f(x)$ as required in (2) and modeled via (7) and (9). Constraints (10)–(14) model the variables' domain.

The mathematical model (4)–(14) is non-linear, non-convex and continuous: It consists of $2B + 2(B-1)I - 2$ continuous variables and $B + 4(B-1)I - 1$ constraints; the objective function (4) as well as constraints (7)–(9) is non-convex.

If the NLP (4)–(14) is infeasible, then there are two possibilities: either M is too small or the combination of δ and B does not allow for the existence of a (δ, B)-overestimator.

The idea of the objective function (4) is intuitive: We minimize the area of the approximator $\ell^+(x)$ and the x-axis; constraints (9) ensure that $\ell^+(x)$ always stays above function $f(x)$. Given a sufficiently large value for M denoted by M^*, we can recover a lower bound $\underline{A}^+$ on the area A between the approximator $\ell^+(x)$ and the original function $f(x)$ via

$$\underline{A}^+ = \tfrac{1}{2}\tilde{A}^+(\delta, B, I, M^*) + F(x_1) - F(x_B). \tag{15}$$

Equation (15) constitutes a lower bound on the area A because both conditions (1) and (2) are relaxed; they hold only on a finite number of (grid) points.

After solving (4)–(14) to (local or global) optimality, we solve (to global optimality)

$$\mu^+(I) := \max_{b=2,\ldots,B} \mu_b^+(I) := \max_{b=2,\ldots,B} \max_{x \in [x_{b-1}, x_b]} (\ell^+(x) - f(x)) \tag{16}$$

in order to compute the maximal vertical deviation between $f(x)$ and the computed approximator $\ell^+(x)$ in the interval $[X^-, X^+]$. If

$$\mu^+(I) \leq \delta, \tag{17}$$

then condition (1) holds true and the computed $\ell^+(x)$ defines a (δ, B)-approximator for $f(x)$.

We further need to check if $\ell^+(x)$ is below function $f(x)$ somewhere in the interval (X^-, X^+). Therefore, we solve (to global optimality)

$$\psi^+(I) := \min_{b=2,\ldots,B} \psi_b^+(I) := \min_{b=2,\ldots,B} \min_{x\in[x_{b-1},x_b]} (\ell^+(x) - f(x)). \tag{18}$$

If

$$\psi(I) \geq 0, \tag{19}$$

then condition (2) holds true. If both (17) and (19) are satisfied, then $\ell^+(x)$ defines an area-tight (δ, B)-overestimator for $f(x)$ with $A = \underline{A}^+$.

If (17) or (19) are violated by more than a pre-defined tolerance $\eta > 0$, then we increase the number of grid points, I, and re-solve (4)–(14) as well as (16) and (18). For any desired precision $\eta > 0$, this process, of increasing I, is finite (granted that the NLP problems can be solved to global optimality).

Corollary 2. *Let $f(x)$ be a continuous function on $[X^-, X^+]$, $\delta > 0$ and $B \in \mathbb{N} \geq 2$ be fixed. Then, for each $\eta > 0$, there exists a finite I^*, such that $\mu(I^*) \leq \delta + \eta$ and $\psi(I^*) \geq -\eta$, given that there exists a (δ, B)-overestimator for $f(x)$.*

The proof of Corollary 2 is based on the continuity of $f(x)$ over a compactum and follows from Rebennack and Kallrath [11, Corollary 6].

Following the same logic as for the area-tight (δ, B)-overestimator, we compute an area-tight (δ, B)-underestimator, $\ell^-(x)$, for $f(x)$ on the interval $[X^-, X^+]$:

$$\tilde{A}^-(\delta, B, I, M) :=$$

$$\max \sum_{b=1}^{B-1} (\phi(x_{b+1}) + \phi(x_b))(x_{b+1} - x_b) \tag{20}$$

$$\text{s.t.} \quad (5)\text{–}(7)\ ,\ (10)\text{–}(13) \tag{21}$$

$$f(x_{bi}) - l_{bi} \leq \delta, \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{22}$$

$$l_{bi} \leq f(x_{bi}), \quad \forall\, b = 2,\ldots,B, \quad i = 1,\ldots,I \tag{23}$$

$$s_b \in [-\delta, 0], \quad \forall\, b = 1,\ldots,B. \tag{24}$$

Analogously, the condition (1) reads for underestimators

$$\mu^-(I) := \max_{b=2,\ldots,B} \mu_b^-(I) := \max_{b=2,\ldots,B} \max_{x\in[x_{b-1},x_b]} (f(x) - \ell^-(x)) \tag{25}$$

and (2) is

$$\psi^-(I) := \min_{b=2,\ldots,B} \psi_b^-(I) := \min_{b=2,\ldots,B} \min_{x\in[x_{b-1},x_b]} (f(x)-\ell^-(x)). \tag{26}$$

Function $\ell^-(x)$ defines an area-tight (δ,B)-underestimator for $f(x)$ with area

$$A = \tfrac{1}{2}\tilde{A}^-(\delta,B,I,M) + F(x_1) - F(x_B),$$

if both

$$\mu^-(I) \le \delta \qquad \text{and} \qquad \psi^-(I) \ge 0. \tag{27}$$

Corollary 2 reads now

Corollary 3. *Let $f(x)$ be a continuous function on $[X^-,X^+]$, $\delta > 0$ and $B \in \mathbb{N} \ge 2$ be fixed. Then, for each $\eta > 0$, there exists a finite I^*, such that $\mu^-(I^*) \le \delta + \eta$ and $\psi^-(I^*) \ge -\eta$, given that there exists a (δ,B)-underestimator for $f(x)$.*

4 Computing an Area-Tight (δ,B)-Tube: (δ,B)-Overestimators and (δ,B)-Underestimators Sharing the Same Breakpoint System

Recall that the purpose of piecewise linear approximations of functions is to replace a non-liner system of constraints or objective function by MILP constructs to be placed in a MILP framework. Therefore, consider a non-convex, continuous, univariate function $f(x)$ which appears as an equation

$$f(x) = b, \quad x \in [X^-,X^+]$$

in the constraints of the MINLP problem to be approximated. In this case, one would compute an area-tight (δ,B)-overestimator, $\ell^+(x)$, and an area-tight (δ,B)-underestimator, $\ell^-(x)$, for $f(x)$. When doing so, there is no guarantee that the breakpoint systems of $\ell^+(x)$ and $\ell^-(x)$ are identical. Most likely, we would require $2(B-1)$ breakpoints for the resulting δ-tube. Notice that the resulting tube might not be an area-tight $(\delta,2B-2)$-tube. For a given number of breakpoints, B, an are-tight (δ,B)-tube can be calculated when the (δ,B)-overestimator and the (δ,B)-underestimator share the same breakpoint system. Notice that the resulting (δ,B)-overestimator and (δ,B)-underestimator might not be area-tight, even though the (δ,B)-tube is.

Just like in the previous section, for notational convenience, we assume that

- $f(x) - \delta \ge 0$ for all $x \in [X^-,X^+]$, and
- $X^- \ge 0$.

For (δ,B)-overestimator, $\ell^+(x)$, and (δ,B)-underestimator, $\ell^-(x)$, sharing the same B breakpoints at x_b, the area of the resulting (δ,B)-tube is derived through

$$\begin{aligned}
&\int_{X^-}^{X^+} (\ell^+(x)-\ell^-(x))dx \\
&\qquad = \sum_{b=1}^{B-1} [L^+(x)-L^-(x)]_{x_b}^{x_{b+1}} \\
&\qquad = \sum_{b=1}^{B-1} (L^+(x_{b+1})-L^+(x_b)-L^-(x_{b+1})+L^-(x_b)) \\
&\qquad = \frac{1}{2}\sum_{b=1}^{B-1} (\ell^+(x_{b+1})+\ell^+(x_b))(x_{b+1}-x_b) \\
&\qquad\qquad -\frac{1}{2}\sum_{b=1}^{B-1} (\ell^-(x_{b+1})+\ell^-(x_b))(x_{b+1}-x_b).
\end{aligned}$$

Similar to (3), we define

$$\phi^+(x_b) := f(x_b)+s_b^+ \quad \text{and} \quad \phi^-(x_b) := f(x_b)+s_b^-, \quad \forall\, b=1,\ldots,B.$$

Following the idea of formulation (4)–(14), we obtain the following continuous, non-convex NLP problem, computing an area-tight (δ,B)-tube for the continuous function $f(x)$ on the interval $[X^-,X^+]$:

$$\tilde{A}^{\pm}(\delta,B,I,M) :=$$

$$\min \quad \frac{1}{2}\sum_{b=1}^{B-1} (\phi^+(x_{b+1})+\phi^+(x_b)-\phi^-(x_{b+1})-\phi^-(x_b))(x_{b+1}-x_b) \tag{28}$$

$$\text{s.t.} \quad (5),(6),\ (10)\text{–}(12) \tag{29}$$

$$l_{bi}^+ = \phi^+(x_{b-1}) + \frac{\phi^+(x_b)-\phi^+(x_{b-1})}{x_b-x_{b-1}}(x_{bi}-x_{b-1}), \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{30}$$

$$l_{bi}^+ - f(x_{bi}) \le \delta, \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{31}$$

$$l_{bi}^+ \ge f(x_{bi}), \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{32}$$

$$l_{bi}^- = \phi^-(x_{b-1}) + \frac{\phi^-(x_b)-\phi^-(x_{b-1})}{x_b-x_{b-1}}(x_{bi}-x_{b-1}), \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{33}$$

$$f(x_{bi}) - l_{bi}^- \le \delta, \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{34}$$

$$l_{bi}^- \le f(x_{bi}), \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{35}$$

$$l_{bi}^+,\ l_{bi}^- \text{ free}, \quad \forall\, b=2,\ldots,B, \quad i=1,\ldots,I \tag{36}$$

$$s_b^+ \in [0,\delta], \quad s_b^- \in [-\delta,0], \quad \forall\, b=1,\ldots,B. \tag{37}$$

Constraint group (29) models the breakpoint system, constraints (30)–(32) model the overestimator and (33)–(35) the underestimator.

The computed $\ell_+(x)$ defines a (δ,B)-overestimator, if both (17) and (19) hold true; $\ell_-(x)$ is a (δ,B)-underestimator, if both conditions in (27) hold. If all four conditions are satisfied, then $\ell^+(x)$ and $\ell^-(x)$ define an area-tight (δ,B)-tube for $f(x)$ on $[X^-,X^+]$ with area $\tilde{A}^{\pm}(\delta,B,I,M)$; otherwise, if at least one of the four conditions is violated, then the grid size I needs to be increased.

We also have a finite convergence argument for tubes.

Corollary 4. *Let $f(x)$ be a continuous function on $[X^-,X^+]$, $\delta>0$ and $B\in\mathbb{N}\geq 2$ be fixed. Then, for each $\eta>0$, there exists a finite I^*, such that $\max\{\mu^+(I^*),\mu^-(I^*)\}\leq \delta+\eta$ and $\min\{\psi^+(I^*),\psi^-(I^*)\}\geq -\eta$, given that there exists a (δ,B)-tube for $f(x)$.*

5 Computing Area-Tight (δ,B)-Approximators

δ-Approximators play the central role in the methodology developed by Rebennack and Kallrath [11], because they allow for the efficient computation of AETT δ-overestimators and δ-underestimators via a simple function value shift; minimality in the number of breakpoints required is preserved as well. The case for area-tight (δ,B)-approximators is different: AT is not preserved after a shifting operation.

Over-, underestimators and tubes are important constructs when replacing NLP problems; approximators are not equally important, as they do not allow for the computation of safe bounds and do not allow for infeasibility detection. Thus, we leave it at a sketch of the idea on how to compute an area-tight (δ,B)-approximator.

Approximators can intersect with the function $f(x)$, unlike over- and underestimators. This poses a challenge, when calculating the area between the approximator and the function. We use the following idea: given that we are working with a grid (the I discrete points) on the x-axis, we evaluate the relative position of the approximator $\ell(x)$ to the function $f(x)$ at these grid points by introducing the binary decision variables γ_{bi} with

$$-\delta\big(1-\gamma_{bi}\big)\leq f(x_{bi})-l_{bi}\leq\delta\gamma_{bi},\quad\forall\, b=2,\ldots,B,\quad i=1,\ldots,I.$$

If $f(x)$ is above (below) the approximator $\ell(x)$ at point x_{bi}, i.e., $f(x_{bi})>l_{bi}$ ($f(x_{bi})<l_{bi}$), then $\gamma_{bi}=1$ ($\gamma_{bi}=0$).

We consider only the case in which the primitive of function of f exists. We distinguish three cases on the relative position of the approximator to the function $f(x)$, to calculate an approximation of the area between $f(x)$ and $\ell(x)$

I: $\gamma_{bi}=\gamma_{b,i+1}=1$

$$F(x_{b,i+1})-F(x_{bi})-L(x_{b,i+1})+L(x_{bi})$$

this formula is precise if $f(x)\geq\ell(x)$ for all $x\in[x_{bi},x_{b,i+1}]$

II: $\gamma_{bi} = \gamma_{b,i+1} = 0$

$$-F(x_{b,i+1}) + F(x_{bi}) + L(x_{b,i+1}) - L(x_{bi})$$

this formula is precise if $f(x) \leq \ell(x)$ for all $x \in [x_{bi}, x_{b,i+1}]$

III: $\gamma_{bi} \neq \gamma_{b,i+1}$ the approximator ℓ intersects with the function f at least once in the interval $x \in [x_{bi}, x_{b,i+1}]$, we assign the area a value of 0.

The three cases above are restricted to the intervals $[x_{b1}, x_{bI}]$, $b = 2, \ldots, B$, and do neither consider the interval $[x_{b-1}, x_{b1}]$ nor $[x_{bI}, x_b]$ located around the breakpoints, $b = 1, \ldots, B$. Therefore, we introduce the binary decision variable γ_b with

$$-\delta\big(1 - \gamma_b\big) \leq s_b \leq \delta\gamma_b, \quad \forall\, b = 1, \ldots, B.$$

and derive the area of the intervals using the three cases above analogously.

The resulting mathematical programming problem is a MINLP, which is non-convex. The number of binary variables depends on the number of breakpoints, B, and the grid size, I. Therefore, we expect that the computation of area-tight (δ, B)-approximators is computationally much harder than the computation of area-tight (δ, B)-overestimators or area-tight (δ, B)-underestimators.

After the resulting MINLP has been solved, we check if the continuums-condition (1) is satisfied, via the solution of the global optimization problem

$$\mu^{\pm}(I) := \max_{b=2,\ldots,B} \mu_b(I) := \max_{b=2,\ldots,B} \max_{x \in [x_{b-1}, x_b]} \Big| \ell(x) - f(x) \Big|.$$

If $\mu(I) > \delta$, then we increase I and start-over; otherwise, $\ell(x)$ is a (δ, B)-approximator. The area computed as described above defines a lower bound on the area of an area-tight (δ, B)-approximator; an upper bound is obtained by evaluating the area between the calculated $\ell(x)$ and $f(x)$. If the lower and the upper bound on the area are close enough together, then we stop, otherwise we increase I further.

6 Computational Results

We execute our computational tests on an Intel(R) i7 @ 2.40 GHz with 8 GB RAM running 64-bit Windows 7. We use GAMS version 23.8 and solve all non-convex NLP problems with the global solver LindoGLOBAL [14] to an absolute gap (i.e., upper bound minus lower bound) of 10^{-5}.

Table 1 One-dimensional test functions taken from Rebennack and Kallrath [11]

#	$f(x)$	X_-	X_+	Comment
01	x^2	-3.5	3.5	Convex function; axial symmetric at $x = 0$
02	$\ln x$	1	32	Concave function
03	$\sin x$	0	2π	Point symmetric at $x = \pi$
04	$\tanh(x)$	-5	5	Strictly monotonically increasing; point symmetric at $x = 0$
05	$\frac{\sin(x)}{x}$	1	12	For numerical stability reason we avoid the removable singularity and the oscillation at 0, the two local minima have an absolute function value difference of ≈ 0.126
06	$2x^2 + x^3$	-2.5	2.5	In $(-\infty, \infty)$, there is one local minimum at $x = 0$ and one local maximum at $x = \frac{4}{3}$
07	$e^{-x}\sin(x)$	-4	4	One global minimum ($x_m \approx -2.356$ with $f(x_m) \approx -7.460$)
08	$e^{-100(x-2)^2}$	0	3	A normal distribution with a sharp peak at $x = 2$
09	$1.03e^{-100(x-1.2)^2} + e^{-100(x-2)^2}$	0	3	The sum of two Gaussians, with two slightly different maxima (their absolute function value difference is ≈ 0.030)
10	Maranas and Floudas [6]	0	2π	Three local minima (the absolute function value difference of the two smallest local minima is ≈ 0.031)

For our computational tests, we made the following selection for the parameters I, M and η. We start with a grid size of $I = 2$ and update the number of grid points according to the following formula

$$\max\{\lfloor 1.5I \rfloor, I + 1\}.$$

We choose $M = 10^{-5}$ as well as $\eta = 0.001$. We use the ten univariate functions, taken from the literature, as summarized in Table 1.

Table 2 summarized the computational results for area-tight (δ, B)-overestimators. We make the following observations: (I) area-tight (δ, B)-overestimators can only be computed for a few number of breakpoints; (II) the number of discretization

points (i.e., I) required to ensure a maximal violation of 0.001 of condition (2) (II.1) varies widely among the tested functions: if the function is convex (e.g., function 01), then any discretization suffices, and (II.2) decreases with an increase in the number of breakpoints; (III) the computational time tends to increase exponentially in the number of breakpoints.

Table 2 Area-tight (δ, B)-overestimators for the functions provided in Table 1

#	B	δ	$\underline{A}^+$	ψ^+	μ^+	I	Sec.
01	3	3.10	14.2917	0.0000000	3.063	2	0.19
	4	1.50	6.3519	0.0000000	1.361	2	0.91
	5	1.10	3.5729	0.0000000	0.766	2	24.35
	6	1.10	2.2867	0.0000000	0.490	2	329.53
	7	0.40	1.5880	0.0000000	0.340	2	0.78
	8	0.40	–	–	–	2	3,600.07[a]
02	3	1.00	2.4186	−0.0005192	0.900	9	4.38
	4	0.85	1.1780	−0.0005961	0.494	9	154.21
	5	0.45	–	–	–	2	3,600.10[a]
03	3	1.50	3.4820	−0.0005656	1.365	28	12.37
	4	0.40	0.7448	−0.0002769	0.278	28	50.67
	5	0.40	0.4484	−0.0004956	0.311	28	1,348.89
	6	0.40	0.2958	−0.0006979	0.125	13	5,965.65
	7	0.40	–	–	–	3	7,081.38[a]
04	3	1.00	3.2294	−0.0002624	0.958	13	4.07
	4	0.30	0.4874	−0.0007642	0.192	3	3.42
	5	0.20	0.2660	−0.0002292	0.172	13	136.22
	6	0.20	0.1819[a]	−0.0010273[a]	–	19	7,808.39[a]
05	3	1.00	1.4856	−0.0007117	0.301	42	30.76
	4	0.40	0.5659	−0.0004862	0.106	13	28.47
	5	0.40	0.3583	−0.0002181	0.102	13	412.04
	6	0.40	0.1849	−0.0007009	0.049	9	1,650.26
	7	0.40	0.1395[a]	−0.0041407[a]	–	6	6,894.49[a]
06	3	5.00	8.4034	−0.0004046	3.959	28	11.81
	4	4.50	4.5613	0.0000000	4.369	63	1,035.67
	5	4.50	3.1492[a]	−0.0027268[a]	–	42	9,040.39[a]
07	3	30.00	17.0289	−0.0005812	7.490	94	87.63
	4	10.00	11.9770	−0.0002707	9.569	42	846.74
	5	4.00	4.8733	−0.0003621	3.603	28	5,184.76
	6	4.00	2.7909[a]	−0.0053520[a]	–	3	5,223.90[a]
08	3	1.00	1.3130[b]	−0.0110870[b]	–	141	562.27[b]
	4	1.00	0.4476	0.0000000	0.785	63	6,338.37
	5	1.00	0.0626	−0.0006100	0.237	13	622.74
	6	1.00	0.0376[a]	−0.0016832[a]	–	28	6,845.49[a]
09	3	1.00	1.9998[b]	−0.0077818[b]	–	141	1,262.31[b]
	4	1.00	1.3293[a]	−0.0862732[a]	–	42	7,569.87[a]
10	3	4.00	10.2380[b]	−0.0069331[b]	–	141	4,376.43[b]
	4	4.00	8.6188[a]	−0.1971054[a]	–	19	9,491.13[a]

[a]Out of time (time limit per model is 3,600 s)

[b]Model size exceeds license limits

The computational results for area-tight (δ, B)-underestimators are provided in Table 3. The concavity of function 02 makes it possible to compute area-tight (δ, B)-underestimators for up to 15 breakpoints within the time limit. Functions 08 and 09 are difficult to tightly underestimate: the value of M needs to be chosen carefully; local solvers might easily miss a global optimum for (20)–(24).

Results for area-tight (δ, B)-tubes for the ten test functions are given in Table 4. The column labeled "$\underline{A}^+ + \underline{A}^-$" reports on the sum of the area of the corresponding area-tight (δ, B)-overestimator and area-tight (δ, B)-underestimator, which is a lower bound on the area of a (δ, B)-tube. Further, $\mu^{\pm} := \max\{\mu^+, \mu^-\}$ provides the maximal absolute vertical deviation of the tube to the original function $f(x)$. Interestingly, the area of an area-tight (δ, B)-tube is only marginally larger (if at all), for the tested functions, compared to the area provided by combining an area-tight (δ, B)-overestimator with an area-tight (δ, B)-underestimator, while the number of

Table 3 Area-tight (δ, B)-underestimators for the functions provided in Table 1

#	B	δ	$\underline{A}^-$	ψ^-	μ^-	I	Sec.
01	3	3.10	7.1458	−0.0001151	3.100	3	0.73
	4	1.50	3.1759	−0.0000429	1.376	9	41.06
	5	1.10	1.7801[a]	−0.0009104[a]	–	28	4,542.38[a]
02	3	1.00	5.9903	0.0000000	0.564	2	0.33
	4	0.85	2.6130	0.0000000	0.319	2	2.03
	5	0.45	1.4598	0.0000000	0.205	2	28.95
	6	0.45	0.9312	0.0000000	0.143	2	5.80
	7	0.25	0.6455	0.0000000	0.106	2	51.80
	8	0.25	0.4738	0.0000000	0.081	2	61.68
	9	0.25	0.3625	0.0000000	0.064	2	44.42
	10	0.25	0.2863	0.0000000	0.052	2	67.59
	11	0.25	0.2318	0.0000000	0.043	2	8.77
	12	0.25	0.1915	0.0000000	0.036	2	299.95
	13	0.25	0.1609	0.0000000	0.031	2	380.18
	14	0.25	0.1371	0.0000000	0.027	2	858.85
	15	0.25	0.1182	0.0000000	0.023	2	526.15
	16	0.25	–	–	–	2	3,601.02[a]
03	3	1.50	3.4820	−0.0005656	1.365	28	13.46
	4	0.40	0.7448	−0.0002769	0.278	28	62.06
	5	0.40	0.4484	−0.0004956	0.311	28	1,118.99
	6	0.40	0.2958[a]	−0.0027497[a]	–	13	7,059.87[a]
04	3	1.00	3.2294	−0.0002941	0.958	13	3.03
	4	0.30	0.4874	−0.0007642	0.192	3	3.06
	5	0.20	0.2661	−0.0000696	0.180	19	202.68
	6	0.20	0.1819[a]	−0.0010272[a]	–	19	6,774.00[a]

(continued)

Table 3 (continued)

#	B	δ	$\underline{A}^-$	ψ^-	μ^-	I	Sec.
05	3	1.00	1.0176	−0.0006447	0.285	9	2.71
	4	0.40	0.3514	−0.0008220	0.157	13	56.77
	5	0.40	0.2615	−0.0002037	0.150	19	1,854.52
	6	0.40	–	–	–	3	6,561.11[a]
06	3	5.00	7.1298	−0.0005952	3.779	3	1.39
	4	4.50	4.0965	−0.0007573	4.351	63	1,319.01
	5	4.50	2.0713[a]	−0.0048858[a]	–	28	7,504.65[a]
07	3	30.00	20.1332[b]	−0.0085689[b]	–	141	196.96[b]
	4	10.00	6.3694[a]	−0.0016387[a]	–	94	6,880.37[a]
08	3	1.00	0.1772	0.0000000	1.000	3	0.86
	4	1.00	0.1764	−0.0009344	0.997	4	67.11
	5	1.00	0.0205	0.0000000	0.108	6	425.71
	6	1.00	0.0142	−0.0000999	0.106	4	354.25
	7	1.00	0.0142	−0.0000999	0.106	4	731.74
	8	1.00	0.0109[a]	−0.0066468[a]	–	9	6,050.35[a]
09	3	1.00	0.3598	0.0000000	1.030	6	4.81
	4	1.00	0.3597	−0.0001966	1.030	9	646.77
	5	1.00	0.1984	0.0000000	1.000	6	1,832.71
	6	1.00	0.1966	−0.0004934	1.030	4	2,752.09
	7	1.00	–	–	–	2	3,601.98[a]
10	3	4.00	7.9921[b]	−0.0017494[b]	–	141	5,844.36[b]
	4	4.00	6.218[a]	−0.2982953[a]	–	13	5,864.30[a]

[a]Out of time (time limit per model is 3,600 s)

[b]Model size exceeds license limits

breakpoints for the area-tight (δ, B)-tubes is almost half compared to the combination of an area-tight (δ, B)-overestimator with an area-tight (δ, B)-underestimator. Computing area-tight (δ, B)-tubes is computationally more challenging than computing area-tight (δ, B)-overestimators and area-tight (δ, B)-underestimators. However, it remains computational tractable to compute area-tight (δ, B)-tubes for a small number of breakpoints.

Figure 1 shows plots of the ten test functions together with an area-tight (δ, B)-overestimator, (δ, B)-underestimator or (δ, B)-tube. The presented over-, underestimators and tubes correspond to the results of Tables 2, 3 and 4.

Table 4 Area-tight (δ, B)-tubes for the functions provided in Table 1

#	B	δ	$\underline{A}^+ + \underline{A}^-$	$\underline{A}^\pm$	ψ^+	ψ^-	$\mu^\pm$	I	Sec.
01	3	3.10	21.4375	21.4375	0.0000000	−0.0001148	3.100	3	1.40
	4	1.50	9.5278	9.5278	0.0000000	−0.0000105	1.369	9	178.39
	5	1.10	[c]	5.3594[a]	0.0000000[a]	−0.0156250[a]	–	9	3,707.36[a]
02	3	1.00	8.4089	8.4292	−0.0008616	0.0000000	0.788	13	19.56
	4	0.85	3.7910	3.7946	−0.0004650	0.0000000	0.449	9	257.74
	5	0.45	[c]	2.1479	−0.0003956	0.0000000	0.282	9	437.53
	6	0.45	[c]	1.3279[a]	−0.0016468[a]	0.0000000[a]	–	4	3833.89[a]
03	3	1.50	6.9639	7.3622	−0.0000396	−0.0004090	1.500	42	91.53
	4	0.40	1.4896	1.5018	−0.0006088	−0.0006088	0.257	19	141.01
	5	0.40	0.8967	1.0616[a]	−0.0020084[a]	−0.0088723[a]	–	9	4,516.01[a]
04	3	1.00	6.4588	7.9908	−0.0006473	−0.0002773	1.000	42	85.96
	4	0.30	0.9748	0.9967	−0.0006409	−0.0006409	0.154	6	42.62
	5	0.20	0.5321	0.7070	−0.0002143	−0.0006732	0.174	13	1,858.18
	6	0.20	[c]	–	–	–	–	2	3,600.11[a]
05	3	1.00	2.5032	2.6914	−0.0006133	−0.0003608	0.453	42	70.57
	4	0.40	0.9173	0.9235	−0.0006387	0.0000000	0.157	13	115.89
	5	0.40	0.6198	0.6192[a]	−0.0028052[a]	−0.0007355[a]	–	13	5,500.06[a]
06	3	5.00	15.5331	15.6470	−0.0007989	−0.0008506	4.466	63	131.09
	4	4.50	8.6578	10.2935[a]	0.0000000[a]	−0.0039028[a]	–	63	5,896.67[a]
07	3	30.00	[c]	37.492[b]	−0.0006538[b]	−0.0073744[b]	–	141	494.11[b]
	4	10.00	[c]	19.2815	−0.0003845	−0.0009017	10.000	42	3,810.09
	5	4.00	[c]	8.518[a]	−0.0241269[a]	−0.0248196[a]	–	13	9,022.88[a]
08	3	1.00	[c]	1.4903[b]	−0.0110513[b]	0.0000000[b]	–	141	3,660.14[b]
	4	1.00	0.6249	0.6221[a]	−0.1077110[a]	0.0000000[a]	–	42	3,889.13[a]
09	3	1.00	[c]	2.3596[a]	−0.0152941[a]	0.0000000[a]	–	94	5,729.48[a]
	4	1.00	[c]	1.7519[a]	−0.0862732[a]	0.0000000[a]	–	42	7,235.23[a]
10	3	4.00	[c]	18.6457[a]	−0.0125147[a]	−0.0071669[a]	–	94	8,319.97[a]
	4	4.00	[c]	13.1937[a]	−1.0792070[a]	−0.0033989[a]	–	9	4,397.01[a]

[a]Out of time (time limit per model is 3,600 s)

[b]Model size exceeds license limits

[c]Over- and/or underestimator problem was not solved to global optimality

7 Conclusions

In this paper, we extend the literature on methodologies which automatically compute optimal piecewise linear overestimators, underestimators and tubes for univariate functions. The computed approximators are optimal among all piecewise

linear, continuous functions in the sense that they minimize the area between the function and the approximator. Our methodology for computing area-tight (δ, B)-overestimators, (δ, B)-underestimators and (δ, B)-tubes require the solution of a series of continuous, non-linear, and non-convex mathematical programming problems.

The computational tests reveal that it is worth-while to compute area-tight (δ, B)-tubes which share the same breakpoint system, rather than computing (δ, B)-overestimators and (δ, B)-underestimators individually, if tubes are desired.

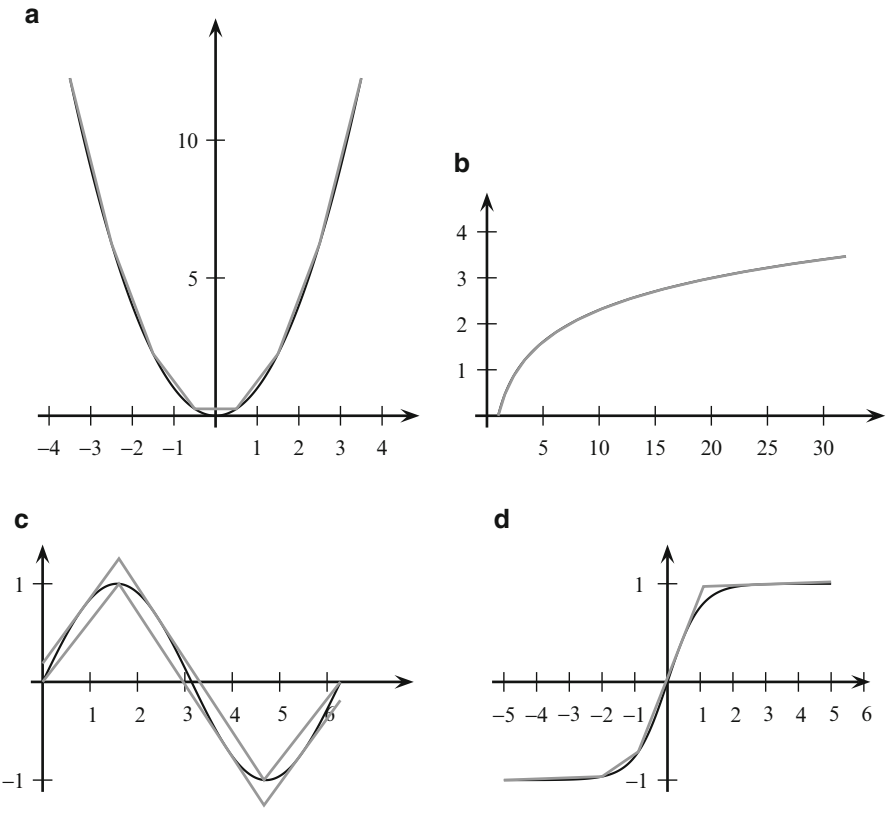


Fig. 1 The ten univariate functions together with computed (δ, B)-overestimator, (δ, B)-underestimator, or (δ, B)-tube; (*black lines*) original function $f(x)$, (*gray lines*) approximator function $\ell^+(x)$, $\ell^-(x)$, or $\ell^\pm(x)$. (**a**) 01: area-tight (0.4,8)-overestimator, (**b**) 02: area-tight (0.25,15)-underestimator, (**c**) 03: area-tight (0.4,4)-tube, (**d**) 04: area-tight (0.2,5)-overestimator, (**e**) 05: area-tight (0.4,4)-underestimator, (**f**) 06: area-tight (5,3)-tube, (**g**) 07: area-tight (10,4)-tube, (**h**) 08: area-tight (1,5)-overestimator, (**i**) 09: area-tight (1,5)-underestimator, (**j**) 10: no area-tight approximation

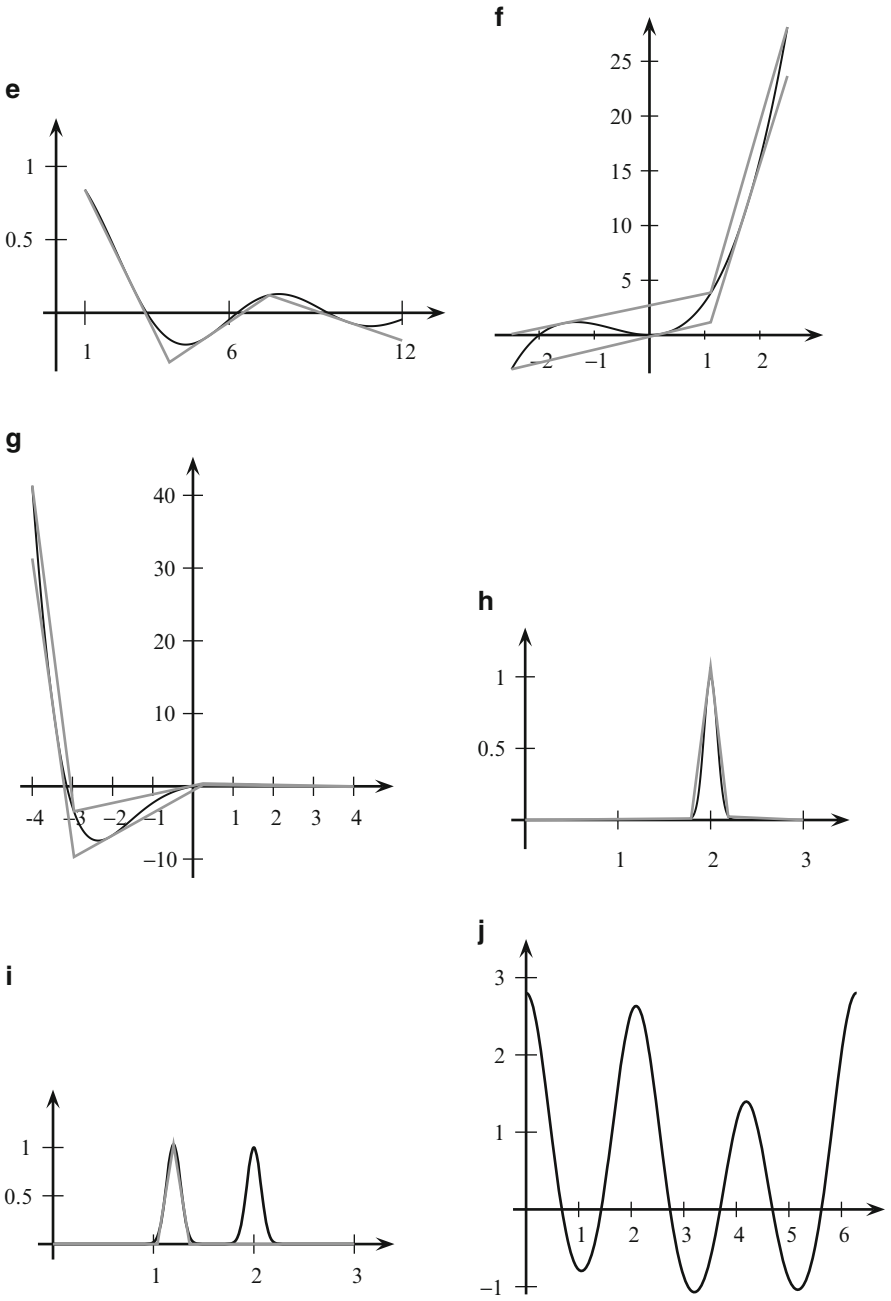


Fig. 1 (continued)

References

1. Geißler, B.: Towards Globally Optimal Solutions for MINLPs by Discretization Techniques with Applications in Gas Network Optimization. Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen-Nürnberg, Germany (2011)
2. Geißler, B., Martin, A., Morsi, A., Schewe, L.: Using piecewise linear functions for solving MINLPs. In: Lee, J., Leyffer, S. (eds.) Mixed Integer Nonlinear Programming. IMA Volumes in Mathematics and its Applications, vol. 154, pp. 287–314. Springer Science+Business Media, LLC (2012)
3. Geyer, A., Hanke, M., Weissensteiner, A.: Life-cycle asset allocation and consumption separable using stochastic linear programming. J. Comput. Finance **12**, 29–50 (2009)
4. Hettich, R., Kortanek, K.O.: Semi-infinite programming. SIAM Rev. **35**, 380–429 (1993)
5. Lopez, M., Still, G.: Semi-infinite programming. Eur. J. Oper. Res. **180**, 491–518 (2007)
6. Maranas, C., Floudas, C.A.: Global minimum potential energy conformations of small molecules. J. Glob. Optim. **4**, 135–170 (1994)
7. Misener, R., Floudas, C.A.: Piecewise-linear approximations of multidimensional functions. J. Optim. Theory Appl. **145**, 120–147 (2010)
8. Misener, R., Floudas, C.A.: Global optimization of mixed-integer quadratically-constrained quadratic programs (MIQCQP) through piecewise-linear and edge-concave relaxations. Math. Program. Ser. B **136**, 155–182 (2012)
9. Misener, R., Floudas, C.A.: GloMIQO: global mixed-integer quadratic optimizer. J. Glob. Optim. (2012, accepted for publication). doi:10.1007/s10898-012-9874-7
10. Pardalos, P.M., Rosen, J.B.: Constrained Global Optimization: Algorithms and Applications. Lecture Notes in Computer Science. Springer, Berlin (1987)
11. Rebennack, S., Kallrath, J.: Continuous Piecewise Linear δ-Approximations for MINLP Problems. I. Minimal Breakpoint Systems for Univariate Functions. CSM Working Paper 2012–12 (2012)
12. Rebennack, S., Kallrath, J.: Continuous Piecewise Linear δ-Approximations for MINLP Problems. II. Bivariate and Multivariate Functions. CSM Working Paper 2012–13 (2012)
13. Rosen, J.B., Pardalos, P.M.: Global minimization of large-scale constrained concave quadratic problems by separable programming. Math. Program. **34**, 163–174 (1986)
14. Schrage, L.: LindoSystems: LindoAPI (2004)
15. Vielma, J.P., Nemhauser, G.: Modeling disjunctive constraints with a logarithmic number of binary variables and constraints. Math. Program. **128**, 49–72 (2011)

Market Graph and Markowitz Model

Valery Kalyagin, Alexander Koldanov, Petr Koldanov, and Viktor Zamaraev

1 Introduction

Market graph is an important part of market network. The concept of the market graph was introduced in [3, 5]. Since, different aspects of the market graph approach (threshold method) were developed in the literature. We mention here some of the references. Dynamics of the US market graphs was studied in [6]. Complexity of the US market graph associated with significant correlations is investigated in [7]. Peculiarity of different financial markets is emphasized in [2, 8, 11, 19, 24]. Market graphs with different measures of similarity were studied in [1, 2, 10, 22]. Statistical procedures for the market graph construction are discussed in [15, 16]. Some efficient algorithms related to the calculation of isolated cliques in a market graph are presented in [9, 12]. The power law phenomenon first observed for US stock market in [5] was then developed in [4, 11, 23].

Markowitz model is the most popular tool for practical portfolio selection and optimization [21]. The main concept of portfolio optimization in the framework of Markowitz model is the efficient frontier of sets of stocks. For a given set of stocks its efficient frontier is the curve in the plane associated with Pareto optimal portfolios according to two criteria: $expected\ return \rightarrow max$, $risk \rightarrow min$. The choice of particular portfolio on the efficient frontier is then determined by the value of risk aversion of investor. However in practice investor is interested to limit the number of stocks in his optimal portfolio. We call it stocks selection problem. Criteria of selection can be different. It can be the stock return, i.e. one selects the stocks with the highest return, it can be the stock volume of trading, i.e. one selects the stocks

V. Kalyagin • A. Koldanov • P. Koldanov • V. Zamaraev (✉)
Laboratory of Algorithms and Technologies for Network Analysis and Department of Applied Mathematics and Informatics, National Research University Higher School of Economics, 136, Rodionova Str., Nizhny Novgorod 603093, Russian Federation
e-mail: vkalyagin@hse.ru; viktor.zamaraev@gmail.com

Partly supported by Russian Federation Government grant N. 11.G34.31.0057.

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_15,

with the highest volume, it can be the stock liquidity, i.e. one selects the stocks with the highest liquidity or other criteria. One needs to make this first selection without big loss of information on efficient portfolios.

In the present paper we investigate a connection between characteristics of the market graph and classical Markowitz portfolio theory. More precisely we consider cliques and independent sets of the market graph. Cliques are sets of highly interconnected stocks and usually are composed by stocks attractive by their return and liquidity [23, 24]. Independent sets are sets of stocks without connections in the market graphs. Independent sets were conjectured in [5] to be useful for the construction of diversified portfolio. Our main result is the following: effective frontier of the market can be well approximated by the effective frontier of the maximum independent set (MIS) of the market graph constructed on the sets of stocks with the highest Sharp ratio. This allows to reduce the number of stocks for portfolio optimization without the loss of quality of obtained portfolio. On the other hand we show that despite some attractiveness the cliques are not suitable for the portfolio optimization. Note that some relations of market network analysis with portfolio theory were already mentioned in [13, 17, 20].

The paper is organized as follows. In Sect. 2 we recall some notions related to market graph and Markowitz theory. In Sect. 3 we discuss some statistical procedures for the stock selection problem. In Sect. 4 we study efficient frontiers of independent sets and cliques for different market graphs and different stock markets. Finally in Sect. 5 we give some comments for the obtained results.

2 Market Graph and Markowitz Theory

Let S be a subset of stocks on financial market, N be the number of stocks in S, and let n be the number of observations. Denote by $p_i(t)$the price of the stock i for the day t, $(i = 1,\dots,N;\ t = 1,\dots,n)$ and define the daily return of the stock i for the period from day $t-1$ to day t as $r_i(t) = \ln(p_i(t)/p_i(t-1))$. We assume $r_i(t)$ to be a realization of the random variable $R_i(t)$. We consider standard assumptions: the random variables $R_i(t), t = 1,\dots,n$are independent with fixed i, have all the same distribution as a random variable $R_i(i = 1,\dots,N)$, and the random vector $(R_1, R_2, \dots, R_N)$has a multivariate distribution with the covariance matrix $\|\sigma_{i,j}\|$. Let

$$\rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

where $\sigma_i^2 = \sigma_{i,i}$, $\sigma_j^2 = \sigma_{j,j}$ Matrix of correlations $\|\rho_{i,j}\|$is the matrix for market graph construction. Each node of the graph corresponds to a stock from S. The edge between two nodes i and j is included in the market graph, if $\rho_{i,j} > \rho_0$(where $\rho_0$is a threshold). Clique in a graph is a subset of nodes connected to each other. Maximum clique (MC) is the clique with the maximal number of nodes. Independent set in a graph is a subset of nodes with no connections. MIS is the independent set

with maximal number of nodes. Cliques are sets of highly interconnected stocks. Independent sets are sets of stocks without connections in the market graphs. As it was mentioned above independent sets were conjectured in [5] to be suitable for the construction of diversified portfolio.

Portfolio of stocks from S is defined by the vector $f = (f_1, f_2, \ldots, f_N)$, where $f_i \geq 0$ is the portion of capital invested in the stock i, $i = 1, 2, \ldots, N$ and $\sum_{i=1}^{N} f_i = 1$. Return of the portfolio f is a random variable $R = \sum_{i=1}^{N} f_i R_i$. Mean-variance theory of Markowitz is based on two characteristics: expected return $E(R)$

$$E(R) = \sum_{i=1}^{N} f_i E(R_i)$$

and risk $\sigma(R)$

$$\sigma^2(R) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{i,j} f_i f_j$$

Efficient frontier of the market is the set of Pareto optimal points in the plane (E, σ) with respect to two criteria

$$E(R) \to \max, \sigma(R) \to \min$$

Investor according to his preferences (utility function, risk aversion, or others) can choose an efficient portfolio associated with a point of the efficient frontier. Efficient frontier of any set of nodes is defined in the same way.

3 Stocks Selection Problem

In this section we discuss the stock selection problem from statistical point of view. Our approach follows the paper [18]. For the set of stocks $S = \{1, \ldots, N\}$ we would like to select a subset according to some criteria. Let $x_i(t)$ be the observation of some characteristic of stock i (return, volume of trading, liquidity or other) for the time t, $t = 1, \ldots, n$, $i = 1, \ldots, N$. We assume $x_i(t)$ to be a realization of the random variable $X_i(t)$. We consider standard assumptions: the random variables $X_i(t), t = 1, \ldots, n$are independent with fixed i, have all the same distribution as a random variable $X_i (i = 1, \ldots, N)$. We assume X_i to be a random variable of the class $N(a_i, \sigma_i^2)$. Let us consider the following selection criteria according to the quality of stocks:

1. the quality of the i-th stock is characterized by parameter a_i, and a stock is said to be *positive* (or good) if $a_i > a_0$, and is said to be *negative* (or bad) if $a_i \leq a_0$.
2. the quality of the i-th stock is characterized by parameter σ_i, and a stock is said to be *positive* (or good) if $\sigma_i < \sigma_0$, and is said to be *negative* (or bad) if $\sigma_i \geq \sigma_0$.

3. the quality of the i-th stock is characterized by parameter $sh_i = \dfrac{a_i}{\sigma_i}$, and a stock is said to be *positive* (or good) if $sh_i > sh_0$, and is said to be *negative* (or bad) if $sh_i \leq sh_0$.

For the case of return the criteria 1 gives the selection of the most profitable stocks, the criteria 2 gives the selection of the least risky stocks, and the criteria 3 gives the selection of the best stocks according to the Sharp ratio. For the case of liquidity the criteria 1 gives the most liquid stocks. All formulated selection problems can be considered as multiple hypothesis testing problems. For each statistical procedure of stock selection there are two possible sources of error. There is the possibility of *false positives*, that is, stocks which are selected although they are negative, and of *false negatives*, that is, populations which are not selected although they are positive. Instead of false negatives we shall focus our attention on *true positives*, that is, on those positive stocks which are included in the selected group.

For measuring how well a statistical procedure carries out its task of identifying the positive stocks we consider:

(a) The expected number of true positives.
(b) The expected proportion of true positives, that is, the quantity (a) divided by the total number of positives.

For measuring how well a procedure carries out its task of identifying the negative stocks we consider:

(c) The expected number of false positives.
(d) The expected proportion of false positives, that is, the quantity (c) divided by the total number of negatives.

As a generic notation for any one of the quantities (a), (b) we shall use $S(\theta, \delta)$ where θ is an element of the parametric space Ω, and δ is the statistical procedure. Similarly, we shall let $R(\theta, \delta)$ denote the quantity (c) or (d). With these definitions of R and S, it is desirable to have $S(\theta, \delta)$ as large and $R(\theta, \delta)$ as small as possible.

A selection procedure is a partition of the sample space into the sets $D_{i_1,\ldots,i_k}$ of those sample points for which the selected group consists of the stocks with subscripts $i_1, \ldots, i_k$ and no others. To these must be added the set D_0 for which none of the stocks is selected. If the number of stocks is N, the number of sets D is 2^N. Let E_i be the set of sample points for which the i-th stock is included in the selected group. Then each of the two systems of sets $\{D\}$ and $\{E\}$ is uniquely expressed in terms of the other. In fact, E_i is the union of all those sets D which have i as one of their subscripts. Conversely,

$$D_{i_1,\ldots,i_k} = E_{i_1} \cap \ldots \cap E_{i_k} \cap \overline{E_{j_1}} \ldots \cap \overline{E_{j_{N-k}}}$$

where $j_1, \ldots, j_{N-k}$ are the subscripts different from $i_1, \ldots, i_k$ and $\overline{E}$ denotes the complement of E. Each E_i is then represented by its characteristic function $\psi_i(x)$. Then selection statistical procedure is characterized by the vector $\psi = (\psi_1, \ldots, \psi_N)$.

In the case of independent random variables (independent returns, volume or liquidity) it is shown in [18] that there exists a statistical procedure of the type

$$\psi_i = \begin{cases} 1, \ T_i \geq c_i \\ 0, \ T_i < c_i \end{cases} \tag{1}$$

which is optimal in the following min–max sense:

$$\text{subject to } \inf_{\theta} S(\theta, \delta) \geq \gamma \quad \text{one has } \sup_{\theta} R(\theta, \delta) \to \min \tag{2}$$

In particular if quality of stocks is characterized by the parameter a then optimal statistical procedure for the stock selection is given by:

$$\psi_i = \begin{cases} 1, \ \overline{x_i} \geq c_i \\ 0, \ \overline{x_i} < c_i \end{cases} \tag{3}$$

where $\overline{x_i} = \dfrac{1}{n} \sum\limits_{t=1}^{n} x_i(t)$.

If the quality of stocks is characterized by the parameter σ, then the optimal statistical procedure for the stock selection is given by:

$$\psi_i = \begin{cases} 0, \ s_i^2 \geq c_i \\ 1, \ s_i^2 < c_i \end{cases} \tag{4}$$

where $s_i^2 = \dfrac{1}{n} \sum\limits_{t=1}^{n} (x_i(t) - \overline{x}_i)^2$

If the quality of stocks is characterized by the parameter $\dfrac{a_i}{\sigma_i}$, then the optimal statistical procedure for the stock selection is given by:

$$\psi_i = \begin{cases} 1, \ \dfrac{\overline{x_i}}{s_i^2} \geq c_i \\ 0, \ \dfrac{\overline{x_i}}{s_i^2} < c_i \end{cases} \tag{5}$$

The assumption of independence of random variables X_i (returns, volume, liquidity) is not realistic for the financial market. Therefore it is important to construct optimal statistical procedures for the stock selection problem for more general cases. The first result in this direction is obtained in [14] where it is shown that the statistical procedure (3) remains optimal in the sense of multiple hypothesis testing for multivariate normal distributions. In what follows we use the tests (3), (4), (5) for the first stage of the stock selection.

4 Efficient Frontiers of Independent Sets and Cliques

In our study of market graphs from the point of view of the portfolio theory we use a two-stage procedure. At the first stage we select a "good" stocks by fixing some criteria and a critical value of the "goodness." At the second stage we construct the market graphs with the selected stocks as nodes with different thresholds of interconnections. We calculate maximum cliques and MISs of the constructed market graphs and study the efficient frontiers of these sets of stocks. Our main goal is to find a small sets of "good" stocks such that the efficient frontiers associated with this sets of stocks are close to the efficient frontier of the entire market. Such sets of stocks can be a basis of the construction of "diversified" portfolios. Our main finding is the following (empirical) conclusion: independent sets are suitable for portfolio optimization in the case when at the first stage one selects the stocks with the highest Sharp ratio. The selection of stocks with highest returns, lowest risk or highest liquidity does not give independent sets with this property. This phenomenon is new and needs a deeper investigation. The situation for the cliques is different. In many markets cliques represent a sets of stocks which dominate market and therefore are attractive for investment. However as it is shown below cliques are not appropriate for portfolio optimization. This result is in some sense expected by the well known principle "don't put all eggs in the same basket" but what is interesting is the fact that the stocks in the maximum cliques produce a very limited efficient frontier in both directions: expected return and risk. Moreover efficient frontiers of the maximum cliques are generated by a very small number of stocks in the clique.

We use data from two stock markets: Nasdaq market (USA), daily returns for the period November 2011 to October 2013 and Moscow interbank currency exchange MICEX (Russian Federation), daily returns for the period October 2008 to October 2010. To make the conclusions more general we take at random 250 stocks from Nasdaq market and 151 stocks from MICEX market. Next we apply the selection procedure according to some criteria, construct the market graphs for the three values of thresholds 0.1, 0.3, 0.4, calculate cliques and independent sets, and compare the efficient frontiers of obtained sets with the efficient frontier of 250 stocks for Nasdaq and 151 stocks for MICEX markets. The results are stable with respect to random choice, similar for both markets and are presented in Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Each figure has some interesting meaning which is described below.

Figure 1 shows the efficient frontiers for three sets of stocks for US market: randomly selected 250 stocks (thick line), MIS (66 stocks) for the market graph constructed on the set of 125 highest Sharp ratio stocks for threshold 0.3 (dashed line), MIS (90 stocks) for the market graph constructed on the set of 125 highest Sharp ratio stocks for threshold 0.4 (thin line). It is clear that efficient frontier of the market (250 stocks) is well approximated by the efficient frontiers of the independent sets (66 and 90 stocks).

The conclusion for Fig. 1 is confirmed in Fig. 2 where only 16 stocks are selected according to the Sharp ratio. Despite a small number of stocks in independent sets (11 and 13 stocks) the approximation of the efficient frontier is still good.

Figure 3 shows the efficient frontiers for three sets of stocks for RF market: randomly selected 151 stocks (thick line), MIS (29 stocks) for the market graph constructed on the set of 76 highest Sharp ratio stocks for threshold 0.3 (dashed line), MIS (40 stocks) for the market graph constructed on the set of 76 highest Sharp ratio stocks for threshold 0.4 (thin line). One can see that the behavior of two market is different but the phenomenon of good approximation is the same.

Figure 4 shows the stocks of MIS in the plane (E, σ) for US (17 stocks) market for the market graphs with threshold 0.1. The stocks in the independent set are enumerated according to their Sharp ratio. One can observe that in fact the efficient frontiers of independent set are constructed only using the stocks number 1, 2, 3, 9, 10. The same conclusion is valid for the values of threshold 0.3, 0.4 (we use here the threshold 0.1 for simplicity of presentation).

The same observation is valid for RF market. Figure 5 shows the stocks of MIS in the plane (E, σ) for RF (eight stocks) market for the market graphs with threshold 0.1. The stocks in the independent sets are enumerated according to their Sharp ratio. One can observe that in fact as for US market the efficient frontiers of independent set are constructed only using the stocks number 1, 2, 4, 6. The same conclusion is valid for the values of threshold 0.3, 0.4.

The phenomenon of good approximation of the efficient frontiers is not observed for another selection criteria. Typical results are Figs. 6 and 7 where the efficient frontiers of independent sets for the selection of the most liquid stocks for the market graph construction are presented.

Our experiments allow to conclude that cliques are not suitable for portfolio optimization. Typical results are given in Figs. 8 and 9. Figure 8 shows the efficient frontiers for three sets of stocks for US market: selected 100 most liquid stocks of the market (thick line), maximum clique (6 stocks) for the market graph constructed on the set of 100 most liquid stocks for threshold 0.5 (dashed line), maximum clique (13 stocks) for the market graph constructed on the set of 100 most liquid stocks for threshold 0.4 (thin line). Figure 9 shows the efficient frontiers for analogous sets of stocks for RF market. Note that for RF market the number of stocks in the cliques is 16 and 21, respectively.

Composition of cliques has some interesting phenomena too. Figure 10 shows the cloud of 21 stocks of the maximum clique of the RF market graph with threshold 0.1 constructed from the set of 100 most liquid stocks. One can observe that in fact only three stocks define the efficient frontier of the maximum clique. This phenomenon is general for the maximum cliques in different situations.

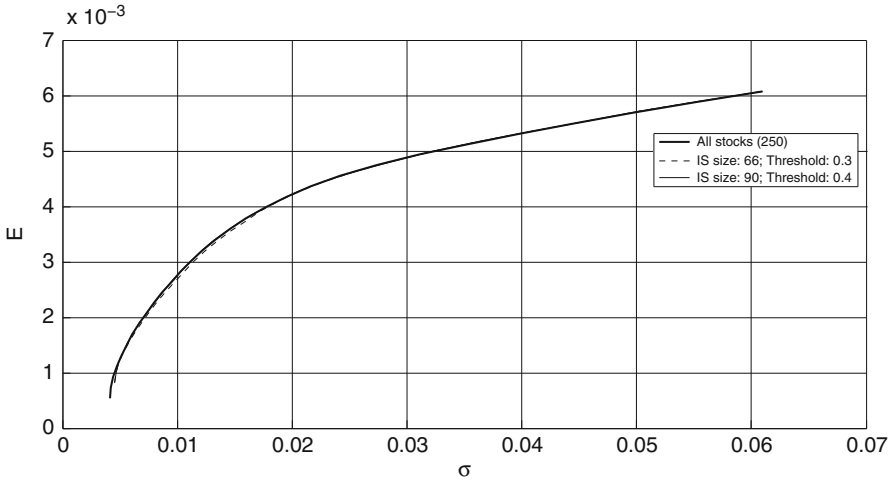


Fig. 1 US market. Efficient frontiers of the maximum independent sets of the market graphs constructed on the set of 125 highest Sharp ratio stocks. Values of threshold 0.3, 0.4

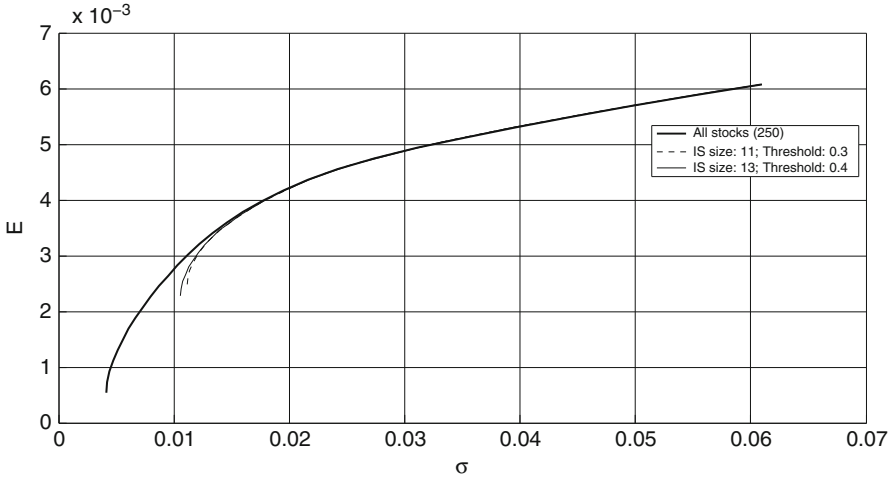


Fig. 2 US market. Efficient frontiers of the maximum independent sets of the market graphs constructed on the set of 16 highest Sharp ratio stocks. Values of threshold 0.3, 0.4

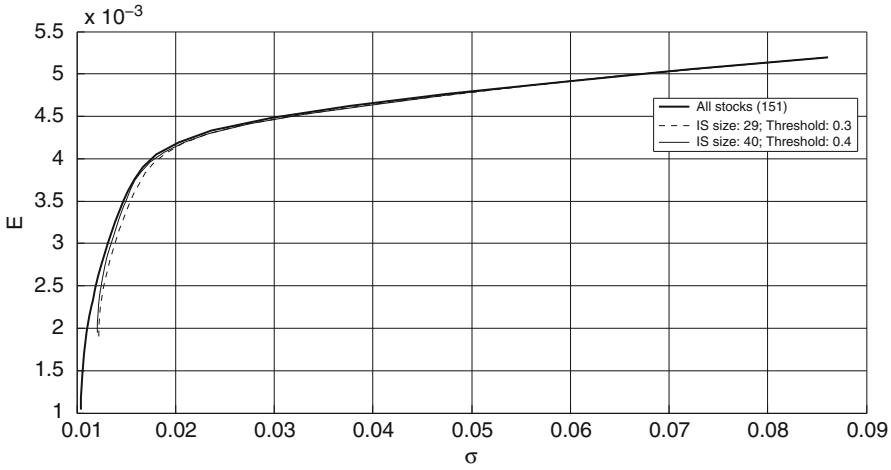


Fig. 3 RF market. Efficient frontiers of the maximum independent sets of the market graphs constructed on the set of 76 highest Sharp ratio stocks. Values of threshold 0.3, 0.4

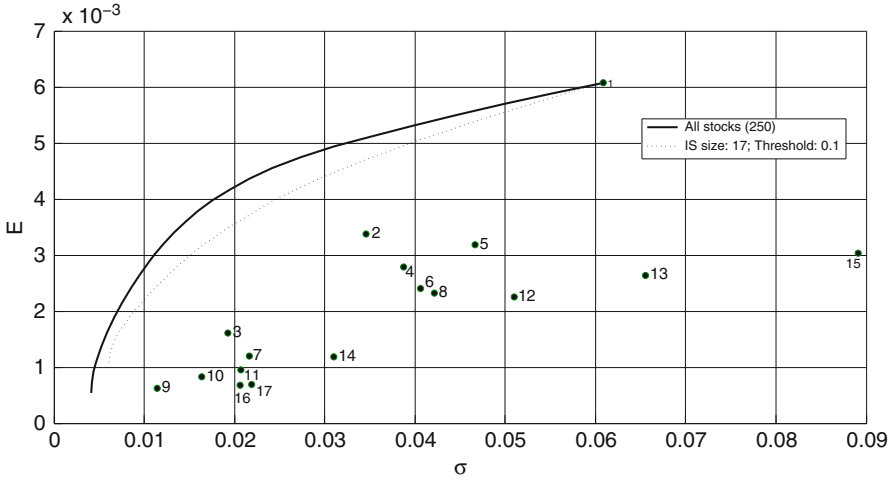


Fig. 4 US market. Efficient frontiers and cloud of the maximum independent set of the market graphs constructed on the set of 125 highest Sharp ratio stocks. Value of threshold 0.1

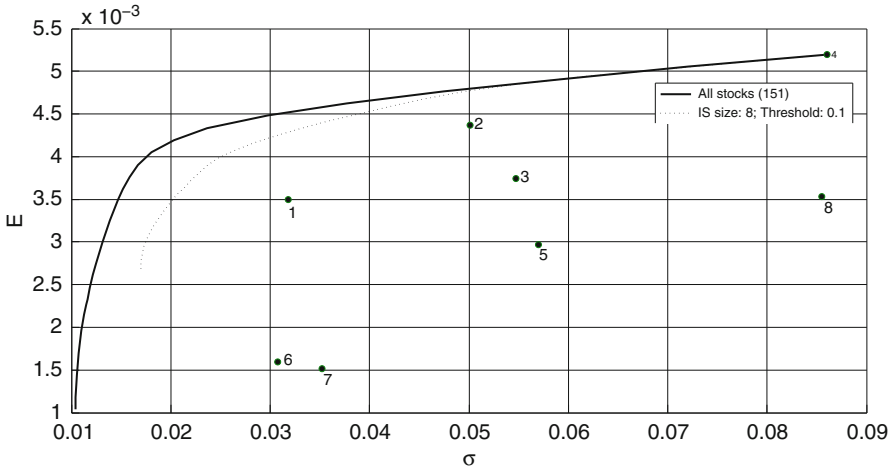


Fig. 5 RF market. Efficient frontiers and cloud of the maximum independent set of the market graphs constructed on the set of 76 highest Sharp ratio stocks. Value of threshold 0.1.

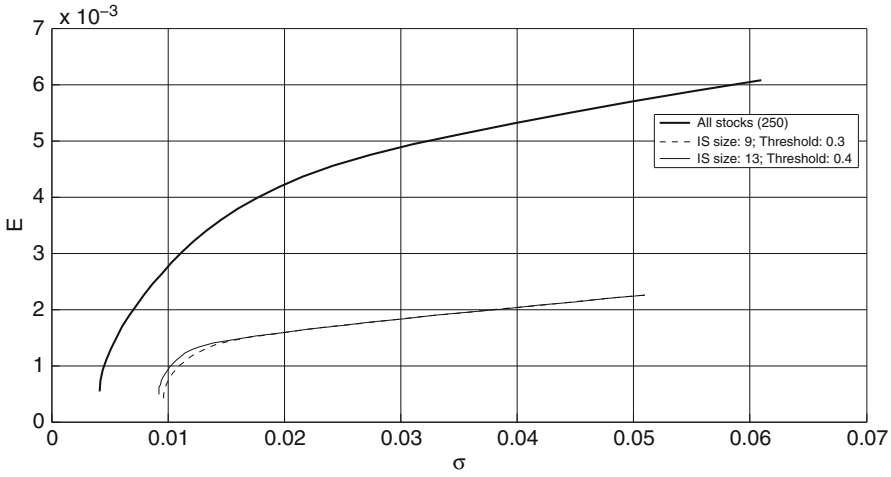


Fig. 6 US market. Efficient frontiers of the maximum independent sets of the market graphs constructed on the set of 16 most liquid stocks. Values of threshold 0.3, 0.4

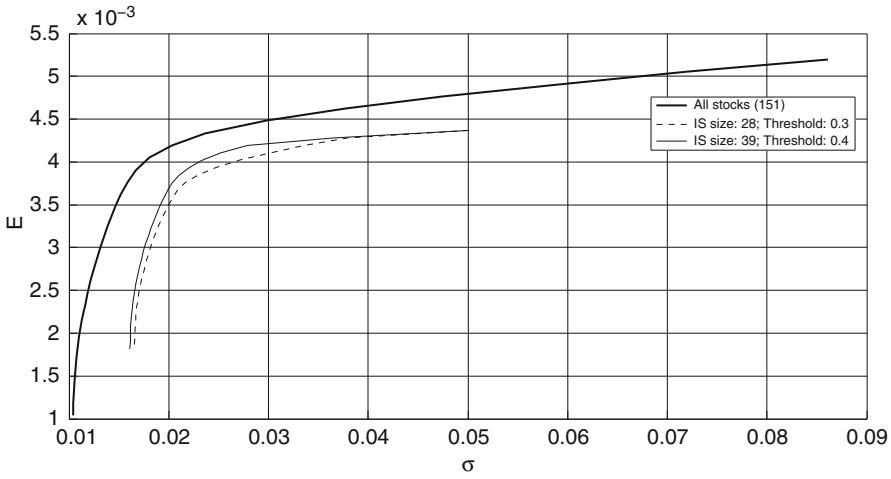


Fig. 7 RF market. Efficient frontiers of the maximum independent sets of the market graphs constructed on the set of 76 most liquid stocks. Values of threshold 0.3, 0.4

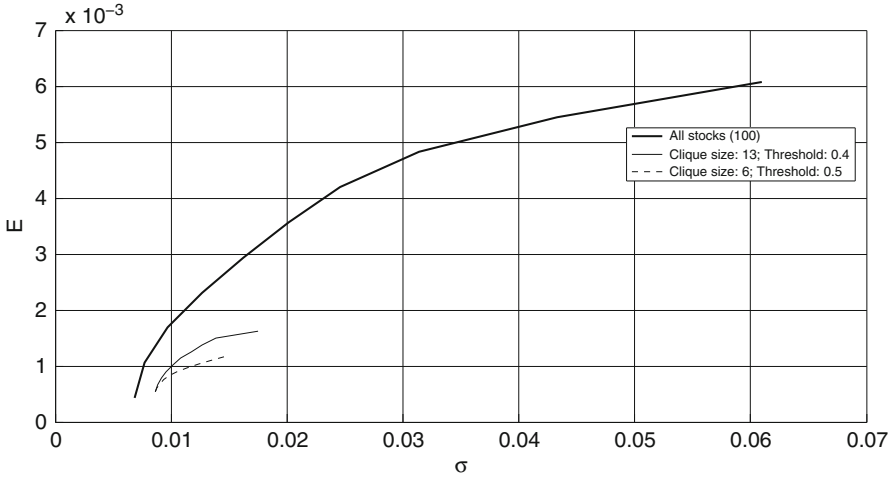


Fig. 8 US market. Efficient frontiers of the maximum cliques of the market graphs constructed on the set of 100 most liquid stocks. Values of threshold 0.4, 0.5

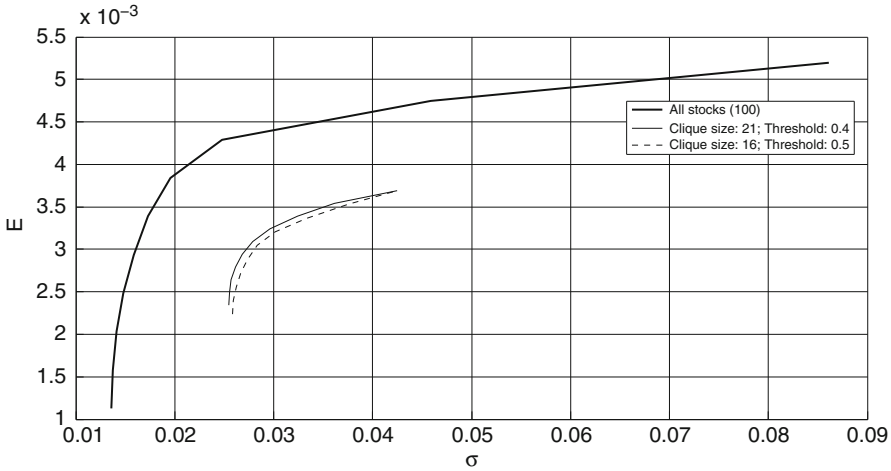


Fig. 9 RF market. Efficient frontiers of the maximum cliques of the market graphs constructed on the set of 100 most liquid stocks. Values of threshold 0.4, 0.5

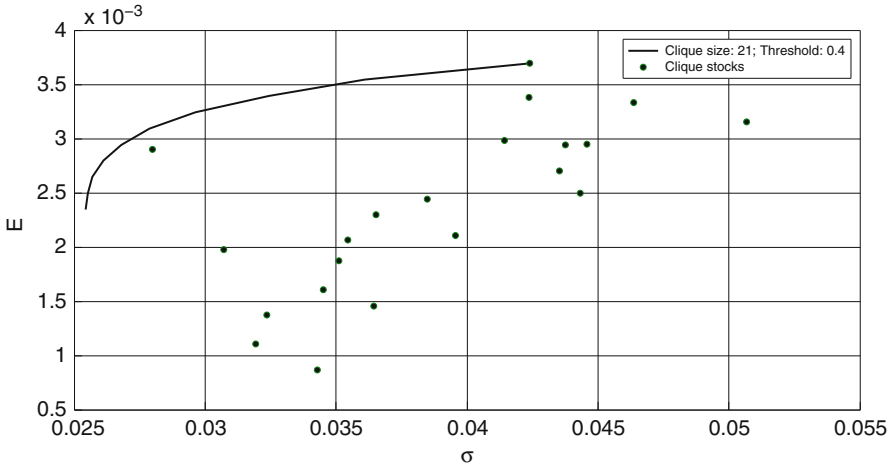


Fig. 10 RF market. Efficient frontier and cloud of the maximum clique of the market graph constructed on the set of 100 most liquid stocks. The value of threshold is 0.4

5 Concluding Remarks

The paper presents some empirical results in the study of connections between market graph approach and Markowitz portfolio theory for financial markets. It is observed that independent sets of the market graphs are suitable for portfolio optimization in the case when at the first stage one selects the stocks with the highest Sharp ratio. The situation for the cliques is different. Despite the fact that in many situations the cliques dominate the market, they are not appropriate for portfolio optimization.

References

1. Bautin, G., Kalyagin, V.A., Koldanov, A.P.: Comparative analysis of two similarity measures for the market graph construction. In: Springer Proceedings in Mathematics and Statistics, vol. 59, pp. 29–41 (2013)
2. Bautin, G.A., Kalyagin, V.A., Koldanov, A.P., Koldanov, P.A., Pardalos, P.M.: Simple measure of similarity for the market graph construction. Comput. Manag. Sci. **10**, 105–124 (2013)
3. Boginski, V., Butenko, S., Pardalos, P.M.: On structural properties of the market graph. In: Nagurney, A. (ed.) Innovations in Financial and Economic Networks, pp. 29–45. Edward Elgar Publishing, Northampton (2003)
4. Boginski, V., Butenko, S., Pardalos, P.M.: Network model of massive data sets. Comput. Sci. Inf. Syst. **1**, 75–89 (2004)
5. Boginski, V., Butenko, S., Pardalos, P.M.: Statistical analysis of financial networks. J. Comput. Stat. Data Anal. **48**(2), 431–443 (2005)
6. Boginski, V., Butenko, S., Pardalos, P.M.: Mining market data: a network approach. J. Comput. Oper. Res. **33**(11), 3171–3184 (2006)
7. Emmert-Streib, F. Dehmer, M.: Identifying critical financial networks of the DJIA: towards a network based index. Complexity **16**(1), 24–33 (2010)
8. Garas, A., Argyrakis, P.: Correlation study of the Athens stock exchange. Physica A **380**, 399–410 (2007)
9. Gunawardena, A.D.A., Meyer, R.R., Dougan, W.L., Monaghan, P.E., Basu, C.: Optimal selection of an independent set of cliques in a market graph. Int. Proc. Econ. Dev. Res. **29**, 281–285 (2012)
10. Hero, A., Rajaratnam, B.: Hub discovery in partial correlation graphs. IEEE Trans. Inf. Theory **58**(9), 6064–6078 (2012)
11. Huang, W.-Q. Zhuang, X.-T. Yao, S.: A network analysis of the Chinese stock market. Physica A **388**, 2956–2964 (2009)
12. Huffner, F., Komusiewicz, C., Moser, H., Niedermeier, R.: Enumerating isolated cliques in synthetic and financial networks. In: Combinatorial Optimization and Applications. Lecture Notes in Computer Science, **5165**, 405–416, Springer (2008)
13. Jones, C.K.: Portfolio size in stochastic portfolio networks using digital portfolio theory. J. Math. Finance **3**, 280–290 (2013)
14. Koldanov, P.A., Bautin, G.A.: Multiple decision problem for stock selection in market network. LNCS (submitted)
15. Koldanov, A.P., Koldanov, P.A: Optimal multiple decision statistical procedure for inverse covariance matrix. Springer Optim. Appl. **87**, 205–216 (2014)
16. Koldanov, A.P., Koldanov, P.A., Kalyagin, V.A., Pardalos, P.M.: Statistical procedures for the market graph construction. Comput. Stat. Data Anal. **68**, 17–29 (2013)

17. Ledoit, O., Wolf, M.: Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J. Empirical Finance **10**, 603–621 (2003)
18. Lehmann, E.L.: Some model I problems of selection. Ann. Math. Stat. **32**(4), 990–1012 (1961)
19. Namaki, A., Shirazi, A.H., Raei, R., Jafari, G.R.: Network analysis of a financial market based on genuine correlation and threshold method. Physica A **390**, 3835–3841 (2011)
20. Onnela, J.-P.: Chakraborti, A., Kaski, K., Kertesz, K., Kanto, A.: Dynamics of market correlations: taxonomy and portfolio analysis. Phys. Rev. E **68**, 56–110 (2003)
21. Panjer, H.H. (ed.): Financial Economics with Applications to Investments, Insurance and Pensions, 2nd edn. The Actuarial Foundation, Schaumburg (2001)
22. Shirokikh, J., Pastukhov, G., Boginski, V., Butenko, S.: Computational study of the US stock market evolution: a rank correlation-based network model. Comput. Manag. Sci. **10**(2–3), 81–103 (2013)
23. Tse C.K., Liu, J., Lau, F.C.M.: A network perspective of the stock market. J. Empirical Finance **17**, 659–667 (2010)
24. Vizgunov, A.N., Goldengorin, B., Kalyagin, V.A., Koldanov, A.P., Koldanov, P., Pardalos, P.M.: Network approach for the Russian stock market. Comput. Manag. Sci. (2013). doi:10.1007/s10287-013-0165-7

Nonconvex Generalized Benders Decomposition

Xiang Li, Arul Sundaramoorthy, and Paul I. Barton

1 Introduction

This chapter is devoted to mixed-integer nonlinear programs (MINLPs) in the following form:

$$\begin{aligned} \min_{x,y} \ & f(x,y) \\ \text{s.t.} \ & g(x,y) \leq 0, \\ & x \in X, \ \ y \in Y, \end{aligned} \tag{P}$$

where $X = \{x = (x_b, x_c) \in \{0,1\}^{n_{x_b}} \times \Pi_{x_c} : p(x) \leq 0\}$, $\Pi_{x_c} \subset \mathbb{R}^{n_{x_c}}$ is convex, $Y = \{y \in \{0,1\}^{n_y} : q(y) \leq 0\}$, $f : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times [0,1]^{n_y} \to \mathbb{R}$, $g : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times [0,1]^{n_y} \to \mathbb{R}^m$, $p : [0,1]^{n_{x_b}} \times \Pi_{x_c} \to \mathbb{R}^{m_p}$, $q : [0,1]^{n_y} \to \mathbb{R}^{m_q}$. Here the subdomains of the functions for binary variables are intervals $[0,1]$ instead of discrete sets $\{0,1\}$, because the functions often need to be defined on these intervals for practical solution of Problem (P) (e.g., via branch-and-bound). y is a vector of *complicating* variables in the sense that Problem (P) is a much easier optimization problem for a fixed y. For example, when Problem (P) is a stochastic program, it may be decomposed into a large number of smaller and easier optimization problems for a fixed y.

X. Li
Department of Chemical Engineering, Queen's University, Kingston, ON, Canada K7L 3N6
e-mail: xiang.li@chee.queensu.ca

A. Sundaramoorthy
Praxair, Inc., Business and Supply Chain Optimization, Tonawanda, NY 14150, USA
e-mail: Arul_Sundaramoorthy@Praxair.com

P.I. Barton (✉)
Process Systems Engineering Laboratory, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: pib@mit.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_16,

The structure of Problem (P) indicates that it may be beneficial to solve the problem by searching the y-space and then the x-space in an iterative manner (instead of searching the xy-space directly). The concept of projection can be used to facilitate this solution strategy, specifically, Problem (P) can be projected onto the y-space as:

$$\begin{aligned} &\min_{y} \ v(y) \\ &\text{s.t.} \ \ y \in Y \cap V, \end{aligned} \tag{P_{proj}}$$

where $v(y) = \inf_{x \in \{x \in X : g(x,y) \leq 0\}} f(x,y)$ and $V = \{y : \exists x \in X,\ g(x,y) \leq 0\}$. When f, g are affine functions and X is a convex polyhedral set, function v and set V in the projected problem can be approximated by cutting planes generated in a dual space, and Problem (P) can be solved by the solution of a sequence of linear programming (LP) and mixed-integer linear programming (MILP) subproblems. This solution method is known as Benders decomposition [1] in the literature. The cutting plane approximation of v and V is also valid when f and g are nonlinear functions that satisfy certain convexity conditions and set X is convex; in this case, Problem (P) can be typically addressed by the solution of a sequence of nonlinear programming (NLP) and MINLP/MILP subproblems, and the solution method is called generalized Benders decomposition (GBD) [2]. However, when functions f, g or set X are nonconvex, neither BD nor GBD can guarantee convergence to the optimal solution due to the loss of strong duality.

This chapter presents an extension of BD/GBD, called nonconvex generalized Benders decomposition (NGBD), to deal with nonconvexity in Problem (P) rigorously. By introducing convex relaxations of nonconvex functions and continuous relaxations of non-complicating binary variables, NGBD can obtain an ε-optimal solution for Problem (P) in finite time. To simplify the discussion, the following assumptions are made for Problem (P).

Assumption 1 *Sets X, Y are nonempty.*

Assumption 2 *Problem (P) either has a minimum or is infeasible for any $y \in Y$.*

Remark 1. Assumption 2 is to exclude the situation in which Problem (P) is feasible but does not have a minimum. This is a mild assumption as it holds when set X is compact and functions f, g are continuous on Π_{x_c} for any feasible x_b, y.

This chapter is organized as follows. The decomposition strategy of NGBD along with the resulting subproblems is introduced in Sect. 2, and important properties of the subproblems are proved in Sect. 3. Then, the NGBD algorithm is given with a proof of the finite convergence property in Sect. 4. In Sect. 5, the application of NGBD to a class of stochastic MINLPs is discussed. The computational advantage of NGBD is demonstrated via case studies of several industrial optimization problems in Sect. 6 and the chapter ends with concluding remarks in Sect. 7.

2 Reformulation and the Subproblems

NGBD is a result of applying the framework of concepts presented by Geoffrion for the design of large-scale mathematical programming techniques [3, 4]. The framework includes two groups of concepts: problem manipulations and solution strategies. Problem manipulations, including projection, dualization, inner and outer linearization, restate a given problem in an alternative form more amenable to solution. Solution strategies, including relaxation and restriction, reduce a complicated problem to a related sequence of simpler subproblems. The following subsections give details on the construction of the NGBD subproblems with the concepts of problem manipulations and solution strategies.

2.1 Convex and Continuous Relaxations: Lower Bounding Problem

One difficulty in decomposing Problem (P) is that dualization does not usually generate an equivalent reformulation for this nonconvex problem. To cope with this difficulty, a surrogate for Problem (P) for which strong duality holds, is constructed via convex relaxations of nonconvex functions and continuous relaxations of noncomplicating binary variables in Problem (P). This new problem, called the lower bounding problem, provides a lower bound for Problem (P), and it can be solved via a procedure similar to GBD. The lower bounding problem can be expressed in the following form:

$$\begin{aligned} &\min_{x,y,e} \; u_f(x,e,y) \\ &\text{s.t.} \;\; u_g(x,e,y) \le 0, \\ &\qquad (x,e) \in D, \;\; y \in Y, \end{aligned} \tag{LBP-NS}$$

where $D = \{(x,e) \in [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e : u_p(x,e) \le 0,\ u_e(x,e) \le 0\}$, Π_e is convex, functions $u_f : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \times [0,1]^{n_y} \to \mathbb{R}$, $u_g : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \times [0,1]^{n_y} \to \mathbb{R}^m$, $u_p : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \to \mathbb{R}^{m_p}$, $u_e : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \to \mathbb{R}^{m_e}$ are all convex on their domains. In addition, the convex functions satisfy the relaxation property, i.e., $\forall \hat{x} \in [0,1]^{n_{x_b}} \times \Pi_{x_c}$ and $\forall \hat{y} \in [0,1]^{n_y}$, $\exists \hat{e} \in \Pi_e$ such that:

$$\begin{aligned} &u_f(\hat{x},\hat{e},\hat{y}) \le f(\hat{x},\hat{y}), \\ &u_g(\hat{x},\hat{e},\hat{y}) \le g(\hat{x},\hat{y}), \\ &u_p(\hat{x},\hat{e}) \le p(\hat{x}), \\ &u_e(\hat{x},\hat{e}) \le 0. \end{aligned} \tag{1}$$

Note that the domain of any binary variable in x has been relaxed into the interval $[0,1]$, and nonconvex functions f, g, p have been replaced with their convex

relaxations u_f, u_g, u_p. The additional variables e and constraints $u_e(x,e) \leq 0$ may be needed if smooth convex relaxations are desired. Standard convex relaxation techniques include McCormick's relaxations [5], outer linearization [6] and αBB [7], and readers can refer to [8] for more discussions on convex relaxation techniques.

Assumption 3 *The relaxed set D is compact.*

Problem (LBP-NS) cannot be practically solved by GBD unless Property P is satisfied [2]. Property P is a strong condition in general, but it trivially holds if the functions in Problem (LBP-NS) are separable in x and y. Therefore, for NGBD to be practical, Problem (LBP-NS) needs to be further relaxed into the following form (if it is not already in this form):

$$\begin{aligned} \min_{x,y,e} \ & u_{f,1}(x,e) + u_{f,2}(y) \\ \text{s.t.} \ & u_{g,1}(x,e) + u_{g,2}(y) \leq 0, \\ & (x,e) \in D, \ \ y \in Y, \end{aligned} \tag{LBP}$$

where functions $u_{f,1} : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \to \mathbb{R}$, $u_{f,2} : [0,1]^{n_y} \to \mathbb{R}$, $u_{g,1} : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \to \mathbb{R}^m$, $u_{g,2} : [0,1]^{n_y} \to \mathbb{R}^m$ are convex on their domains. In addition, $\forall (\hat{x}, \hat{e}, \hat{y}) \in [0,1]^{n_{x_b}} \times \Pi_{x_c} \times \Pi_e \times [0,1]^{n_y}$,

$$\begin{aligned} u_{f,1}(\hat{x}, \hat{e}) + u_{f,2}(\hat{y}) &\leq u_f(\hat{x}, \hat{e}, \hat{y}), \\ u_{g,1}(\hat{x}, \hat{e}) + u_{g,2}(\hat{y}) &\leq u_g(\hat{x}, \hat{e}, \hat{y}). \end{aligned} \tag{2}$$

If functions u_f and u_g are continuous, functions $u_{f,1}$, $u_{f,2}$, $u_{g,1}$, $u_{g,2}$ can always be obtained through outer linearization using their gradient or subgradient information [9, 10].

Assumption 4 *Functions $u_{f,1}$, $u_{f,2}$, $u_{g,1}$, $u_{g,2}$ are continuous.*

Remark 2. Assumptions 3 and 4 imply that Problem (LBP) has a compact feasible set and a continuous objective function, so Problem (LBP) either has finite optimal objective value or is infeasible.

Assumption 5 *Problem (LBP) satisfies Slater's condition for y fixed to those elements in Y for which Problem (LBP) is feasible.*

Remark 3. Assumption 5 implies that strong duality holds for Problem (LBP) for y fixed to those elements in Y for which Problem (LBP) is feasible. This validates the dualization manipulation of the problem in the next subsection.

2.2 Projection/Dualization: Master Problem

Direct solution of Problem (LBP) is generally difficult, as complicating variables y are still present and coupled with non-complicating variables x. Therefore, it is

solved in NGBD via a decomposition procedure that is very similar to classical GBD method. The first step of the decomposition is to project the problem to the y space (as explained in Sect. 1), and then express the objective function and feasible set with the cutting planes in a dual space. Readers can refer to [2] for details on the dualization manipulation. The resulting problem is called the master problem, which can be written in the following form:

$$\begin{aligned}
&\min_{\eta,y} \quad \eta \\
&\text{s.t. } \eta \geq \inf_{(x,e)\in D} \left[u_{f,1}(x,e) + \lambda^{\mathrm{T}} u_{g,1}(x,e)\right] + u_{f,2}(y) + \lambda^{\mathrm{T}} u_{g,2}(y), \quad \forall \lambda \geq 0, \\
&\qquad 0 \geq \inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(y), \quad \forall \mu \in M1, \\
&\qquad y \in Y,\ \eta \in \mathbb{R},
\end{aligned} \tag{MP1}$$

where $\lambda, \mu \in \mathbb{R}^m$ and $M1 = \{\mu \in \mathbb{R}^m : \mu \geq 0, \sum_{i=1}^m \mu_i = 1\}$. For convenience in establishing valid subproblems later, Problem (MP1) is further reformulated into the following form (by replacing set $M1$ with set M):

$$\begin{aligned}
&\min_{\eta,y} \quad \eta \\
&\text{s.t. } \eta \geq \inf_{(x,e)\in D} \left[u_{f,1}(x,e) + \lambda^{\mathrm{T}} u_{g,1}(x,e)\right] + u_{f,2}(y) + \lambda^{\mathrm{T}} u_{g,2}(y), \quad \forall \lambda \geq 0, \\
&\qquad 0 \geq \inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(y), \quad \forall \mu \in M, \\
&\qquad y \in Y,\ \eta \in \mathbb{R},
\end{aligned} \tag{MP}$$

where $M = \{\mu \in \mathbb{R}^m : \mu \geq 0, \sum_{i=1}^m \mu_i > 0\}$. The equivalence of Problems (MP1) and (MP) is proved in the next section.

2.3 Restriction: Primal Problem, Primal Bounding Problem and Feasibility Problem

The primal problem is obtained through restricting y in Problem (P) to an element $y^{(l)}$ in Y, where the superscript l enumerates the sequence of integer realizations visited by the primal problem (i.e. the integer realizations for which the primal problem is constructed and solved). This problem can be written as follows:

$$\begin{aligned}
\mathrm{obj}_{\mathrm{PP}}(y^{(l)}) = \min_x \ & f(x, y^{(l)}) \\
\text{s.t. } & g(x, y^{(l)}) \leq 0, \\
& x \in X,
\end{aligned} \tag{PP^l}$$

where $\text{obj}_{\text{PP}}(y^{(l)})$ denotes the optimal objective value of Problem (PP^l) (which depends on the integer realization $y^{(l)}$).

Remark 4. Problem (PP^l) is a NLP, MILP or MINLP, which can be solved to ε-optimality in finite time by state-of-the-art solvers, such as CPLEX [11] or BARON [6], provided suitable convex underestimators of the participating functions can be constructed.

Similarly, the primal bounding problem is obtained through restricting y in Problem (LBP) to an element $y^{(k)}$ in Y, where the superscript k enumerates the sequence of integer realizations visited by the primal bounding problem. This problem can be written as follows:

$$\begin{aligned} \text{obj}_{\text{PBP}}(y^{(k)}) = \min_{x,e}\ & u_{f,1}(x,e) + u_{f,2}(y^{(k)}) \\ \text{s.t.}\ & u_{g,1}(x,e) + u_{g,2}(y^{(k)}) \le 0, \\ & (x,e) \in D, \end{aligned} \tag{PBP^k}$$

where $\text{obj}_{\text{PBP}}(y^{(k)})$ denotes the optimal objective value of Problem (PBP^k). If Problem (PBP^k) is infeasible, the following feasibility problem is solved:

$$\begin{aligned} \text{obj}_{\text{FP}}(y^{(k)}) = \min_{x,e,z}\ & ||z|| \\ \text{s.t.}\ & u_{g,1}(x,e) + u_{g,2}(y^{(k)}) \le z, \\ & (x,e) \in D, \quad z \in Z, \end{aligned} \tag{FP^k}$$

where $\text{obj}_{\text{FP}}(y^{(k)})$ denotes the optimal objective value of Problem (FP^k), $||z||$ denotes an arbitrary norm of the slack variable vector z, set $Z \subset \{z \in \mathbb{R}^m : z \ge 0\}$ and it has three additional properties:

1. Z is a convex set;
2. Z is a pointed cone, i.e., $0 \in Z$, and $\forall \alpha > 0, z \in Z$ implies $\alpha z \in Z$;
3. There exists $\hat{z} \in Z$ such that $\hat{z} > 0$ (therefore the cone Z is unbounded from above in each dimension).

Each component of z measures the violation of a constraint, so the norm of z is minimized for minimum violation of the constraints. Since any norm function is convex, Problem (FP^k) is convex.

Remark 5. If the convex subproblems (PBP^k) and (FP^k) are smooth, they can be solved by gradient-based optimization solvers such as CONOPT [12], SNOPT [13], CPLEX [11] (only for linear programs, convex quadratic programs and convex quadratically constrained programs). Otherwise, they may be solved by nonsmooth optimization methods such as bundle methods [14].

2.4 Relaxation: Relaxed Master Problem

The master problem (MP) is difficult to solve directly because of the infinite number of constraints involved. Therefore, it is relaxed by only keeping a finite number of constraints. The resulting subproblem is called the relaxed master problem; at the kth iteration, this subproblem can be written in the following form:

$$\begin{aligned}
&\min_{\eta,y} \quad \eta \\
&\text{s.t. } \eta \geq \inf_{(x,e)\in D} \left[u_{f,1}(x,e) + \left(\lambda^{(j)}\right)^{\mathrm{T}} u_{g,1}(x,e)\right] + u_{f,2}(y) + \left(\lambda^{(j)}\right)^{\mathrm{T}} u_{g,2}(y), \quad \forall j \in T^k, \\
&\quad 0 \geq \inf_{(x,e)\in D} \left(\mu^{(i)}\right)^{\mathrm{T}} u_{g,1}(x,e) + \left(\mu^{(i)}\right)^{\mathrm{T}} u_{g,2}(y), \quad \forall i \in S^k, \\
&\quad \sum_{r\in R_1^t} y_r - \sum_{r\in R_0^t} y_r \leq |R_1^t| - 1, \quad \forall t \in T^k \cup S^k, \\
&\quad y \in Y, \ \eta \in \mathbb{R},
\end{aligned} \tag{RMP1^k}$$

where the index sets

$$\begin{aligned}
T^k &= \{j \in \{1,\ldots,k\} : \text{Problem (PBP) is feasible for } y = y^{(j)}\}, \\
S^k &= \{i \in \{1,\ldots,k\} : \text{Problem (PBP) is infeasible for } y = y^{(i)}\}, \\
R_1^t &= \{r \in \{1,\ldots,n_y\} : y_r^{(t)} = 1\}, \\
R_0^t &= \{r \in \{1,\ldots,n_y\} : y_r^{(t)} = 0\}.
\end{aligned}$$

$\lambda^{(j)}$ denotes Lagrange multipliers for Problem (PBP^j), which form an optimality cut for iteration $j \in T^k$. $\mu^{(i)}$ denotes Lagrange multipliers for Problem (FP^i), which form a feasibility cut for iteration $i \in S^k$. To be precise, the definition of a Lagrange multiplier is given below.

Definition 1. λ^* is a Lagrange multiplier for the optimization problem

$$\begin{aligned}
&\min_x f(x) \\
&\text{s.t.} \quad g(x) \leq 0, \\
&\qquad x \in X,
\end{aligned}$$

if $\lambda^* \geq 0$ and $f(x^*) = \inf_{x\in X} \left[f(x) + (\lambda^*)^{\mathrm{T}} g(x)\right]$, where x^* denotes an optimal solution of the problem.

Remark 6. Definition 1 for Lagrange multipliers follows from [15] in the context of duality theory (where they are called geometric multipliers instead). This definition is consistent with the one used by Geoffrion for the GBD method [2] and duality theory [16] (where they are called optimal multipliers). Note that the Lagrange multipliers defined here are in general different from the multipliers that satisfy the

Karush–Kuhn–Tucker (KKT) conditions, which are usually called KKT multipliers. However, for convex program (PBP^k) or (FP^k), KKT multipliers are also Lagrange multipliers, as implied by the theorem on [17, p. 211]. State-of-the-art optimization solvers, such as CONOPT, SNOPT, CPLEX, return such multiplier values at a solution, so there is no need to develop an additional algorithm to obtain the Lagrange multipliers for Problem (PBP^k) or (FP^k) in NGBD.

The third group of constraints in Problem (RMP1^k), which does not appear in the master problem (MP), represents a set of canonical integer cuts that prevent the previously examined integer realizations from becoming a solution [18].

When $T^k = \emptyset$, Problem (RMP1^k) is unbounded; in this case, the following feasibility relaxed master problem is solved instead:

$$\begin{aligned}
&\min_{y} \ \sum_{i=1}^{n_y} y_i \\
&\text{s.t. } 0 \geq \inf_{(x,e)\in D} \left(\mu^{(i)}\right)^{\mathrm{T}} u_{g,1}(x,e) + \left(\mu^{(i)}\right)^{\mathrm{T}} u_{g,2}(y), \quad \forall i \in S^k, \\
&\quad \sum_{r\in R_1^t} y_r - \sum_{r\in R_0^t} y_r \leq |R_1^t| - 1, \quad \forall t \in S^k, \\
&\quad y \in Y.
\end{aligned} \tag{FRMP1^k}$$

The inner optimization problems in Problems (RMP1^k) and (FRMP1^k) can be replaced by the solution information of the previously solved primal bounding problems and feasibility problems (which is to be explained in the next section). As a result, Problem (RMP1^k) is equivalent to the following single-level optimization problem:

$$\begin{aligned}
&\min_{\eta,y} \quad \eta \\
&\text{s.t.} \quad \eta \geq \mathrm{obj}_{\mathrm{PBP}}(y^{(j)}) + u_{f,2}(y) - u_{f,2}(y^{(j)}) \\
&\qquad + \left(\lambda^{(j)}\right)^{\mathrm{T}} \left(u_{g,2}(y) - u_{g,2}(y^{(j)})\right), \forall j \in T^k, \\
&\qquad 0 \geq \mathrm{obj}_{\mathrm{FP}}(y^{(i)}) + \left(\mu^{(i)}\right)^{\mathrm{T}} \left(u_{g,2}(y) - u_{g,2}(y^{(i)})\right), \quad \forall i \in S^k, \\
&\qquad \sum_{r\in R_1^t} y_r - \sum_{r\in R_0^t} y_r \leq |R_1^t| - 1, \quad \forall t \in T^k \cup S^k, \\
&\qquad y \in Y, \ \ \eta \in \mathbb{R}.
\end{aligned} \tag{RMP^k}$$

and Problem (FRMP1^k) is:

$$\begin{aligned}
\min_{y} \quad & \sum_{i=1}^{n_y} y_i \\
\text{s.t.} \quad & 0 \geq \text{obj}_{\text{FP}}(y^{(i)}) + \left(\mu^{(i)}\right)^{\text{T}} \left(u_{g,2}(y) - u_{g,2}(y^{(i)})\right), \quad \forall i \in S^k, \qquad (\text{FRMP}^k) \\
& \sum_{r \in R_1^t} y_r - \sum_{r \in R_0^t} y_r \leq |R_1^t| - 1, \quad \forall t \in S^k, \\
& y \in Y.
\end{aligned}$$

Remark 7. Problem (RMP^k) or (FRMP^k) is a convex MINLP or a MILP. Commercial solvers are available for solving these problems, such as DICOPT [19] (for convex MINLP), CPLEX (for MILP).

This section details the reformulation of the original MINLP into a collection of subproblems through convex and continuous relaxations, projection, dualization, restriction and relaxation. The subproblems to be solved directly in NGBD include Problems (PP^l), (PBP^k), (FP^k), (RMP^k) and (FRMP^k). The lower bounding problem and the master problem are also generated in the reformulation, but they are not solved directly in the NGBD procedure. In the next section, a set of propositions regarding the properties and the relationships of the subproblems are presented and proved. Based on these results, the NGBD algorithm is developed with a convergence proof in Sect. 4.

3 Properties of the Subproblems

Proposition 1. *The optimal objective value of Problem (LBP) represents a lower bound on the optimal objective value of Problem (P).*

Proof. Let $(\hat{x}, \hat{y})$ be a minimum of Problem (P). Then $\hat{x} \in [0,1]^{n_{x_b}} \times \Pi_{x_c}$ and $\hat{y} \in [0,1]^{n_y}$. According to (1), $\exists \hat{e} \in \Pi_e$ such that

$$\begin{aligned}
& u_f(\hat{x}, \hat{e}, \hat{y}) \leq f(\hat{x}, \hat{y}), \\
& u_g(\hat{x}, \hat{e}, \hat{y}) \leq g(\hat{x}, \hat{y}) \leq 0, \\
& u_p(\hat{x}, \hat{e}) \leq p(\hat{x}) \leq 0, \\
& u_e(\hat{x}, \hat{e}) \leq 0.
\end{aligned}$$

So point $(\hat{x}, \hat{e}, \hat{y})$ is feasible for Problem (LBP-NS), and the objective value of Problem (LBP-NS) at this point is no larger than the optimal objective value of Problem (P).

According to (2), $(\hat{x}, \hat{e}, \hat{y})$ also satisfy

$$\begin{aligned}
& u_{f,1}(\hat{x}, \hat{e}) + u_{f,2}(\hat{y}) \leq u_f(\hat{x}, \hat{e}, \hat{y}) \leq f(\hat{x}, \hat{y}), \\
& u_{g,1}(\hat{x}, \hat{e}) + u_{g,2}(\hat{y}) \leq u_g(\hat{x}, \hat{e}, \hat{y}) \leq 0.
\end{aligned}$$

So this point is also feasible for Problem (LBP) and the objective value of Problem (LBP) at this point is also no larger than the optimal objective value of Problem (P). Therefore, the optimal objective value of Problem (LBP) is no larger than that of Problem (P). □

Proposition 2. *Problems (LBP) and (MP1) are equivalent in the sense that:*

(1) Problem (LBP) is feasible iff Problem (MP1) is feasible;
(2) The optimal objective values of Problems (LBP) and (MP1) are the same;
(3) The optimal objective value of Problem (LBP) is attained with an integer realization iff the optimal objective value of Problem (MP1) is attained with the same integer realization.

Proof. Given Assumption 5, the results follow immediately from Theorems 2.1, 2.2 and 2.3 in [2]. □

Proposition 3. *Problems (MP1) and (MP) are equivalent in the sense that:*

(1) Problem (MP1) is feasible iff Problem (MP) is feasible;
(2) The optimal objective values of Problems (MP1) and (MP) are the same;
(3) The optimal objective value of Problem (MP1) is attained with an integer realization iff the optimal objective value of Problem (MP) is attained with the same integer realization.

Proof. The results can be proved by showing that Problems (MP1) and (MP) have the same feasible set. Denote the feasible regions of Problems (MP1) and (MP) by F_{MP1} and F_{MP}, respectively. $F_{\mathrm{MP1}} = F_{\mathrm{MP}}$ can be proved by showing $F_{\mathrm{MP}} \subset F_{\mathrm{MP1}}$ and $F_{\mathrm{MP1}} \subset F_{\mathrm{MP}}$.

First, for any $(\hat{y}, \hat{\eta}) \in F_{\mathrm{MP}}$,

$$0 \geq \inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(\hat{y}), \quad \forall \mu \in M,$$

so

$$0 \geq \inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(\hat{y}), \quad \forall \mu \in M1,$$

because $M1 \subset M$. Therefore, $F_{\mathrm{MP}} \subset F_{\mathrm{MP1}}$.

Second, for any $(\hat{y}, \hat{\eta}) \in F_{\mathrm{MP1}}$,

$$0 \geq \inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(\hat{y}), \quad \forall \mu \in M1. \tag{3}$$

For such $(\hat{y}, \hat{\eta})$, consider any $\hat{\mu} \in M$,

$$\sum_{i=1}^{m} \hat{\mu}_i > 0. \tag{4}$$

So $\hat{\mu}$ can be used to define new multipliers

$$\tilde{\mu}_i = \hat{\mu}_i / \sum_{i=1}^{m} \hat{\mu}_i, \quad \forall i \in \{1, \ldots, m\}, \tag{5}$$

then

$$\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_m) \in M1. \tag{6}$$

From (3) and (6),

$$\inf_{(x,e)\in D} \tilde{\mu}^{\mathrm{T}} u_{g,1}(x,e) + \tilde{\mu}^{\mathrm{T}} u_{g,2}(\hat{y}) \leq 0. \tag{7}$$

Considering (4), (5) and (7)

$$\begin{aligned}
&\inf_{(x,e)\in D} \hat{\mu}^{\mathrm{T}} u_{g,1}(x,e) + \hat{\mu}^{\mathrm{T}} u_{g,2}(\hat{y}) \\
&= \left(\sum_{i=1}^{m} \hat{\mu}_i\right) \inf_{(x,e)\in D} \left(\hat{\mu} / \left(\sum_{i=1}^{m} \hat{\mu}_i\right)\right)^{\mathrm{T}} u_{g,1}(x,e) + \left(\hat{\mu} / \left(\sum_{i=1}^{m} \hat{\mu}_i\right)\right)^{\mathrm{T}} u_{g,2}(\hat{y}) \\
&= \left(\sum_{i=1}^{m} \hat{\mu}_i\right) \left(\inf_{(x,e)\in D} \tilde{\mu}^{\mathrm{T}} u_{g,1}(x,e) + \tilde{\mu}^{\mathrm{T}} u_{g,2}(\hat{y})\right) \\
&\leq 0
\end{aligned}$$

Therefore,

$$\inf_{(x,e)\in D} \mu^{\mathrm{T}} u_{g,1}(x,e) + \mu^{\mathrm{T}} u_{g,2}(\hat{y}) \leq 0, \quad \forall \mu \in M,$$

and $(\hat{y}, \hat{\eta}) \in F_{\mathrm{MP}}$ too. So $F_{\mathrm{MP1}} \subset F_{\mathrm{MP}}$. □

Proposition 4. *For y fixed to any element in Y, if Problem (PP^l) is feasible, its optimal objective value is no less than the optimal objective value of Problem (P).*

Proof. This result trivially holds due to the construction of Problem (PP^l) and the principle of restriction. □

Proposition 5. *If the primal problem (PP^k) is feasible, the corresponding primal bounding problem (PBP^k) is feasible as well. In this case, the optimal objective value of Problem (PP^k) is no less than that of Problem (PBP^k).*

Proof. This can be proved according to the construction of these problems in the same way to prove Proposition 1. □

Remark 8. Proposition 5 implies that, if the optimal objective value of Problem (PBP^k) is worse than that of Problem (P), there is no need to solve Problem (PP^l) because $y = y^{(k)}$ cannot lead to an optimum of Problem (P). This property will be exploited in the NGBD algorithm to reduce the number of the primal problems to be solved, since obtaining a global optimum for the primal problem is computationally expensive.

Proposition 6. *Problem (FP^k) satisfies Slater's condition and it always has a minimum.*

Proof. According to Assumption 5, set D has at least one Slater point, say $(\hat{x},\hat{e})$. Due to the continuity of functions $u_{g,1}$, $u_{g,2}$, $u_{g,1}(\hat{x},\hat{e})+u_{g,2}(y^{(k)})$ is finite, so there exists $\hat{z}\in Z$ such that $u_{g,1}(\hat{x},\hat{e})+u_{g,2}(y^{(k)})<\hat{z}$. Then $(\hat{x},\hat{e},\hat{z})$ is a Slater point of Problem (FP^k). In addition, Problem (FP^k) has a closed feasible set and $||z||$ is continuous and coercive on Z, so Problem (FP^k) has a minimum according to Weierstrass' Theorem [15]. □

Proposition 7. *Let μ^* be Lagrange multipliers of Problem (FP^k). If Problem (PBP^k) is infeasible, $\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right]$ is a finite positive value and $\sum_{i=1}^{m}\mu_i^*>0$.*

Proof. As Lagrange multipliers,

$$\mu^*\geq 0. \tag{8}$$

Let (x^*,e^*,z^*) be a minimum of Problem (FRMP1^k), then due to strong duality (implied by Proposition 6 and the convexity of the problem),

$$\begin{aligned}||z^*|| &= \inf_{(x,e,z)\in D\times Z}\left[||z||+(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})-z\right)\right]\\ &= \inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]+\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right].\end{aligned} \tag{9}$$

First, $\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right]$ is finite due to the compactness of the feasible set and continuity of the objective function of the problem.

Second, we will show that $\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]=0$ by contradiction. Suppose that

$$\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]<0, \tag{10}$$

then $\exists\varepsilon>0$ such that

$$\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]<-\varepsilon. \tag{11}$$

Hence, $\forall\alpha>0$,

$$\alpha\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]<-\alpha\varepsilon, \tag{12}$$

which is

$$\inf_{z\in Z}\left[||\alpha z||-(\mu^*)^{\mathrm{T}}\alpha z\right]<-\alpha\varepsilon. \tag{13}$$

Since $\forall z\in Z$, $\alpha z\in Z$ as well,

$$\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]=\inf_{z\in Z}\left[||\alpha z||-(\mu^*)^{\mathrm{T}}(\alpha z)\right]<-\alpha\varepsilon \tag{14}$$

and therefore

$$\inf_{z\in Z}\left[||z||-(\mu^*)^{\mathrm{T}}z\right]=-\infty. \tag{15}$$

According to (9), (15) and finiteness of $\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right]$, $||z^*|| = -\infty$, which contradicts the definition of a norm. Therefore, (10) is not true and

$$\inf_{z\in Z}\left[||z|| - (\mu^*)^{\mathrm{T}} z\right] \geq 0. \tag{16}$$

On the other hand, when $z = 0$ ($\in Z$), $||z|| - (\mu^*)^{\mathrm{T}} z = 0$, so

$$\inf_{z\in Z}\left[||z|| - (\mu^*)^{\mathrm{T}} z\right] \leq 0. \tag{17}$$

Inequalities (16) and (17) imply

$$\inf_{z\in Z}\left[||z|| - (\mu^*)^{\mathrm{T}} z\right] = 0. \tag{18}$$

Finally, according to (9) and (18),

$$\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right] = ||z^*||. \tag{19}$$

If Problem (PBP^k) is infeasible, $z^* \neq 0$ and therefore $||z^*|| > 0$, then (19) implies

$$\inf_{(x,e)\in D}\left[(\mu^*)^{\mathrm{T}}\left(u_{g,1}(x,e)+u_{g,2}(y^{(k)})\right)\right] > 0, \tag{20}$$

which further implies

$$\mu^* \neq 0. \tag{21}$$

So according to (8) and (21),

$$\sum_{i=1}^{m} \mu_i^* > 0. \tag{22}$$

□

Proposition 8. *Problem (RMP1^k) is a relaxation of the master problem (MP) when (MP) is augmented with the relevant canonical integer cuts excluding the previously examined integer realizations.*

Proof. As Lagrange multipliers, $\lambda^{(j)} \geq 0$, $\forall j \in T^k$. According to Proposition 7, $\mu^{(i)} \in M$, $\forall i \in S^k$. Therefore, Problem (RMP1^k) is a relaxation of the master problem (MP) excluding all the previously examined integer variables (i.e. the master problem augmented with the integer cuts). □

Proposition 9. *Problems (RMP1^k) and (RMP^k) are equivalent.*

Proof. This follows from the separability of the functions in the continuous and the integer variables. Detailed proof can be found in [2]. □

Corollary 1. *Problem (RMP^k) or (FRMP^k) never generates the same integer solution twice.*

Corollary 2. *The optimal objective value of Problem (RMP^k) is a valid lower bound for the lower bounding problem (LBP) (or the master problem (MP)) augmented with the relevant canonical integer cuts and the original problem (P) augmented with the relevant canonical integer cuts.*

4 NGBD Algorithm

4.1 Algorithm

Initialize:

1. Iteration counters $k = 0$, $l = 1$ and the index sets $T^0 = \emptyset$, $S^0 = \emptyset$, $U^0 = \emptyset$.
2. Upper bound for Problem (P) $UBD = +\infty$, upper bound for Problem (LBP) $UBDPB = +\infty$, lower bound for Problems (LBP) and (P) $LBD = -\infty$.
3. Tolerance ε is set and initial integer realization $y^{(1)}$ is given.

repeat
if $k = 0$ or (Problem (RMP^k) is feasible and $LBD < UBDPB$ and $LBD < UBD - \varepsilon$) **then**
repeat
Set $k = k + 1$

1. Solve Problem (PBP^k). If Problem (PBP^k) is feasible and has Lagrange multipliers $\lambda^{(k)}$, add an optimality cut to Problem (RMP^k) with $\lambda^{(k)}$, set $T^k = T^{k-1} \cup \{k\}$. If $\text{obj}_{\text{PBP}}(y^{(k)}) < UBDPB$, update $UBDPB = \text{obj}_{\text{PBP}}(y^{(k)})$, $y^* = y^{(k)}$, $k^* = k$.
2. If Problem (PBP^k) is infeasible, set $S^k = S^{k-1} \cup \{k\}$. Then, solve Problem (FP^k) and obtain the corresponding Lagrange multipliers $\mu^{(k)}$. Add a feasibility cut to Problem (RMP^k) with $\mu^{(k)}$.
3. If $T^k \neq \emptyset$, solve Problem (RMP^k); otherwise, solve Problem (FRMP^k). In the former case, set LBD to the optimal objective value of Problem (RMP^k) if Problem (RMP^k) is feasible. In either case, set $y^{(k+1)}$ to the y value at the solution of either problem.

until $LBD \geq UBDPB$ or (Problem (RMP^k) or (FRMP^k) is infeasible).
end if
if $UBDPB < UBD - \varepsilon$

1. Solve Problem (PP*) (i.e., for $y = y^*$) to ε-optimality, set $U^l = U^{l-1} \cup \{k^*\}$. If Problem (PP*) has a minimum x^* and $\text{obj}_{\text{PP}}(y^*) < UBD$, update $UBD = \text{obj}_{\text{PP}}(y^*)$ and set $y_p^* = y^*$, $x_p^* = x^*$.
2. If $T^k \setminus U^l = \emptyset$, set $UBDPB = +\infty$.
3. If $T^k \setminus U^l \neq \emptyset$, pick $i \in T^k \setminus U^l$ such that $\text{obj}_{\text{PBP}}(y^{(i)}) = \min_{j \in T^k \setminus U^l}\{\text{obj}_{\text{PBP}}(y^{(j)})\}$. Update $UBDPB = \text{obj}_{\text{PBP}}(y^{(i)})$, $y^* = y^{(i)}$, $k^* = i$. Set $l = l + 1$.

end if
until $UBDPB \geq UBD - \varepsilon$ and (Problem (RMP^k) or (FRMP^k) is infeasible or $LBD \geq UBD - \varepsilon$).
Problem (P) has an ε-optimal solution (x_p^*, y_p^*) or it is infeasible.

4.2 Finite Convergence

Assumption 6 *Compared to Problem (PP^l), Problems (PBP^k) and (FP^k) (which are convex NLPs or LPs) and Problems (RMP^k) and (FRMP^k) (which are convex MINLPs or MILPs) can be solved with a much tighter tolerance, which is then negligible for the discussion of the ε-optimality of the NGBD algorithm.*

Assumption 7 *The optimal objective value of a problem returned by a global optimizer is worse than or equal to the real optimal objective value.*

Remark 9. Note that $UBDPB$ is neither the upper bound, nor the lower bound for Problem (P). $UBDPB$ has two functions in the algorithm. One is to control the "inner loop" of the algorithm (which is a GBD-like procedure). The other is to prevent solving Problem (PP) for any integer realization that will not lead to a global solution of Problem (P), and this is explained in the following Lemma 1.

Lemma 1. *If the NGBD algorithm terminates finitely with a feasible solution of Problem (P), this feasible solution is an ε-optimal solution of Problem (P).*

Proof. Note that the algorithm terminates with "$UBDPB \geq UBD - \varepsilon$ and (Problem (RMP^k) or Problem (FRMP^k) is infeasible" or "$LBD \geq UBD - \varepsilon$)". First, it is demonstrated that this termination condition ensures that an integer realization which leads to an ε-optimal solution of Problem (P) has been visited by Problem (PBP). Second, it is demonstrated that if one such integer realization has been visited by Problem (PBP), the termination condition ensures that $y = y_p^*$ is one such integer realization and UBD is an ε-optimal objective value of Problem (P).

Consider the case in which Problem (RMP^k) or (FRMP^k) is infeasible. Since Problem (P) is feasible, Problem (FRMP^k) cannot be infeasible and the infeasibility of Problem (RMP^k) implies that all the feasible integer realizations have been visited by Problem (PBP), so any integer realization leading to an ε-optimal solution of Problem (P) has been visited by Problem (PBP).

Consider the case in which $LBD \geq UBD - \varepsilon$. Denote the real optimal objective value of the original problem (P) by $\widehat{\text{obj}}_{\text{P}}$. Denote the real optimal objective value of Problem (PP) for $y = y_p^*$ by $\widehat{\text{obj}}_{\text{PP}}^*$ and the one returned by the solver by obj_{PP}^*. Obviously,

$$\widehat{\text{obj}}_{\text{PP}}^* \geq \widehat{\text{obj}}_{\text{P}}. \tag{23}$$

According to Assumption 7,

$$UBD = \text{obj}_{\text{PP}}^* \geq \widehat{\text{obj}}_{\text{PP}}^*. \tag{24}$$

From (23) and (24),

$$UBD \geq \widehat{\text{obj}}_{\text{P}}. \tag{25}$$

Assume any integer realization that leads to an ε-optimal solution of Problem (P) has not been visited by Problem (PBP), then any such integer realization has not been excluded by the canonical integer cuts in Problem (RMP^k). According to Corollary 2,

$$\widehat{\text{obj}}_{\text{P}} \geq LBD \geq UBD - \varepsilon. \tag{26}$$

Inequalities (25) and (26) imply that $y = y_p^*$ obtained at the termination of the algorithm leads to an ε-optimal solution of Problem (P) and this integer realization has been visited by Problem (PBP), which contradicts the assumption. Therefore, in the case in which $LBD \geq UBD - \varepsilon$, at least one integer realization that leads to an ε-optimal solution of Problem (P) has been visited by Problem (PBP) as well.

Finally, the algorithm ensures that $UBDPB$ equals the minimum optimal objective value of Problem (PBP) for those integer realizations that have been visited by Problem (PBP) but not by Problem (PP) (and $UBDPB = +\infty$ if no such integer realizations exist). Then at the termination when $UBDPB \geq UBD - \varepsilon$ always holds, such integer realizations cannot lead to a global optimal solution of Problem (P) due to Proposition 5. Therefore, an integer realization that leads to an ε-optimal solution of Problem (P) has been visited by Problem (PP), which has been recorded by $y = y_p^*$ and $UBD = \text{obj}_{\text{PP}}^*$ is an ε-optimal objective value of Problem (P).

Theorem 1. *If all the subproblems can be solved to ε-optimality in a finite number of steps, then the NGBD algorithm terminates in a finite number of steps with an ε-optimal solution of Problem (P) or an indication that Problem (P) is infeasible.*

Proof. Notice that all the integer realizations are generated by solving Problem (RMP^k) or (FRMP^k) in the algorithm. According to Corollary 1, no integer realizations will be generated twice. Since the cardinality of set Y is finite by definition and all the subproblems are terminated in finite number of steps, the algorithm terminates in a finite number of steps.

Lemma 1 shows that if Problem (P) is feasible, the algorithm terminates with its ε-optimal solution. If Problem (P) is infeasible, the algorithm terminates with $UBD = +\infty$ because UBD can only be updated with an ε-optimal solution of Problem (PP), which is infeasible for any integer realization in Y (and therefore UBD is never updated).

5 Application to Stochastic MINLPs

MINLP is widely adopted to model problems that involve discrete and continuous decisions and nonlinearities. Over the past several decades there has been a tremendous amount of work on the development and solution of MINLP models in various engineering areas [20, 21], from product and process design to process operation and control [22]. While these problems have been traditionally solved with

deterministic MINLP models, recently more attention has been paid to including uncertainty considerations in the model, typically using a stochastic programming approach [23]. This section discusses the application of NGBD to scenario-based, two-stage stochastic MINLPs in the following form:

$$\begin{aligned} \min_{x_1,\ldots,x_s,y} \ & \sum_{h=1}^{s} f_h(x_h,y) \\ \text{s.t.} \ & g_h(x_h,y) \le 0, \quad \forall h \in \{1,\ldots,s\}, \\ & x_h \in X_h, \quad \forall h \in \{1,\ldots,s\}, \\ & y \in Y, \end{aligned} \tag{P-SMIP}$$

where $X_h = \{x_h \in \{0,1\}^{n_{x_b}} \times \Pi_{x_c} : p_h(x_h) \le 0\}$, $\Pi_{x_c} \subset \mathbb{R}^{n_{x_c}}$ is convex, $Y = \{y \in \{0,1\}^{n_y} : q(y) \le 0\}$, $f_h : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times [0,1]^{n_y} \to \mathbb{R}$, $g_h : [0,1]^{n_{x_b}} \times \Pi_{x_c} \times [0,1]^{n_y} \to \mathbb{R}^m$, $p_h : [0,1]^{n_{x_b}} \times \Pi_{x_c} \to \mathbb{R}^{m_p}$, $q_h : [0,1]^{n_y} \to \mathbb{R}^{m_q}$. Here uncertainties are characterized by s different uncertainty realizations, also called *scenarios* [23, 24], which are indexed by h. y involves binary variables representing first-stage decisions that are made before realization of the uncertainties. x_h involves binary and/or continuous variables presenting second-stage decisions made after the outcome of scenario h. $f_h(x_h,y)$ in the objective function is related to a cost associated with the realization of scenario h. Problem (P-SMIP) is assumed to satisfy all assumptions made for Problem (P), so it is a special case of Problem (P) and inherits all the properties of Problem (P).

The size of Problem (P-SMIP) depends on the number of scenarios (s) addressed. When s is large, Problem (P) is a large-scale MINLP even if the model with one scenario is small. Obviously, y is a vector of complicating variables for Problem (P-SMIP), as the problem can naturally be decomposed into s subproblems if y is fixed. General-purpose deterministic global optimization methods, such as branch-and-reduce [6], SMIN-αBB and GMIN-αBB [25], and nonconvex outer approximation [26], cannot fully exploit the decomposable structure of Problem (P-SMIP). These methods have to solve a sequence of subproblems whose sizes grow with the number of scenarios in the problem, so they are usually not practical for Problem (P-SMIP) with large numbers of scenarios.

It is not difficult to find that Problem (P-SMIP) can be solved by NGBD. Primal problem, primal bounding problem, and feasibility problem for Problem (P-SMIP) can all be decomposed over the scenarios. The decomposed primal subproblem, primal bounding subproblem and feasibility subproblem are given below as Problem (PP_h^l-SMIP), Problem (PBP_h^k-SMIP), Problem (FP_h^k-SMIP) for any scenario h.

$$\begin{aligned} \text{obj}_{\text{PP}_h}(y^{(l)}) = \min_{x_h} \ & f_h(x_h,y^{(l)}) \\ \text{s.t.} \ & g_h(x_h,y^{(l)}) \le 0, \\ & x_h \in X_h. \end{aligned} \tag{PP_h^l-SMIP}$$

$$\begin{aligned}\mathrm{obj}_{\mathrm{PBP}_h}(y^{(k)}) = \min_{x_h,e_h}\ & u_{f,1,h}(x_h,e_h)+u_{f,2,h}(y^{(k)}) \\ \text{s.t. }\ & u_{g,1,h}(x_h,e_h)+u_{g,2,h}(y^{(k)})\le 0, \\ & (x_h,e_h)\in D_h.\end{aligned} \tag{PBP$_h^k$-SMIP}$$

$$\begin{aligned}\mathrm{obj}_{\mathrm{FP}_h}(y^{(k)}) = \min_{x_h,e_h,z_h}\ & ||z_h|| \\ \text{s.t. }\ & u_{g,1,h}(x_h,e_h)+u_{g,2,h}(y^{(k)})\le z_h, \\ & (x_h,e_h)\in D_h,\quad z_h\in Z_h.\end{aligned} \tag{FP$_h^k$-SMIP}$$

Note that functions $u_{f,1,h}$, $u_{f,2,h}$, $u_{g,1,h}$, $u_{g,2,h}$ and sets D_h are obtained through convex and continuous relaxations as described in Sect. 2. Problem (PP$_h^l$-SMIP) needs to be solved to ε_h-optimality to ensure ε-optimality of the NGBD algorithm, where $\sum_{h=1}^m \varepsilon_h \le \varepsilon$.

It is also easy to derive customized (RMP^k) and (FRMP^k) for Problem (P-SMIP). They are given below as Problem (RMP^k-SMIP) and Problem (FRMP$_h^k$-SMIP).

$$\begin{aligned}\min_{\eta,y}\ & \eta \\ \text{s.t.}\ & \eta \ge \sum_{h=1}^{s}\left[\mathrm{obj}_{\mathrm{PBP}_h}(y^{(j)})+u_{f,2,h}(y)-u_{f,2,h}(y^{(j)})+\left(\lambda_h^{(j)}\right)^{\mathrm{T}}\left(u_{g,2,h}(y)-u_{g,2,h}(y^{(j)})\right)\right], \\ & \qquad \forall j\in T^k, \\ & 0\ge \sum_{h=1}^{s}\left[\mathrm{obj}_{\mathrm{FP}_h}(y^{(i)})+\left(\mu_h^{(i)}\right)^{\mathrm{T}}\left(u_{g,2,h}(y)-u_{g,2,h}(y^{(i)})\right)\right],\quad \forall i\in S^k, \\ & \sum_{r\in R_1^t} y_r-\sum_{r\in R_0^t} y_r\le |R_1^t|-1,\quad \forall t\in T^k\cup S^k, \\ & y\in Y,\ \ \eta\in\mathbb{R}.\end{aligned} \tag{RMP^k-SMIP}$$

$$\begin{aligned}\min_{y}\ & \sum_{i=1}^{n_y} y_i \\ \text{s.t.}\ & 0\ge \sum_{h=1}^{s}\left[\mathrm{obj}_{\mathrm{FP}_h}(y^{(i)})+\left(\mu_h^{(i)}\right)^{\mathrm{T}}\left(u_{g,2,h}(y)-u_{g,2,h}(y^{(i)})\right)\right],\quad \forall i\in S^k, \\ & \sum_{r\in R_1^t} y_r-\sum_{r\in R_0^t} y_r\le |R_1^t|-1,\quad \forall t\in S^k, \\ & y\in Y.\end{aligned} \tag{FRMP$_h^k$-SMIP}$$

Note that $\lambda_h^{(j)}$, $\mu_h^{(i)}$ are Lagrange multipliers for Problem (PBP$_h^k$-SMIP) and Problem (FP$_h^k$-SMIP), respectively.

Obviously, the sizes of the subproblems to be solved in NGBD for Problem (P-SMIP) are all independent of the number of scenarios. This brings tremendous computational advantage for problems with large numbers of scenarios. On the one hand, the number of subproblems to be solved in NGBD grows linearly with s, so the

NGBD solution time grows roughly linearly with s if the total number of iterations does not change significantly with s. On the other hand, the computational complexity of a general-purpose optimization method is usually worse than linear, e.g. polynomial for linear programs and some convex programs, worst-case exponential for (global optimization of) nonconvex and/or mixed-integer programs. So the solution times of these methods usually increase with s much faster than NGBD does. This is demonstrated in the case study results in the next section. In addition, the primal subproblems in one iteration can be solved simultaneously, because the solution of one of the subproblems is not dependent on the solution of the others. The primal bounding and feasibility subproblems have the same feature. Therefore, the NGBD solution time can be readily reduced through parallel computation.

A large number of stochastic MINLPs in the literature arise from integrated system design and operation problems, in which y represents design decisions that are related to the development of the infrastructure of the system, and x_h represents operational decisions that are related to the operation of the system for scenario h. In this case, functions f_h, g_h in Problem (P-SMIP) are often separable in y and x_h and affine in y. In other words, Problem (P-SMIP) can be expressed in the following form:

$$\begin{aligned} \min_{x_1,\ldots,x_s,y} \ & \sum_{h=1}^{s} f_h(x_h) + c_h^{\mathrm{T}} y \\ \text{s.t.} \ & g_h(x_h) + B_h y \le 0, \quad \forall h \in \{1,\ldots,s\}, \\ & x_h \in X_h, \quad \forall h \in \{1,\ldots,s\}, \\ & y \in Y. \end{aligned} \tag{P-SMIP-S}$$

The application of NGBD to Problem (P-SMIP-S) is a lot easier than the general case. As y and x_h are already separable in Problem (P-SMIP-S), the lower bounding problem (LBP) can be constructed without the construction of the intermediate problem (LBP-NS). In addition, Problems (RMP^k) and (FRMP^k) are always MILPs (that are usually easier to solve than MINLPs).

Furthermore, when all functions in Problem (P-SMIP) are affine, the problem becomes a MILP and it can be solved by NBGD via the solution of a sequence of MILP and LP subproblems. Note that this MILP cannot be solved by BD or GBD in general, as set X_h is nonconvex due to the binary variables involved.

6 Case Studies

6.1 Case Study Problems

Three industrial problems are studied here to demonstrate the computational advantage of NGBD over state-of-the-art commercial solvers. Brief descriptions of the case study problems are given below.

6.1.1 Pump Network Configuration Problem

This problem is to find the optimal configuration of a centrifugal pump network that achieves a prespecified pressure rise based on a given total flow rate. The objective of the optimization is to minimize annualized cost. The deterministic version of the problem was initially presented in [27] and then updated in [25] with a set of additional linear constraints for tighter relaxation in global optimization. Here the problem is further reformulated to reduce the number of nonlinear functions and then it is extended into a two-stage stochastic problem which explicitly addresses the uncertainty in the pump performance models and minimizes an expected annualized cost. More details of the problem can be found in [28].

6.1.2 Sarawak Gas Production Subsystem Design Problem

This problem comes from a real industrial system, the Sarawak Gas Production System (SGPS) [29]. In [30], optimal operation of a subsystem of the SGPS is studied. This problem is extended into an integrated design and operation problem under uncertainty. The uncertainties in the system include gas product demand, gas product price and the pressure-flow relationship in a pipeline. The objective of the optimization is to maximize the expected net present value while satisfying the demand constraints at the end node over the scenarios of consideration. More details of the problem can be found in [28].

6.1.3 Capacity Planning Problem

This is a capacity planning problem in continuous pharmaceutical manufacturing under clinical trials uncertainty [31, 32]. The problem considers the building of new facilities or expansion of existing facilities for future manufacturing of new drugs that are still in the clinical trial stages. The problem is modelled as a two-stage stochastic MILP to achieve the best expected profits. The first-stage decisions are to determine the timing of facility development and the discrete sizes of the facilities to be developed before product launch. The second-stage decisions are to determine the discrete sizes of the facilities to be developed after product launch and the operation of the facilities. More details of the problem can be found in [31, 32].

All the three case study problems are two-stage stochastic programs exhibiting the structure of Problem (P-SMIP-S) discussed in the previous section, in which the complicating variables are the first-stage decisions. In the first two problems, the second-stage decisions are all continuous and the nonconvexity of the problem comes from the nonconvex functions involved. In the third problem, the second-stage decisions involve both continuous and binary variables (which make set X_h nonconvex) but all functions involved are affine, so the overall problem is a MILP.

6.2 Implementation

The first two case study problems are nonconvex MINLPs, so BARON 9.0 was used to compare with NGBD for global solution. Convex underestimators for constructing the lower bounding problems in NGBD were generated through McCormick relaxation [5, 10] and auxiliary variables were introduced for generating smooth underestimators [8]. The problem instances were solved on a computer allocated a single 2.83 GHz CPU, 2 GB memory and running Linux Kernel. GAMS 23.4 [33] was used to formulate the problems, program the NGBD algorithm and interface the solvers for the subproblems. The NGBD method employed BARON 9.0 for solving nonconvex NLP subproblems, CPLEX 12.1 for LP/MILP subproblems and CONOPT 3 for convex NLP subproblems. BARON 9.0 itself employed CPLEX 12.1 as the LP solver and CONOPT 3 the local NLP solver.

The third case study problem is a MILP, so CPLEX 12.3 was used to compare with NGBD. The problem instances were solved on a computer with a 3.2 GHz Intel Xeon CPU, 12 GB memory, and Windows platform. GAMS 23.7 was used to formulate the problems, program the NGBD algorithm and interface the solvers for the subproblems. The NGBD method employed CPLEX 12.3 to solve LP and MILP subproblems.

The relative and absolute termination criteria were set to be 10^{-3} for all the three case study problems. The NGBD solution times reported here are the total times reported by the GAMS solvers for solving all the subproblems.

6.3 Results and Discussion

The results for the case study problems with different numbers of scenarios are summarized in Tables 1, 2 and 3. It can be seen that when the number of scenarios is small, NGBD may not be faster than the commercial solvers; it is slower than BARON 9.0 for the pump network problem with 1 scenario and slower than CPLEX 12.3 for the capacity planning problem with 16 scenarios. However, when more scenarios are considered, the solution time with BARON 9.0 or CPLEX 12.3 increases rapidly with the number of scenarios while the solution time with NGBD increases slowly. For the pump network problem with 125 scenarios and the SGPS problem with 27 scenarios, BARON cannot return a global solution within 10,000 s, while NGBD can within several minutes. For the capacity planning problem with 4,094 scenarios, CPLEX 12.3 cannot return an optimal solution within 20,000 s, while NGBD can within 15 min. For all the three case study problems, NGBD is at least an order of magnitude faster than the commercial solvers in most cases. In addition, computational results of two large problem instances in the tables are notable. One is the SGPS problem with 1,331 scenarios. This nonconvex MINLP has nearly 150,000 variables and it was solved to global optimality by NGBD within only 80 min. The other is the capacity planning problem with 65,536 scenarios. This

Table 1 Results for the pump network configuration problem

Number of scenarios	1	27	125	343	729	1,331
Number of continuous variables	38	1,026	4,750	13,034	27,702	50,578
Number of binary variables	18	18	18	18	18	18
Total time with BARON 9.0 (s)	0.5	28.9	–[a]	–	–	–
Total time with NGBD (s)	7.7	60.9	328.8	754.1	1,497.0	2,794.8
Detailed results for NGBD						
Time for PBP&FP (s)	0.4	5.7	9.9	31.1	150.2	304.1
Time for RMP&FRMP (s)	1.8	1.0	0.7	0.7	1.1	1.1
Time for PP (s)	5.4	54.2	3,18.3	722.4	1,345.7	2,489.6
UBD at termination (FIM)[b]	128.9	136.6	136.3	145.3	145.3	145.3
LBD at termination (FIM)	$+\infty$[c]	$+\infty$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
Integer realizations visited by PBP	100	72	77	75	77	80
Integer realizations visited by PP	41	21	20	19	19	19

[a] No ε-optimal solution returned within 10,000 s
[b] *UBD* at termination is returned as the ε-optimal objective value of the problem
[c] Represented by a large number (10^{10}) in NGBD, which indicates that PBP has visited all feasible integer realizations

Table 2 Results for the SGPS problem

Number of scenarios	1	27	125	343	729	1,331
Number of continuous variables	110	2,970	13,750	37,730	80,190	146,410
Number of binary variables	20	20	20	20	20	20
Total time with BARON 9.0 (s)	123.2	–[a]	–	–	–	–
Total time with NGBD (s)	0.6	78.7	372.0	1,081.2	2,253.1	4,234.8
Detailed results for NGBD						
Time for PBP&FP (s)	0.1	5.5	31.8	82.3	175.4	263.4
Time for RMP&FRMP (s)	0.1	0.8	0.8	0.9	0.8	0.6
Time for PP (s)	0.5	72.4	339.3	998.1	2,076.9	3,970.8
UBD at termination (Billion \$)[b]	−7.209	−7.189	−7.187	−7.188	−7.188	−7.188
LBD at termination (Billion \$)	−7.209	−7.189	−7.189	−7.189	−7.189	−7.189
Integer realizations visited by PBP_h	14	70	70	71	70	68
Integer realizations visited by PP_h	1	16	16	16	16	16

[a] No ε-optimal solution returned within 10,000 s
[b] *UBD* at termination is returned as the ε-optimal objective value of the problem

MILP has about 65 million binary variables, 240 million continuous variables, 250 million constraints, and it was solved by NGBD within only 6 h.

The computational advantage of NGBD comes from the fact that (1) the total number of NGBD iterations does not change significantly with the number of scenarios (as indicated by integer realizations visited by PP and PBP shown in the tables), and (2) the NGBD solution time is dominated by Problem (PP) and Problem (PBP) that are decomposable over the scenarios. So the NGBD solution time increases roughly linearly with the number of scenarios, while BARON 9.0 and

Table 3 Results for the capacity planning problem

Number of scenarios	16	256	4,096	16,384	65,536
Number of first-stage binary variables	124	124	124	124	124
Number of second-stage binary variables (per scenario)	248	496	744	868	992
Number of continuous variables (per scenario)	1,054	1,922	2,790	3,224	3,658
Total time with CPLEX 12.3 (s)	1.8	1,127.5	–[a]	–	–
Total time with NGBD (s)	2.4	39.3	878.1	4,103.7	19,235.5
Detailed results for NGBD					
Time for PBP&FP (s)	1.3	6.5	128.4	611.4	2,797.8
Time for RMP&FRMP (s)	0.2	0.1	0.1	0.1	0.3
Time for PP (s)	0.9	32.7	749.6	3,492.2	16,437.4
UBD at termination (Billion $)[b]	−32.198	−63.803	−83.410	−88.953	−99.664
LBD at termination (Billion $)	−32.216	−63.789	−83.365	−88.890	−99.579
Integer realizations visited by PBP	5	2	2	2	2
Integer realizations visited by PP	1	1	1	1	1

[a] No ε-optimal solution returned within 20,000 s
[b] *UBD* at termination is returned as the ε-optimal objective value of the problem

CPLEX 12.3 cannot achieve this empirical linear complexity. In addition, Problem (PP) is the most difficult subproblem in NGBD, and the solution of it is postponed as much as possible. It can be seen from the three tables that the number of integer realizations visited by Problem (PP) (i.e. the number of Problem (PP) solved) is smaller than that by Problem (PBP). This is because the solution of Problem (PBP) helps to eliminate some integer realizations that will never lead to a global optimum (as indicated by Proposition 5).

Tables 1, 2 and 3 also show *UBD* and *LBD* at the NGBD termination for all the cases. *UBD* stands for an upper bound as well as an ε-optimal objective value of Problem (P). *LBD* stands for the lower bound of the Problem (P) excluding all the visited integer realizations. *LBD* can be significantly larger than *UBD* at the termination and $LBD = +\infty$ implies that all the feasible integer realizations have been visited.

7 Conclusions

By using the concepts of problem manipulation and solution strategy used in BD/GBD and the notion of convex/continuous relaxations, NGBD solves Problem (P) by solving a sequence of subproblems that lead to an ε-optimal solution in a finite number of steps. NGBD has tremendous computational advantages when the subproblems are much easier to solve than the original problem, such as for the stochastic programs (P-SMIP) and (P-SMIP-S). Case studies of several stochastic MINLP/MILP problems from industry demonstrate that NGBD is more efficient than state-of-the-art commercial solvers by at least one order of

magnitude. In addition, the solution time with NGBD increases moderately with the number of scenarios involved in the stochastic program.

As the lower bounding problem serves as a surrogate of the original problem, its closeness to the original problem has a dominant impact on the convergence rate of NGBD. The efficiency of NGBD can be improved by introducing tighter convex and continuous relaxations for constructing the lower bounding problem. It has been shown that the efficiency of NGBD can be improved by an order of magnitude through the integration of a piecewise convex relaxation framework [34]. Since most of the subproblems in a NGBD iteration can be solved without exchanging information among them, the performance of NGBD can also be improved by exploitation of a parallel computation architecture. Some preliminary study has demonstrated that a parallel NGBD algorithm can reduce the NGBD solution time by several times on an 8-core processor [35].

The NGBD method proposed in this work only deals with binary complicating variables (i.e. y in Problem (P) is binary). An interesting future work is to extend the method to continuous/mixed integer and continuous complicating variables. While the convergence of the current NGBD method is established on the finite number of realizations for the complicating variables, the extension will require a new mechanism to guarantee convergence to an ε-optimal solution.

References

1. Benders, J.F.: Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik **4**, 238–252 (1962)
2. Geoffrion, A.M.: Generalized Benders decomposition. J. Optim. Theory Appl. **10**(4), 237–260 (1972)
3. Geoffrion, A.M.: Elements of large-scale mathematical programming: Part I: concepts. Manag. Sci. **16**(11), 652–675 (1970)
4. Geoffrion, A.M.: Elements of large-scale mathematical programming: Part II: synthesis of algorithms and bibliography. Manag. Sci. **16**(11), 652–675 (1970)
5. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I - convex underestimating problems. Math. Program. **10**, 147–175 (1976)
6. Tawarmalani, M., Sahinidis, N.V.: Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. Math. Program. **99**, 563–591 (2004)
7. Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, α-BB, for general twice-differentiable constrained NLPs – I. Theoretical advances. Comput. Chem. Eng. **22**(9), 1137–1158 (1998)
8. Gatzke, E.P., Tolsma, J.E., Barton, P.I.: Construction of convex relaxations using automated code generation technique. Optim. Eng. **3**, 305–326 (2002)
9. Fletcher, R., Leyffer, S.: Solving mixed integer nonlinear programs by outer approximation. Math. Program. **66**, 327–349 (1994)
10. Mitsos, A., Chachuat, B., Barton, P.I.: McCormick-based relaxations of algorithms. SIAM J. Optim. **20**(2), 573–601 (2009)
11. IBM. IBM ILOG CPLEX Optimization Studio. http://www-03.ibm.com/software/products/us/en/ibmilogcpleoptistud/ Accessed April 13, 2014
12. ARKI Consulting and Development. http://www.gams.com/docs/conopt3.pdf Accessed April 13, 2014

13. Gill, P.E., Murray, W., Saunders, M.A.: SNOPT: an SQP algorithm for large-scale constrained optimization. SIAM Rev. **47**, 99–131 (2005)
14. Lemaréchal, C., Sagastizábal, C.: Variable metric bundle methods: from conceptual to implementable forms. Math. Program. **76**, 393–410 (1997)
15. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Cambridge (1999)
16. Geoffrion, A.M.: Duality in nonlinear programming: a simplified applications-oriented development. SIAM Rev. **13**(1), 1–37 (1971)
17. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms, 2nd edn. Wiley, New York (1993)
18. Balas, E., Jeroslow, R.: Canonical cuts on the unit hypercube. SIAM J. Appl. Math. **23**(1), 61–69 (1972)
19. Grossmann, I.E., Raman, R., Kalvelagen, E.: DICOPT User's Manual. http://www.gams.com/dd/docs/solvers/dicopt.pdf Accessed April 13, 2014
20. Grossmann, I.E.: Review of nonlinear mixed-integer and disjunctive programming techniques. Optim. Eng. **3**, 227–252 (2002)
21. Floudas, C.A.: Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications. Oxford University Press, Oxford (1995)
22. Bieglerm, L.T., Grossmann, I.E.: Retrospective on optimization. Comput. Chem. Eng. **28**, 1169–1192 (2004)
23. Birge, J.R., Louveaux, F.: Introduction to Stochastic Programming. Springer, New York (1997)
24. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2009)
25. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: Global optimization of mixed-integer nonlinear problems. AIChE J. **46**(9), 1769–1797 (2000)
26. Kesavan, P., Allgor, R.J., Gatzke, E.P., Barton, P.I.: Outer approximation algorithms for separable nonconvex mixed-integer nonlinear programs. Math. Program. Ser. A **100**, 517–535 (2004)
27. Westerlund, T., Pettersson, F., Grossmann, I.E.: Optimization of pump configurations as a MINLP problem. Comput. Chem. Eng. **18**(9), 845–858 (1994)
28. Li, X., Tomasgard, A., Barton, P.I.: Nonconvex generalized Benders decomposition for stochastic separable mixed-integer nonlinear programs. J. Optim. Theory Appl. **151**, 425–454 (2011)
29. Selot, A., Kuok, L.K., Robinson, M., Mason, T.L., Barton, P.I.: A short-term operational planning model for natural gas production systems. AIChE J. **54**(2), 495–515 (2008)
30. Selot, A.: Short-Term Supply Chain Management in Upstream Natural Gas Systems. Ph.D. thesis, Massachusetts Institute of Technology (2009)
31. Sundaramoorthy, A., Evans, J.M.B., Barton, P.I.: Capacity planning under clinical trials uncertainty in continuous pharmaceutical manufacturing, 1: mathematical framework. Ind. Eng. Chem. Res. **51**, 13692–13702 (2012)
32. Sundaramoorthy, A., Li, X., Evans, J.M.B., Barton, P.I.: Capacity planning under clinical trials uncertainty in continuous pharmaceutical manufacturing, 2: solution method. Ind. Eng. Chem. Res. **51**, 13703–13711 (2012)
33. GAMS. General Algebraic and Modeling System. http://www.gams.com/ Accessed April 13, 2014
34. Li, X., Chen, Y., Barton, P.I.: Nonconvex generalized Benders decomposition with piecewise convex relaxations for global optimization of integrated process design and operation problems. Ind. Eng. Chem. Res. **51**, 7287–7299 (2012)
35. Li, X.: Parallel nonconvex generalized Benders decomposition for natural gas production network planning under uncertainty. Comput. Chem. Eng. **55**, 97–108 (2013)

On Nonsmooth Multiobjective Optimality Conditions with Generalized Convexities

Marko M. Mäkelä, Ville-Pekka Eronen, and Napsu Karmitsa

1 Introduction

Optimality conditions are an essential part of mathematical optimization theory, affecting heavily, for example to the method development both in local and global optimization [21]. When constructing optimality conditions convexity has been the most important concept during the last decades. Recently there have been numerous attempts to generalize the concept of convexity in order to weaken the assumptions of the attained results (see, e.g., [1, 4, 8, 13, 16, 26, 30, 32]).

Different kinds of generalized convexities have proved to be the main tool when constructing optimality conditions, particularly sufficient conditions. There exist a wide amount of papers published for smooth (continuously differentiable) single-objective case (see [26] and references therein). For nonsmooth (not continuously differentiable) problems there is an additional degree of freedom in choosing the way how to deal with the nonsmoothness. There are many different generalized directional derivatives to do this. For example, necessary and sufficient conditions for nonsmooth single-objective optimization by using the Dini directional derivatives were developed in [8]. These results were extended for nonsmooth multiobjective problems in [3].

Another degree of freedom is how to generalize convexity. In [22] sufficient conditions for nonsmooth multiobjective programs were derived by using the $(\mathcal{F}, \rho)$-convexity defined by Preda [27] and its extension for nonsmooth case defined by Bhatia and Jain [4]. Recently, the concept of invexity defined by Hanson [9] has become a very popular research concept. It was used to formulate necessary and sufficient conditions for differentiable multiobjective case in [25], for arcwise connected functions in [5] and for nonsmooth multiobjective programming in [6, 12, 23, 24].

M.M. Mäkelä (✉) • V.-P. Eronen • N. Karmitsa
Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland
e-mail: makela@utu.fi; vpoero@utu.fi; napsu@karmitsa.fi

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_17,

In this paper, we present optimality conditions for nonsmooth multiobjective problems with locally Lipschitz continuous functions. Three types of constraint sets are considered. First, we discuss general set constraint, then, only inequality constraints and, finally, both inequality and equality constraints. To deal with the nonsmoothness we use the Clarke subdifferential as a generalization to gradient. For the necessary condition we require that certain constraint qualifications holds. For sufficient conditions we use f°-pseudo- and quasiconvexities [13] as a generalization to convexity. The necessary conditions with inequality constraints rely mainly on [14]. In [12] a sufficient condition was presented which differs from ours mainly by the formulation of object function. Moreover, f°-quasiconcave inequality constraints were not considered in [12].

Nonsmooth problems with locally Lipschitz continuous functions were considered also in [24, 29, 31]. Our presentation differs from [24, 29] by constraint qualifications and the formulation of KKT conditions. Also, in [24] the necessary optimality condition relied on a theorem, which required the subdifferential of equality constraint functions to be a singleton. For the sufficient conditions we need generalized pseudo- and quasiconvexities. Contrary to [24], the invexity and its generalizations are not used here. In [31] general constraint set was used in the derivation of conditions for weak Pareto optimality. Our presentation has different, more specific formulation for these conditions.

This article is organized as follows. In Sect. 2 we recall some basic tools from nonsmooth analysis. In Sect. 3 results concerning generalized pseudo- and quasiconvexity are presented. In Sect. 4 we present Karush–Kuhn–Tucker (KKT) type necessary and sufficient conditions of weak Pareto optimality for nonsmooth multiobjective optimization problems with different constraint sets. Finally, some concluding remarks are given in Sect. 5.

2 Nonsmooth Analysis

In this section we collect some notions and results from nonsmooth analysis. Most of the proofs of this section are omitted, since they can be found, for example in [7, 15, 17]. Nevertheless, we start by recalling the notion of convexity and Lipschitz continuity. The function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$ we have

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y).$$

A function is *locally Lipschitz continuous at a point* $x \in \mathbb{R}^n$ if there exist scalars $K > 0$ and $\delta > 0$ such that

$$|f(y) - f(z)| \leq K\|y - z\| \quad \text{for all } y, z \in B(x;\delta),$$

where $B(x;\delta) \subset \mathbb{R}^n$ is an open ball with center x and radius δ. If a function is locally Lipschitz continuous at every point then it is called *locally Lipschitz continuous*. Note that both convex and smooth functions are always locally Lipschitz continuous

(see, e.g. [7]). In what follows the considered functions are assumed to be locally Lipschitz continuous.

Definition 1. [7] Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at $x \in S \subset \mathbb{R}^n$. The *Clarke generalized directional derivative* of f at x in the direction of $d \in \mathbb{R}^n$ is defined by

$$f^\circ(x;d) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+td)-f(y)}{t}$$

and the *Clarke subdifferential* of f at x by

$$\partial f(x) = \{\xi \in \mathbb{R}^n \mid f^\circ(x;d) \geq \xi^T d \text{ for all } d \in \mathbb{R}^n\}.$$

Each element $\xi \in \partial f(x)$ is called a *subgradient* of f at x.

Note that the Clarke generalized directional derivative $f^\circ(x;d)$ always exists for a locally Lipschitz continuous function f. Furthermore, if f is smooth $\partial f(x)$ reduces to $\partial f(x) = \{\nabla f(x)\}$ and if f is convex $\partial f(x)$ coincides with the classical subdifferential of convex function (cf. [28]), in other words the set of $\xi \in \mathbb{R}^n$ satisfying

$$f(y) \geq f(x) + \xi^T (y-x) \quad \text{for all } y \in \mathbb{R}^n.$$

The following properties derived in [7, 17] are characteristic to the generalized directional derivative and subdifferential.

Theorem 1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous at $x \in \mathbb{R}^n$, then*

a) $d \mapsto f^\circ(x;d)$ is positively homogeneous, subadditive and Lipschitz continuous function such that $f^\circ(x;-d) = (-f)^\circ(x;d)$.
b) $\partial f(x)$ is a nonempty, convex and compact set such that $\partial(-f)(x) = -\partial f(x)$.
c) $f^\circ(x;d) = \max\{\xi^T d \mid \xi \in \partial f(x)\}$ for all $d \in \mathbb{R}^n$.
d) $f^\circ(x;d)$ is upper semicontinuous as a function of (x,d).

In order to maintain equalities instead of inclusions in subderivation rules we need the following regularity property.

Definition 2. The function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *subdifferentially regular* at $x \in \mathbb{R}^n$ if it is locally Lipschitz continuous at x and for all $d \in \mathbb{R}^n$ the classical directional derivative

$$f'(x;d) = \lim_{t \downarrow 0} \frac{f(x+td)-f(x)}{t}$$

exists and $f'(x;d) = f^\circ(x;d)$.

Note, that the equality $f'(x;d) = f^\circ(x;d)$ is not necessarily valid in general even if $f'(x;d)$ exists. This is the case, for instance, with concave nonsmooth functions. However, convexity, as well as smoothness implies subdifferential regularity [7]. Furthermore, it is easy to show that a necessary and sufficient condition for convexity is that for all $x, y \in \mathbb{R}^n$ we have

$$f(y) - f(x) \geq f^{\circ}(x; y-x) = f'(x; y-x). \tag{1}$$

Next we present two subderivation rules of composite functions, namely the finite maximum and positive linear combination of subdifferentially regular functions.

Theorem 2. *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at x for all $i = 1, \ldots, m$. Then the function*

$$f(x) = \max\{f_i(x) \mid i = 1, \ldots, m\}$$

is locally Lipschitz continuous at x and

$$\partial f(x) \subset \operatorname{conv}\{\partial f_i(x) \mid f_i(x) = f(x),\ i = 1, \ldots, m\}, \tag{2}$$

where conv *denotes the convex hull of a set. In addition, if f_i is subdifferentially regular at x for all $i = 1, \ldots, m$, then f is also subdifferentially regular at x and equality holds in (2).*

Theorem 3. *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at x and $\lambda_i \in \mathbb{R}$ for all $i = 1, \ldots, m$. Then the function*

$$f(x) = \sum_{i=1}^{m} \lambda_i f_i(x)$$

is locally Lipschitz continuous at x and

$$\partial f(x) \subset \sum_{i=1}^{m} \lambda_i \partial f_i(x). \tag{3}$$

In addition, if f_i is subdifferentially regular at x and $\lambda_i \geq 0$ for all $i = 1, \ldots, m$, then f is also subdifferentially regular at x and equality holds in (3).

In the following, for a given set $S \subset \mathbb{R}^n$ we denote by d_S the *distance function* of S, that is,

$$d_S(x) = \inf\{\|x - s\| \mid s \in S\}. \tag{4}$$

If S is nonempty, then d_S is locally Lipschitz continuous with the constant one [7]. The closure of a set S is denoted $\operatorname{cl} S$. By the Weierstrass Theorem we may replace inf by min in (4) if $S \neq \emptyset$ is closed. Note also that $d_S(x) = 0$ if $x \in \operatorname{cl} S$.

A set $C \subset \mathbb{R}^n$ is a *cone* if $\lambda x \in C$ for all $\lambda \geq 0$ and $x \in C$. We also denote

$$\operatorname{ray} S = \{\lambda s \mid \lambda \geq 0, s \in S\} \quad \text{and} \quad \operatorname{cone} S = \operatorname{ray} \operatorname{conv} S.$$

In other words $\operatorname{ray} S$ is the smallest cone containing S and $\operatorname{cone} S$ is the smallest convex cone containing S.

Definition 3. The *Clarke normal cone* of the set $S \subset \mathbb{R}^n$ at $x \in S$ is given by the formula

$$N_S(x) = \operatorname{cl} \operatorname{ray} \partial d_S(x).$$

It is easy to derive that $N_S(x)$ is a closed convex cone (see, e.g. [7]). In convex case the normal cone can be expressed by the following simple inequality condition.

Theorem 4. *If S is a convex set and $x \in S$, then*

$$N_S(x) = \{z \in \mathbb{R}^n \mid z^T(y-x) \leq 0 \quad \textit{for all} \quad y \in S\}.$$

The *contingent cone*, *polar cone* and *strict polar cone* of set $S \in \mathbb{R}^n$ at point x are defined respectively as

$$\begin{aligned} T_S(x) &= \{d \in \mathbb{R}^n \mid \text{there exist } t_i \downarrow 0 \text{ and } d_i \to d \text{ with } x + t_i d_i \in S\} \\ S^{\leq} &= \{d \in \mathbb{R}^n \mid s^T d \leq 0, \text{ for all } s \in S\} \\ S^{<} &= \{d \in \mathbb{R}^n \mid s^T d < 0, \text{ for all } s \in S\}. \end{aligned}$$

Next we will present some basic results that are useful in Sect. 4. The proofs of the following two lemmas can be found in [17].

Lemma 1. *Let $S_i \subset \mathbb{R}^n$, $i = 1, 2, \ldots, m$ be convex sets and $C \subset \mathbb{R}^n$ be a convex cone. Assume that all the sets are nonempty. Then*

a) $\operatorname{conv} \bigcup_{i=1}^m S_i = \{\sum_{i=1}^m \lambda_i s_i \mid s_i \in S_i, \lambda_i \geq 0, \sum_{i=1}^m \lambda_i = 1\}$.
b) $\operatorname{cone} \bigcup_{i=1}^m S_i = \{\sum_{i=1}^m \mu_i s_i \mid s_i \in S_i, \mu_i \geq 0\} = \sum_{i=1}^m \operatorname{ray} S_i$.
c) $\bigcup_{i=1}^m (S_i + C) = \bigcup_{i=1}^m S_i + C$.
d) $\operatorname{conv} \bigcup_{i=1}^m (S_i + C) = \operatorname{conv} \bigcup_{i=1}^m S_i + C$.

Lemma 2. *Let $S_i \subset \mathbb{R}^n$, $i = 1, 2, \ldots, m$ be convex compact sets. Then* $\operatorname{conv} \bigcup_{i=1}^m S_i$ *is a compact set.*

To the end of this section we recall the classical necessary and sufficient nonsmooth unconstrained optimality condition.

Theorem 5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz continuous at x^*. If f attains its local minimum at x^*, then*

$$0 \in \partial f(x^*).$$

If, in addition, f is convex, then the above condition is sufficient for x^* to be a global minimum.

3 Generalized Convexities

In this section we present some generalizations of convexity, namely f°-pseudoconvexity, quasiconvexity and f°-quasiconvexity, that are used later. We also define f°-quasiconcavity. A famous generalization of convexity is pseudoconvexity introduced in [18]. For a pseudoconvex function f a point $x \in \mathbb{R}$ is a global minimum if and only if $\nabla f(x) = 0$. The classical pseudoconvexity requires the

function to be smooth and, thus, it is not suitable for our purposes. However, with some modifications pseudoconvexity can be defined for nonsmooth functions as well. One such definition is presented in [10]. This definition requires the function to be merely locally Lipschitz continuous.

Definition 4. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *f°-pseudoconvex*, if it is locally Lipschitz continuous and for all $x, y \in \mathbb{R}^n$

$$f(y) < f(x) \qquad \text{implies} \qquad f^\circ(x; y-x) < 0.$$

Note that due to (1) a convex function is always f°-pseudoconvex. Sometimes the reasoning chain in the definition of f°-pseudoconvexity needs to be converted.

Lemma 3. *A locally Lipschitz continuous function f is f°-pseudoconvex, if and only if for all $x, y \in \mathbb{R}^n$*

$$f^\circ(x; y-x) \geq 0 \qquad \textit{implies} \qquad f(y) \geq f(x).$$

Proof. Follows directly from the definition of f°-pseudoconvexity. □

The important sufficient extremum property of pseudoconvexity remains also for f°-pseudoconvexity.

Theorem 6. *An f°-pseudoconvex function f attains its global minimum at x^*, if and only if*

$$0 \in \partial f(x^*).$$

Proof. If f attains its global minimum at x^*, then by Theorem 5 we have $0 \in \partial f(x^*)$. On the other hand, if $0 \in \partial f(x^*)$ and $y \in \mathbb{R}^n$, then by Definition 1 we have

$$f^\circ(x^*; y-x^*) \geq 0^T (y-x^*) = 0$$

and, thus by Lemma 3 we have

$$f(y) \geq f(x^*).$$

□

Note that it follows from Theorem 6 that pseudoconvexity implies f°-pseudoconvexity.

The notion of quasiconvexity is the most widely used generalization of convexity and, thus, there exist various equivalent definitions and characterizations. Next we recall the most commonly used definition of quasiconvexity (see [1]).

Definition 5. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *quasiconvex*, if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\}.$$

Note that, unlike pseudoconvexity, the previous definition of quasiconvexity does not require differentiability nor continuity. We give also a useful result concerning a finite maximum of quasiconvex functions.

Theorem 7. *Let $f_i : \mathbb{R}^n \to \mathbb{R}$ be quasiconvex at x for all $i = 1,\dots,m$. Then the function*

$$f(x) = \max\{f_i(x) \mid i = 1,\dots,m\}$$

is also quasiconvex.

Proof. Follows directly from the definition of quasiconvexity. □

Analogously to the Definition 4 we can define the corresponding generalized concept, which is a special case of h-quasiconvexity defined by Komlósi [13] when h is the Clarke generalized directional derivative.

Definition 6. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *f°-quasiconvex*, if it is locally Lipschitz continuous and for all $x, y \in \mathbb{R}^n$

$$f(y) \le f(x) \qquad \text{implies} \qquad f^\circ(x; y-x) \le 0.$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is f°-quasiconcave if $-f$ is f°-quasiconvex.

Theorem 8. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is f°-quasiconcave if it is locally Lipschitz continuous and for all $x, y \in \mathbb{R}^n$*

$$f(y) \le f(x) \qquad \textit{implies} \qquad f^\circ(y; y-x) \le 0.$$

Proof. By Definition 6 we have

$$-f(x) \le -f(y) \qquad \text{implies} \qquad (-f)^\circ(y; x-y) \le 0.$$

Using Theorem 1(a) we obtain

$$f(y) \le f(x) \qquad \text{implies} \qquad f^\circ(y; y-x) \le 0$$

which proves the theorem. □

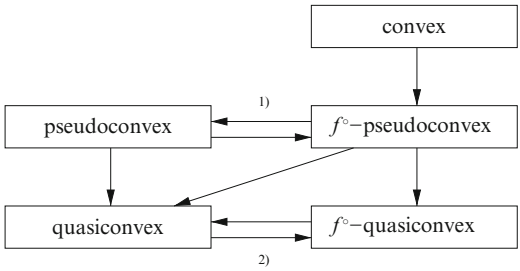
Next, we give few results concerning relations between the previously presented generalized convexities. The proofs for these results can be found in [16].

Theorem 9. *If $f : \mathbb{R}^n \to \mathbb{R}$ is f°-pseudoconvex, then f is f°-quasiconvex and quasiconvex.*

Theorem 10. *If $f : \mathbb{R}^n \to \mathbb{R}$ is f°-quasiconvex, then f is quasiconvex.*

Theorem 11. *If $f : \mathbb{R}^n \to \mathbb{R}$ is subdifferentially regular and quasiconvex then f is f°-quasiconvex.*

Figure 1 illustrates the relations between different convexities.



[1] demands continuous differentiability

[2] demands subdifferential regularity

Fig. 1 Relations between different types of generalized convexities

4 Optimality Conditions for Nonsmooth Multiobjective Problem

In this section we present some necessary and sufficient optimality conditions for multiobjective optimization.

Consider first a general multiobjective optimization problem

$$\begin{cases} \text{minimize} & \{f_1(x),\ldots,f_q(x)\} \\ \text{subject to} & x \in S, \end{cases} \tag{5}$$

where $f_k : \mathbb{R}^n \to \mathbb{R}$ for $k = 1,2,\ldots,q$ are locally Lipschitz continuous functions and $S \subset \mathbb{R}^n$ is an arbitrary nonempty set. Denote

$$F(x) = \bigcup_{k\in Q} \partial f_k(x) \quad \text{and} \quad Q = \{1,2,\ldots,q\}.$$

We start the consideration by defining the notion of optimality for the multiobjective problem (5).

Definition 7. A vector x^* is said to be a *global Pareto optimum* of (5), if there does not exist $x \in S$ such, that $f_k(x) \le f_k(x^*)$ for all $k = 1,\ldots,q$ and $f_l(x) < f_l(x^*)$ for some l. Vector x^* is said to be a *global weak Pareto optimum* of (5), if there does not exist $x \in S$ such that $f_k(x) < f_k(x^*)$ for all $k = 1,\ldots,q$. Vector x^* is a *local (weak) Pareto optimum* of (5), if there exists $\delta > 0$ such that x^* is a global (weak) Pareto optimum on $B(x^*;\delta) \cap S$.

Next we will present some optimality conditions of problem (5) in terms of cones. We also consider the unconstrained case, that is, when $S = \mathbb{R}^n$. We begin the considerations with the following lemma which can be found in [14, Lemma 4.2].

Lemma 4. *If x^* is a local weak Pareto optimum of problem (5), then $F^<(x^*) \cap T_S(x^*) = \emptyset$.*

Proof. Let x^* be a local weak Pareto optimum. Then, there exists $\varepsilon > 0$ such that for every $y \in S \cap B(x^*, \varepsilon)$ there exists $k \in Q$ such that inequality $f_k(y) \geq f_k(x^*)$ holds. Let $d \in T_S(x^*)$ be arbitrary. Then, there exist sequences (d_i) and (t_i) such that $d_i \to d$, $t_i \downarrow 0$ and $x^* + t_i d_i \in S$ for all $i \in \mathbb{N}$. Also, there exists an index I_1 such that $x^* + t_i d_i \in S \cap B(x^*, \varepsilon)$ for all $i > I_1$. Then for every $i > I_1$ there exists k_i such that $f_{k_i}(x^* + t_i d_i) \geq f_{k_i}(x^*)$. Since the set Q is finite, there exists $\bar{k} \in Q$ and subsequences $(d_{i_j}) \subset (d_i)$ and $(t_{i_j}) \subset (t_i)$ such that

$$f_{\bar{k}}(x^* + t_{i_j} d_{i_j}) \geq f_{\bar{k}}(x^*) \tag{6}$$

for all i_j with $j \in \mathbb{N}$ large enough. Denote $I_2 = \{i_j \mid i_j > I_1, j \in \mathbb{N}\}$. The Mean-Value Theorem (see, e.g., [7]) implies that for all $\bar{i} \in I_2$ there exists $\tilde{t}_{\bar{i}} \in (0, t_{\bar{i}})$ such that

$$f_{\bar{k}}(x^* + t_{\bar{i}} d_{\bar{i}}) - f_{\bar{k}}(x^*) \in \partial f_{\bar{k}}(x^* + \tilde{t}_{\bar{i}} d_{\bar{i}})^T t_{\bar{i}} d_{\bar{i}}. \tag{7}$$

From the definition of generalized directional derivative (Definition 1), (6) and (7) we obtain

$$f^\circ_{\bar{k}}(x^* + \tilde{t}_{\bar{i}} d_{\bar{i}}; d_{\bar{i}}) = \max_{\xi \in \partial f_{\bar{k}}(x^* + \tilde{t}_{\bar{i}} d_{\bar{i}})} \xi^T d_{\bar{i}} \geq \frac{1}{t_{\bar{i}}} (f_{\bar{k}}(x^* + t_{\bar{i}} d_{\bar{i}}) - f_{\bar{k}}(x^*)) \geq 0.$$

Thus, for all $\bar{i} \in I_2$ we have $f^\circ_{\bar{k}}(x^* + \tilde{t}_{\bar{i}} d_{\bar{i}}; d_{\bar{i}}) \geq 0$. Since $d_{\bar{i}} \to d$ and $x^* + \tilde{t}_{\bar{i}} d_{\bar{i}} \to x^*$ the upper semicontinuity of function $f^\circ_{\bar{k}}$ [Theorem 1, (d)] implies

$$f^\circ_{\bar{k}}(x^*, d) \geq \lim_{\bar{i} \to \infty} f^\circ_{\bar{k}}(x^* + \tilde{t}_{\bar{i}} d_{\bar{i}}; d_{\bar{i}}) \geq 0.$$

Thus, there exists $\xi \in \partial f_{\bar{k}}(x^*) \subset F(x^*)$ such that $\xi^T d \geq 0$ implying $d \notin F^<(x^*)$. □

Next, we will present a result for the unconstrained case. The result is analogous to Theorem 5.

Theorem 12. *Let f_k be locally Lipschitz continuous for all $k \in Q$ and $S = \mathbb{R}^n$. If x^* is a local weak Pareto optimum of problem (5), then*

$$0 \in \operatorname{conv} F(x^*)$$

Proof. Since $S = \mathbb{R}^n$ we have $T_S(x^*) = \mathbb{R}^n$ as well. Then by Lemma 4 we have $F^<(x^*) = \emptyset$. Hence, for any $d \in \mathbb{R}^n$ there exists $\xi \in F(x^*) \subset \operatorname{conv} F(x^*)$ such that

$$d^T \xi \geq 0. \tag{8}$$

Suppose that $0 \notin \operatorname{conv} F(x^*)$. Since the sets $\operatorname{conv} F(x^*)$ and $\{0\}$ are closed convex sets, there exists $d \in \mathbb{R}^n$ and $a \in \mathbb{R}$ such that

$$0 = d^T 0 \geq a \quad \text{and} \quad d^T \xi < a \quad \text{for all } \xi \in \operatorname{conv} F(x^*)$$

according to the Separation Theorem (see, e.g. [2]). From the first inequality we see that $a \leq 0$. Then the second inequality contradicts with inequality (8). Hence, $0 \in \operatorname{conv} F(x^*)$. □

In the following we shall present the necessary optimality condition of problem (5) in terms of Clarke normal cone. The proof is quite similar to the proof for single objective case in [15, pp. 72–73]. Before the condition we will present a useful lemma.

Lemma 5. *If x^* is a local weak Pareto optimum of problem (5), then it is local weak Pareto optimum of unconstrained problem*

$$\min_{x \in \mathbb{R}^n} \{f_1(x) + Kd_S(x), f_2(x) + Kd_S(x), \ldots, f_q(x) + Kd_S(x)\}, \tag{9}$$

where $K = \max\{K_1, K_2, \ldots, K_q\}$ and K_k is the Lipschitz constant of function f_k at point x^*.

Proof. From the definition of K and local weak Pareto optimality we see that there exists $\varepsilon > 0$ such that the Lipschitz condition holds for all f_k at $B(x^*; \varepsilon)$ and x^* is weak Pareto optimum at $B(x^*; \varepsilon) \cap S$. Suppose on the contrary that x^* is not a local weak Pareto optimum of problem (9). Then there exists $y \in B(x^*; \frac{\varepsilon}{2})$ such that

$$f_k(y) + Kd_S(y) < f_k(x^*) + Kd_S(x^*) = f_k(x^*) \quad \text{for all } k \in Q. \tag{10}$$

Suppose $y \in \operatorname{cl} S$. Then $Kd_S(y) = 0$ and by the continuity of f_k there exists $\delta > 0$ such that $f_k(z) < f_k(x^*)$ for all $k \in Q$ and $z \in B(y; \delta) \subset B(x^*; \frac{\varepsilon}{2})$. Since $y \in \operatorname{cl} S$ we have $S \cap B(y; \delta) \cap B(x^*; \frac{\varepsilon}{2}) \neq \emptyset$ and, thus, x^* is not a weak Pareto optimum of (5) in $S \cap B(x^*; \varepsilon)$ contradicting the assumption. Hence, $y \notin \operatorname{cl} S$ and $d_S(y) > 0$.

By the definition of $d_S(y)$ there exists $c \in \operatorname{cl} S$ such that $d_S(y) = \|y - c\|$. Furthermore,

$$\|c - y\| \leq \|x^* - y\| < \frac{\varepsilon}{2}.$$

Thus,

$$\|c - x^*\| \leq \|c - y\| + \|y - x^*\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

implying $c \in B(x^*; \varepsilon)$. By inequality (10) and local weak Pareto optimality of x^* there exists $k_1 \in Q$ such that

$$f_{k_1}(y) < f_{k_1}(x^*) \leq f_{k_1}(c).$$

Hence,

$$\left|f_{k_1}(x^*) - f_{k_1}(y)\right| \leq \left|f_{k_1}(c) - f_{k_1}(y)\right| \leq K\|y - c\| = Kd_S(y)$$

implying $f_{k_1}(x^*) \leq f_{k_1}(y) + Kd_S(y)$. This contradicts with inequality (10). Thus, x^* is a local weak Pareto optimum of problem (9). □

Finally, we can state the necessary optimality condition of problem (5) with arbitrary nonempty feasible set $S \subset \mathbb{R}^n$.

Theorem 13. *If x^* is a local weak Pareto minimum of* (5)*, then*

$$0 \in \operatorname{conv} F(x^*) + N_S(x^*). \tag{11}$$

Proof. By Lemma 5 x^* is a local weak Pareto optimum of unconstrained problem (9). Consider kth objective function of the unconstrained problem. By Theorem 3 we have

$$\partial(f_k(x) + Kd_S(x)) \subset \partial f_k(x) + K\partial d_S(x).$$

The Definition 3 of normal cone implies $K\partial d_S(x) \subset N_S(x)$. Since x^* is a local weak Pareto optimum of problem (9), Theorem 12 implies

$$0 \in \operatorname{conv} \bigcup_{k\in Q} \partial(f_k(x^*) + Kd_S(x^*)) \subset \operatorname{conv} \bigcup_{k\in Q} (\partial f_k(x^*) + N_S(x^*)).$$

By Lemma 1(d) we have

$$\operatorname{conv} \bigcup_{k\in Q} (\partial f_k(x^*) + N_S(x^*)) = \operatorname{conv} F(x^*) + N_S(x^*),$$

as desired. □

Since Pareto optimality implies weak Pareto optimality we get immediately the following consequence.

Corollary 1. *Condition* (11) *is also necessary for x^* to be a local Pareto optimum of* (5)*.*

To prove a sufficient condition for global optimality we need the assumptions that S is convex and f_k are f°-pseudoconvex for all $k \in Q$.

Theorem 14. *Let f_k be f°-pseudoconvex for all $k \in Q$ and S convex. Then $x^* \in S$ is a global weak Pareto minimum of* (5)*, if and only if*

$$0 \in \operatorname{conv} F(x^*) + N_S(x^*).$$

Proof. The necessity follows directly from Theorem 13. For sufficiency let $0 \in \operatorname{conv} F(x^*) + N_S(x^*)$. Then there exist $\xi_* \in \operatorname{conv} F(x^*)$ and $z_* \in N_S(x^*)$ such that $\xi_* = -z_*$. Then by Theorem 4 we have for all $x \in S$ that

$$0 \le -z_*^T(x - x^*) = \xi_*^T(x - x^*) = \sum_{k=1}^{q} \lambda_k \xi_k^T(x - x^*),$$

where $\lambda_k \ge 0$, $\xi_k \in \partial f_k(x^*)$ for all $k \in Q$ and $\sum_{k=1}^{q} \lambda_k = 1$. Thus, there exists k_1 such that $f_{k_1}^\circ(x^*, x - x^*) \ge \xi_{k_1}^T(x - x^*) \ge 0$. Then by Lemma 3 the f°-pseudoconvexity of f_{k_1} implies $f_{k_1}(x) \ge f_{k_1}(x^*)$. Thus, there exists no feasible point $x \in S$ with $f_k(x) < f_k(x^*)$ for all $k \in Q$ implying x^* is a global weak Pareto optimum. □

4.1 Inequality Constraints

Now we shall consider problem (5) with inequality constraints:

$$\begin{cases} \text{minimize} & \{f_1(x),\ldots,f_q(x)\} \\ \text{subject to} & g_i(x) \le 0 \quad \text{for all } i = 1,\ldots,m, \end{cases} \tag{12}$$

where also $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1,\ldots,m$ are locally Lipschitz continuous functions. Denote $M = \{1,2,\ldots,m\}$ and the *total constraint function* by

$$g(x) = \max\{g_i(x) \mid i = 1,\ldots,m\}.$$

Problem (12) can be seen as a special case of (5), where

$$S = \{x \in \mathbb{R}^n \mid g(x) \le 0\}.$$

Denote also

$$G(x) = \bigcup_{i \in I(x)} \partial g_i(x), \text{ where } I(x) = \{i \mid g_i(x) = 0\}.$$

For necessary conditions we need some constraint qualifications. We restrict ourselves to constraint qualifications that give conditions in terms of feasible set or constraint functions. This makes the constraint qualifications easily applicable to both single and multiobjective problems. There are many constraint qualifications involving the objective functions too (see, e.g. [14]), but they are not considered here.

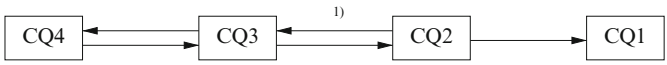
In order to formulate KKT type optimality conditions we need one of the following constraint qualifications

$$\begin{aligned} &\text{(CQ1)} \quad G^{\le}(x) \subset T_S(x) \\ &\text{(CQ2)} \quad 0 \notin \partial g(x) \\ &\text{(CQ3)} \quad G^{<}(x) \neq \emptyset \\ &\text{(CQ4)} \quad 0 \notin \operatorname{conv} G(x), \end{aligned}$$

where we assume $I(x) \neq \emptyset$ for all the constraint qualifications. Due to Theorem 1(b) the assumption $I(x) \neq \emptyset$ guarantees that $G(x) \neq \emptyset$. Note that the sets $G^{\le}(x)$ and $G^{<}(x)$ can be defined also in terms of generalized directional derivatives. For example

$$\begin{aligned} G^{\le}(x) &= \{d \mid \xi^T d \le 0, \text{ for all } \xi \in \bigcup_{i \in I(x)} \partial g_i(x)\} \\ &= \{d \mid g_i^\circ(x;d) \le 0, \text{ for all } i \in I(x)\}. \end{aligned}$$

In [14] CQ1 and CQ3 were called nonsmooth analogs of Abadie qualification and Cottle qualification respectively, while both CQ4 and CQ2 were called Cottle constraint qualifications in [19] and [15] respectively. In [14] it was shown that CQ1 follows from CQ3. In the appendix we will show that the following relations hold between the given constraint qualifications.



[1)] if all constraint functions are subdifferentially regular or f° -pseudoconvex

Fig. 2 Relations between different constraint qualifications

Next, we will prove a KKT Theorem in the case where the constraint qualification is CQ1. As seen in Fig. 2, CQ1 is the weakest condition of the above qualifications. Thus, CQ1 can be replaced by any of CQ2, CQ3 or CQ4. The proof of the KKT Theorem is in practice the same as in [14]. The idea is quite similar to the proof in [2, p. 165] for differentiable single objective case. The outline of the proof goes as follows. First we characterize a necessary condition for (weak Pareto) optimality in terms of contingent cone and objective function(s). Then, by some constraint qualification we replace the contingent cone by another cone, related to constraint functions and, finally, by some alternative theorem we may express the optimality in the form of KKT conditions. The main difference between the differentiable and nondifferentiable case is that the cones are defined with generalized directional derivatives (or subdifferentials) instead of classical gradients.

The weak Pareto optimality was expressed in terms of contingent cone and objective functions in Lemma 4. Let us then prove the theorem of alternatives needed in the proof of the KKT Theorem.

Lemma 6. *Let $S \subset \mathbb{R}^n$ be a nonempty closed convex set and let $C \subset \mathbb{R}^n$ be a nonempty closed convex cone. Then one and only one of the following relations hold*

a) $S \cap C \neq \emptyset$
b) $S^< \cap -C^\leq \neq \emptyset$.

Proof. Assume that $S \cap C \neq \emptyset$. If $S^< = \emptyset$ then trivially $S^< \cap -C^\leq = \emptyset$. If $d \in S^< \neq \emptyset$, we have $s^T d < 0$ for all $s \in S \cap C$. Thus, $d \notin -C^\leq = \{x \mid x^T c \geq 0, \forall c \in C\}$ and $S^< \cap -C^\leq = \emptyset$.

Assume next that $S \cap C = \emptyset$. Since S and C are closed convex sets the Separation Theorem (see e.g. [2]) implies there exist $d \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that

$$d^T s < \alpha \quad \text{for all } s \in S \tag{13}$$
$$d^T c \geq \alpha \quad \text{for all } c \in C. \tag{14}$$

Since C is a cone, $0 \in C$ and C is unbounded, we can choose $\alpha = 0$. Then, Eq. (13) means that $d \in S^<$ and Eq. (14) means that $d \in -C^\leq$. Thus, $d \in S^< \cap -C^\leq \neq \emptyset$. □

The following results are useful in the proof of necessary conditions.

Lemma 7. *Let f_k, $k \in Q$ and g_i, $i \in M$ be locally Lipschitz continuous and $S \subset \mathbb{R}^n$ an arbitrary set. Then*

$$S^\leq = (\mathrm{cl}\, S)^\leq, \quad F^<(x) = (\mathrm{conv}\, F(x))^< \quad \textit{and} \quad G^\leq(x) = (\mathrm{cone}\, G(x))^\leq.$$

Proof. Since

$$S \subset \mathrm{cl}\, S, \quad F(x) \subset \mathrm{conv}\, F(x) \quad \text{and} \quad G(x) \subset \mathrm{cone}\, G(x)$$

clearly

$$(\mathrm{cl}\, S)^\leq \subset S^\leq, \quad (\mathrm{conv}\, F(x))^< \subset F^<(x) \quad \text{and} \quad (\mathrm{cone}\, G(x))^\leq \subset G^\leq(x).$$

Suppose that $d \in S^\leq$. If $d \notin (\mathrm{cl}\, S)^\leq$ then $d^T s > 0$ for some $s \in \mathrm{cl}\, S$. By the continuity of function $d^T s$ there exists $\varepsilon > 0$ such that $d^T b > 0$ for all $b \in B(s;\varepsilon)$. This contradicts with assumption $d \in S^\leq$ as $B(s;\varepsilon) \cap S \neq \emptyset$.

Suppose that $d \in F^<(x)$. Then for every $\xi \in \bigcup_{k \in Q} \partial f_k(x)$ we have $d^T \xi < 0$. Then

$$d^T \left(\sum_{k=1}^{q} \lambda_k \xi_k \right) = \sum_{k=1}^{q} \lambda_k d^T \xi_k < 0,$$

for all $\xi_k \in \partial f_k(x)$ and $\lambda_k \geq 0$, $\sum_{k=1}^{q} \lambda_k = 1$. Hence, $d \in (\mathrm{conv}\, F(x))^<$.

Suppose that $d \in G^\leq(x)$. Likewise to the previous case we can show that $d \in (\mathrm{conv}\, G(x))^\leq$. Then

$$d^T \xi \leq 0 \quad \text{implies} \quad d^T \lambda \xi \leq 0$$

for all $\lambda \geq 0$ and $\xi \in \mathrm{conv}\, G(x)$. Hence, $d \in (\mathrm{cone}\, G(x))^\leq$. □

Now, we are ready to formulate the necessary condition for local weak Pareto optimality.

Theorem 15. *If x^* is a local weak Pareto optimum and CQ1 holds then*

$$0 \in \mathrm{conv}\, F(x^*) + \mathrm{cl}\, \mathrm{cone}\, G(x^*). \tag{15}$$

Proof. By Lemma 4 $F^<(x^*) \cap T_S(x^*) = \emptyset$. Since the CQ1 holds we have

$$F^<(x^*) \cap G^\leq(x^*) \subset F^<(x^*) \cap T_S(x^*) = \emptyset.$$

By Lemma 7 we have

$$\begin{aligned} F^<(x^*) \cap G^\leq(x^*) &= (\mathrm{conv}\, F(x^*))^< \cap (\mathrm{cone}\, G(x^*))^\leq \\ &= (\mathrm{conv}\, F(x^*))^< \cap (\mathrm{cl}\, \mathrm{cone}\, G(x^*))^\leq = \emptyset. \end{aligned}$$

Since $F(x^*)$ and $G(x^*)$ are nonempty ($I(x^*) \neq \emptyset$), $\operatorname{conv} F(x^*)$ is a closed convex set (Lemma 2) and $\operatorname{cl}\operatorname{cone} G(x^*)$ is a closed convex cone. Then Lemma 6 implies

$$\operatorname{conv} F(x^*) \cap -\operatorname{cl}\operatorname{cone} G(x^*) \neq \emptyset.$$

This is equivalent with $0 \in \operatorname{conv} F(x^*) + \operatorname{cl}\operatorname{cone} G(x^*)$. □

Since Pareto optimality implies weak Pareto optimality we get immediately the following consequence.

Corollary 2. *Condition* (15) *is also necessary for* x^* *to be a local Pareto optimum of* (12).

In Theorem 15 it was assumed that $I(x) \neq \emptyset$. If this is not the case, then we have $g(x) < 0$. By continuity of g there exists $\varepsilon > 0$ such that $B(x;\varepsilon)$ belongs to the feasible set. Then $N_S(x) = \{0\}$ and with Theorem 13 we may deduce that condition in Theorem 12 holds. From that we may deduce that assumption $I(x) \neq \emptyset$ could be omitted if in (15) $\operatorname{cl}\operatorname{cone} G(x^*)$ is replaced by $\{0\} \cup \operatorname{cl}\operatorname{cone} G(x^*)$.

A condition stronger than (15) was developed for CQ3 in [14, 19]. Next we shall study the stronger condition. For that we need the following lemma.

Lemma 8. *If CQ4 (or equivalently CQ3) holds at* $x \in \mathbb{R}^n$, *then* $\operatorname{cone} G(x)$ *is closed.*

Proof. Let $(d_j) \subset \operatorname{cone} G(x)$ be an arbitrary converging sequence such that $\lim_{j\to\infty} d_j = \hat{d}$. For every j there exists $\lambda_j \geq 0$ and $\xi_j \in \operatorname{conv} G(x)$ such that $d_j = \lambda_j \xi_j$. By Lemma 2 $\operatorname{conv} G(x)$ is a compact set. Then there exists a converging subsequence (ξ_{j_i}) such that $\lim_{i\to\infty} \xi_{j_i} = \hat{\xi}$. By closedness of $\operatorname{conv} G(x)$ we have $\hat{\xi} \in \operatorname{conv} G(x)$. Since $0 \notin \operatorname{conv} G(x)$ sequence

$$\lambda_{j_i} = \frac{\|d_{j_i}\|}{\|\xi_{j_i}\|}$$

is converging too. Denote $\lim_{i\to\infty} \lambda_{j_i} = \hat{\lambda}$. Then

$$\hat{d} = \hat{\lambda}\hat{\xi} \in \operatorname{cone} G(x)$$

implying that $\operatorname{cone} G(x)$ is closed. □

Theorem 16. *If* x^* *is a local weak Pareto optimum and CQ3 holds, then*

$$0 \in \operatorname{conv} F(x^*) + \operatorname{cone} G(x^*).$$

Proof. From Lemma 8 it follows that if CQ3 holds then $\operatorname{cl}\operatorname{cone} G(x^*) = \operatorname{cone} G(x^*)$. Then the result follows directly from Theorem 15. □

Consider then the sufficient conditions of problem (12). It is well-known that the convexity of the functions f_k, $k \in Q$, and g_i, $i \in M$, guarantees the sufficiency of the

KKT optimality condition for global weak Pareto optimality in Theorem 16 (see [19, p. 51]). We will present the sufficient conditions in more detail later. Namely, they can be obtained as a special case of sufficient conditions for problems with both inequality and equality constraints.

4.2 Equality Constraints

Consider problem (5) with both inequality and equality constraints.

$$\begin{cases} \text{minimize} & \{f_1(x),\ldots,f_q(x)\} \\ \text{subject to} & g_i(x) \le 0 \quad \text{for all } i=1,\ldots,m, \\ & h_j(x) = 0 \quad \text{for all } j=1,\ldots,p, \end{cases} \tag{16}$$

where all functions are supposed to be locally Lipschitz continuous. Denote $H(x) = \bigcup_{j=1}^{p} \partial h_j(x)$ and $J = \{1,2,\ldots,p\}$. By Theorem 1(b) we see that

$$-H(x) = -\bigcup_{j\in J} \partial h_j(x) = \bigcup_{j\in J} \partial(-h_j)(x).$$

A straightforward way to deal with an equality constraint $h_j(x) = 0$ is to replace it with two inequality constraints

$$h_j(x) \le 0 \quad \text{and} \quad -h_j(x) \le 0. \tag{17}$$

Then, we may use the results obtained for problem (12) to derive results for problem (16). However, some constraint qualifications are not satisfied if this kind of operation is done as we will see soon.

Consider first the CQ1. Denote

$$\begin{aligned} G_*^{\le}(x) &= \{d \mid g_i^\circ(x;d) \le 0,\, i \in I(x),\, h_j^\circ(x;d) \le 0,\, (-h_j)^\circ(x;d) \le 0,\, j \in J\} \\ &= G^{\le}(x) \cap H^{\le}(x) \cap (-H)^{\le}(x). \end{aligned}$$

It is good to note that we can replace $(-h_j)^\circ(x;d) \le 0$ by $h_j^\circ(x;-d) \le 0$ in the definition of $G_*^{\le}(x)$ according to Theorem 1(a). We can use a new cone instead of the cone $H^{\le}(x) \cap (-H)^{\le}(x)$ as the next lemma shows.

Lemma 9. *Let $h : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. Then*

$$\begin{aligned} \partial h(x)^{\le} \cap (-\partial h(x))^{\le} &= \{d \mid h^\circ(x;d) \le 0,\, h^\circ(x;-d) \le 0\} \\ &\subset \{d \mid h^\circ(x;d) = 0\} \end{aligned}$$

Proof. Suppose $d \in \partial h(x)^{\le} \cap (-\partial h(x))^{\le}$. By the subadditivity of h° [Theorem 1(a)] we have

$$0 = h^\circ(x;0) \le h^\circ(x;-d) + h^\circ(x;d) \le 0, \tag{18}$$

which is possible only if $h^\circ(x;-d) = h^\circ(x;d) = 0$. Namely, if one would be strictly negative the other should be strictly positive in order to satisfy inequality (18). This is impossible since $d \in \partial h(x)^{\leq} \cap (-\partial h(x))^{\leq}$. □

Denote

$$H^0(x) = \{d \mid h_j^\circ(x;d) = 0 \text{ for all } j \in J\}.$$

From Lemma 9 we can easily deduce that $H^{\leq}(x) \cap (-H)^{\leq}(x) \subset H^0(x)$. However, in general $H^0(x) \not\subset H^{\leq}(x) \cap (-H)^{\leq}(x)$. To see this, consider a function

$$h(x) = \begin{cases} -x, & \text{if } x \leq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Then $h^\circ(0,1) = 0$ and $h^\circ(0,-1) = 1$. Thus, $1 \in H^0(0)$ but $1 \notin H^{\leq}(0) \cap (-H)^{\leq}(0)$.

Now we can present two constraint qualifications for problem (16):

$$\begin{aligned} &\text{(CQ5)} \quad G^{\leq}(x) \cap H^{\leq}(x) \cap (-H)^{\leq}(x) \subset T_S(x) \\ &\text{(CQ6)} \quad G^{\leq}(x) \cap H^0(x) \subset T_S(x), \end{aligned}$$

where again $I(x) \neq \emptyset$. From Lemma 9 we see that CQ6 implies CQ5. Thus, we can derive KKT conditions with CQ6 if we can do so for CQ5.

Consider next the constraint qualification CQ2. Assume our problem has an equality constraint $h_1(x) = 0$. Then, at the feasible points the total constraint function will be

$$g(x) = \max\{h_1(x), -h_1(x), l(x)\} = \max\{\max\{h_1(x), -h_1(x)\}, l(x)\},$$

where $l(x)$ contains the other terms. It is clear that function $\max\{h_1(x), -h_1(x)\}$ is non-negative. Consequently, g is non-negative too. Then, 0 is minimum value for g and it is attained at every feasible point of problem (16). Thus, for any feasible x we have $0 \in \partial g(x)$ according to Theorem 5 and, thus, CQ2 does not hold. Hence, CQ2 is not suitable for problems with equality constraints.

Next, we shall consider CQ3. Denote

$$\begin{aligned} G_*^{<}(x) &= \{d \mid g_i^\circ(x;d) < 0, i \in I(x), h_j^\circ(x;d) < 0, (-h_j)^\circ(x;d) < 0, j \in J\} \\ &= G^{<}(x) \cap \{d \mid h_j^\circ(x;d) < 0, h_j^\circ(x;-d) < 0, j \in J\}. \end{aligned}$$

Let $x, d \in \mathbb{R}^n$ and $j \in J$ be arbitrary. By the subadditivity of h_j° we have

$$0 = h_j^\circ(x;0) \leq h_j^\circ(x;d) + h_j^\circ(x;-d). \tag{19}$$

From inequality (19) it is easy to see that $\{d \mid h_j^\circ(x;d) < 0, h_j^\circ(x;-d) < 0\} = \emptyset$. Hence, CQ3 does not hold implying that the constraint qualification CQ3 (or CQ4) is not suitable for equality constraints.

Before the proof of the KKT Theorem of problem (16) we need the following lemma.

Lemma 10. *If A and B are nonempty cones then* $\mathrm{cl}(A+B) \subset \mathrm{cl}A + \mathrm{cl}B$.

Proof. Since $A \subset \mathrm{cl}A$ and $B \subset \mathrm{cl}B$ we have $A+B \subset \mathrm{cl}A + \mathrm{cl}B$. By Lemma 2 in [20] $\mathrm{cl}A + \mathrm{cl}B$ is closed. Thus, $\mathrm{cl}(A+B) \subset \mathrm{cl}A + \mathrm{cl}B$. □

Finally, we can state the theorem corresponding to Theorem 15 with constraint qualification CQ5.

Theorem 17. *If x^* is a local weak Pareto optimum of (16) and CQ5 holds at x^*, then*

$$0 \in \mathrm{conv}\, F(x^*) + \mathrm{cl}\,\mathrm{cone}\, G(x^*) + \mathrm{cl}\,\mathrm{cone}\, H(x^*) - \mathrm{cl}\,\mathrm{cone}\, H(x^*). \tag{20}$$

Proof. From Theorem 15 and previous considerations we see that

$$0 \in \mathrm{conv}\, F(x^*) + \mathrm{cl}\,\mathrm{cone}(G(x^*) \cup H(x^*) \cup -H(x^*)). \tag{21}$$

By using Lemma 1(b) twice and Lemma 10 we obtain

$$\begin{aligned}
&\mathrm{cl}\,\mathrm{cone}(G(x^*) \cup H(x^*) \cup -H(x^*)) \\
&= \mathrm{cl}\left(\sum_{i \in I(x^*)} \mathrm{ray}\, \partial g_i(x^*) + \sum_{j \in J} \mathrm{ray}\, \partial h_j(x^*) + \sum_{j \in J} \mathrm{ray}\, \partial(-h_j(x^*)) \right) \\
&= \mathrm{cl}(\mathrm{cone}\, G(x^*) + \mathrm{cone}\, H(x^*) - \mathrm{cone}\, H(x^*)) \\
&\subset \mathrm{cl}\,\mathrm{cone}\, G(x^*) + \mathrm{cl}\,\mathrm{cone}\, H(x^*) - \mathrm{cl}\,\mathrm{cone}\, H(x^*).
\end{aligned}$$

Combining this with relation (21) proves the theorem. □

There are papers dealing with equality constraints in nonsmooth problems without turning them into inequality constraints (see, e.g., [11]). However, the conditions are expressed in terms of generalized Jacobian of multivalued mapping $h : \mathbb{R}^m \to \mathbb{R}^n$. We shall not consider generalized Jacobians here and, thus, will not discuss these type of conditions further.

There are also papers where closures are not needed in conditions in Theorem 17 (see, e.g., [29]). But there they used constraint qualifications including objective functions which we shall not consider either.

After the necessary conditions we shall now study sufficient conditions. For that we do not need the constraint qualifications but we have to make some assumptions on objective and constraint functions. More accurately, we assume that objective functions are f°-pseudoconvex and inequality constraint functions are f°-quasiconvex. The equality constraints may be f°-quasiconvex or f°-quasiconcave. Denote

$$H_+(x) = \bigcup_{j \in J_+} \partial h_j(x) \quad \text{and} \quad H_-(x) = \bigcup_{j \in J_-} \partial h_j(x),$$

where $J_- \cup J_+ = J$ and h_j is f°-quasiconvex if $j \in J_+$ and h_j is f°-quasiconcave if $j \in J_-$.

Theorem 18. *Let x^* be a feasible point of problem (16). Suppose f_k are f°-pseudoconvex for all $k \in Q$, g_i are f°-quasiconvex for all $i \in M$, h_j are f°-quasiconvex for all $j \in J_+$ and f°-quasiconcave for all $j \in J_-$. If*

$$0 \in \operatorname{conv} F(x^*) + \operatorname{cone} G(x^*) + \operatorname{cone} H_+(x^*) - \operatorname{cone} H_-(x^*), \tag{22}$$

then x^* is a global weak Pareto optimum of (16).

Proof. Note that if (22) is satisfied then $I(x^*) \neq \emptyset$. Let $x \in \mathbb{R}^n$ be an arbitrary feasible point. Then $g_i(x) \leq g_i(x^*)$ if $i \in I(x^*)$, $h_j(x) = h_j(x^*)$ for all $j \in J_+ \cup J_-$ and f°-quasiconvexity implies that

$$g_i^\circ(x^*; x - x^*) \leq 0 \text{ for all } i \in I(x^*) \tag{23}$$
$$h_j^\circ(x^*; x - x^*) \leq 0 \text{ for all } j \in J_+. \tag{24}$$

The f°-quasiconcavity implies that

$$h_j^\circ(x^*; x^* - x) \leq 0 \text{ for all } j \in J_-. \tag{25}$$

According to (22) there exist $\xi_k \in \partial f_k(x^*)$, $\zeta_i \in \partial g_i(x^*)$, $\eta_j \in \partial h_j(x^*)$ and coefficients $\lambda_k, \mu_i, \nu_j \geq 0$, for all $k \in Q$, $i \in I(x^*)$ and $j \in J$ such that $\sum_{k=1}^q \lambda_k = 1$ and

$$0 = \sum_{k \in Q} \lambda_k \xi_k + \sum_{i \in I(x^*)} \mu_i \zeta_i + \sum_{j \in J_+} \nu_j \eta_j - \sum_{j \in J_-} \nu_j \eta_j. \tag{26}$$

Multiplying equation (26) by $x - x^*$, using Definition 1 and Eqs. (23), (24) and (25) we obtain

$$\begin{aligned}
&- \sum_{k \in Q} \lambda_k \xi_k^T (x - x^*) \\
&= \sum_{i \in I(x^*)} \mu_i \zeta_i^T (x - x^*) + \sum_{j \in J_+} \nu_j \eta_j^T (x - x^*) + \sum_{j \in J_-} \nu_j \eta_j^T (x^* - x) \\
&\leq \sum_{i \in I(x^*)} \mu_i g_i^\circ(x^*; x - x^*) + \sum_{j \in J_+} \nu_j h_j^\circ(x^*; x - x^*) + \sum_{j \in J_-} \nu_j h_j^\circ(x^*; x^* - x) \\
&\leq \sum_{i \in I(x^*)} \mu_i \cdot 0 + \sum_{j \in J_+} \nu_j^+ \cdot 0 + \sum_{j \in J_-} \nu_j \cdot 0 = 0.
\end{aligned}$$

Thus,

$$0 \leq \sum_{k \in Q} \lambda_k \xi_k^T (x - x^*) \leq \sum_{k \in Q} \lambda_k f_k^\circ(x^*; x - x^*).$$

Since $\lambda_k \geq 0$ for all $k \in Q$ and $\sum_{k \in Q} \lambda_k = 1 > 0$ there exists $k_1 \in Q$ such that

$$0 \leq f_{k_1}^\circ(x^*; x - x^*).$$

Then, f°-pseudoconvexity of f_{k_1} implies that $f_{k_1}(x^*) \leq f_{k_1}(x)$. Since x is an arbitrary feasible point there exists no feasible point $y \in \mathbb{R}^n$ such that $f_k(y) < f_k(x^*)$ for all $k \in Q$. Thus, x^* is a global weak Pareto optimum of problem (16). □

Note, that due to Theorems 9 and 11 the previous result is valid also for f°-pseudoconvex and subdifferentially regular quasiconvex inequality constraint functions. Also, the implicit assumption $I(x^*) \neq \emptyset$ could be omitted by replacing $\operatorname{cone} G(x^*)$ by $\{0\} \cup \operatorname{cone} G(x^*)$.

Finally, by modifying somewhat the proof we get the sufficient KKT optimality condition for global Pareto optimality with an extra assumption for the multipliers.

Corollary 3. *The condition of Theorem 18 is also sufficient for x^* to be a global Pareto optimum of (16), if in addition $\lambda_j > 0$ for all $k \in Q$.*

Proof. By the proof of Theorem 18 we know that inequality

$$0 \leq \sum_{k \in Q} \lambda_k \xi_k^T (x - x^*) \leq \sum_{k \in Q} \lambda_k f_k^\circ(x^*; x - x^*) \tag{27}$$

holds for arbitrary feasible x. Suppose there exists $k_1 \in Q$ such that $f_{k_1}^\circ(x^*; x - x^*) < 0$. Because $\lambda_k > 0$ for all $k \in Q$, by inequality (27) there must be also $k_2 \in Q$ such that $f_{k_2}^\circ(x^*; x - x^*) > 0$. By Theorem 9 f_{k_2} is f°-quasiconvex and by Definition 6 we have $f_{k_2}(x) > f_{k_2}(x^*)$. Since x were arbitrary, x^* is Pareto optimal.

Suppose then that $f_k^\circ(x^*; x - x^*) \geq 0$ for all $k \in Q$. Then the f°-pseudoconvexity implies that $f_k(x^*) \leq f_k(x)$ and, thus, x^* is Pareto optimal. □

As the next example shows a global minimum x^* does not necessarily satisfy the conditions in Theorem 18.

Example 1. Consider the problem

$$\begin{aligned} \text{minimize } & f(x) = -x_1 \\ \text{subject to } & g(x) = (x_1 - 2)^2 + (x_2 + 2)^2 - 2 \leq 0 \\ & h(x) = (x_1 - 4)^2 + x_2^2 - 10 = 0. \end{aligned}$$

All the functions are convex and, thus, the assumptions of Theorem 18 are satisfied.

The global minimum to this problem is $x^* = (3, -3)^T$. The gradients at this point are

$$\nabla f(x^*) = (-1, 0)^T, \quad \nabla g(x^*) = (2, -2)^T \text{ and } \quad \nabla h(x^*) = (-2, -6)^T.$$

The gradients are illustrated in Fig. 3. The lengths of the gradients in figure are scaled for clarity. The bolded curve represents the feasible set. It is easy to see that $0 \notin \nabla f(x^*) + \operatorname{cone} \nabla g(x^*) + \operatorname{cone} \nabla h(x^*)$. Thus we have a global optimum but the sufficient condition is not satisfied.

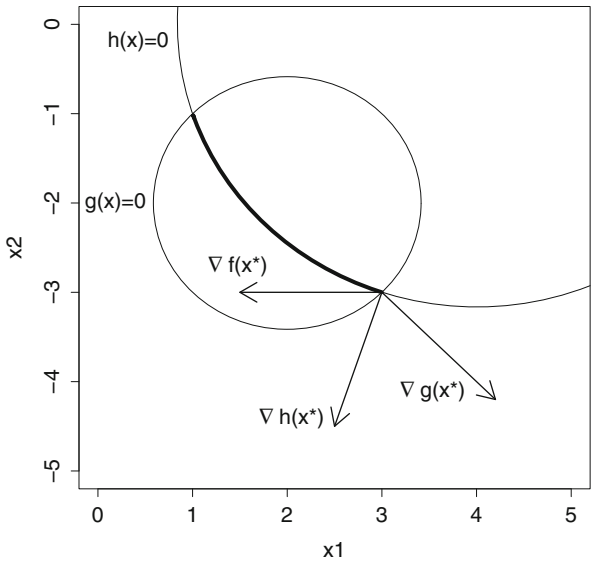


Fig. 3 Gradients at the global minimum

Let us then apply necessary conditions (Theorem 17) to the given example. It is easy to see that qualifications CQ5 and CQ6 are equivalent if functions h_j are differentiable for all $j \in J$. Clearly,

$$\begin{aligned} T_S(x^*) &= \{\lambda(-3,1) \mid \lambda \geq 0\}, \\ H^0(x^*) &= \{\lambda(-3,1) \mid \lambda \in \mathbb{R}\} \text{ and} \\ G^{\leq}(x^*) &= \{(d_1,d_2) \mid d_1,d_2 \in \mathbb{R}, d_1 \leq d_2\}. \end{aligned}$$

Thus, $G^{\leq}(x^*) \cap H^0(x^*) = T_S(x^*)$ implying that CQ6 is satisfied. According to Theorem 17, relation (20) should hold at global minimum x^*. Indeed,

$$\begin{aligned} 0 &= \nabla f(x^*) + \frac{3}{8}\nabla g(x^*) + 0\nabla h(x^*) - \frac{1}{8}\nabla h(x^*) \\ &\subset \operatorname{conv} F(x^*) + \operatorname{cl}\operatorname{cone} G(x^*) + \operatorname{cl}\operatorname{cone} H(x^*) - \operatorname{cl}\operatorname{cone} H(x^*). \end{aligned}$$

5 Concluding Remarks

We have considered KKT type necessary and sufficient conditions for nonsmooth multiobjective optimization problems. Both inequality and equality constraints were considered. The optimality were characterized as a weak Pareto optimality. In necessary conditions CQ1–CQ6 constraint qualifications were needed. In sufficient conditions the main tools used were the generalized pseudo- and quasiconvexities based on the Clarke generalized directional derivative. It was assumed that the objective functions are f°-pseudoconvex and the constraint functions are f°-quasiconvex.

Due to relations between different generalized convexities the results are valid also for f°-pseudoconvex and subdifferentially regular quasiconvex constraint functions.

Appendix

Consider problem (12), that is, problem

$$\begin{cases} \text{minimize} & \{f_1(x),\ldots,f_q(x)\} \\ \text{subject to} & g_i(x)\le 0 \quad \text{for all } i\in M=\{1,\ldots,m\}. \end{cases} \tag{28}$$

Next, we will study some relations between the constraint qualifications. From now on, we assume that $I(x)\neq\emptyset$.

In [14] it was shown that CQ1 follows from CQ3. Next we will prove that CQ1 follows also from CQ2.

Theorem 19. *Let $x\in\mathbb{R}^n$ be a feasible point of problem (28) such that $I(x)\neq\emptyset$. If $0\notin\partial g(x)$ then $G^{\leq}(x)\subset T_S(x)$.*

Proof. Assume that there exists $d^*\in G^{\leq}(x)$ such that $d^*\notin T_S(x)$. Since a contingent cone is a closed set there exists $\varepsilon>0$ such that $\operatorname{cl}B(d^*;\varepsilon)\cap T_S(x)=\emptyset$. Since $d\notin T_S(x)$, for every $d\in\operatorname{cl}B(d^*;\varepsilon)$ there exists $t(d)>0$ such that $g(x+t_1d)>g(x)$ when $0<t_1<t(d)$. Thus,

$$g^\circ(x;d)\ge 0,\ \text{ for all } d\in\operatorname{cl}B(d^*;\varepsilon). \tag{29}$$

Since $d^*\in G^{\leq}(x)$ we have

$$\begin{aligned} g^\circ(x;d^*) &= \max\left\{\zeta^Td^*\mid\zeta\in\partial g(x)\right\} \\ &\le \max\left\{\zeta^Td^*\mid\zeta\in\operatorname{conv}\{\partial g_i(x)\mid i\in I(x)\}\right\} \\ &= \max\{g_i^\circ(x,d^*)\mid i\in I(x)\}\le 0. \end{aligned} \tag{30}$$

Then for all $\zeta\in\partial g(x)$ we have $\zeta^Td^*\le 0$. Since we have $0\notin\partial g(x)$ the Separation Theorem (see, e.g. [2]) implies that there exist $\alpha\in\mathbb{R}$ and z, $\|z\|=1$ such that

$$z^T0>\alpha \quad\text{and}\quad z^T\zeta\le\alpha$$

for all $\zeta\in\partial g(x)$. Since $z^T0=0$ we see that $z^T\zeta<0$ for all $\zeta\in\partial g(x)$. If $\bar d=d^*+\varepsilon z$, then $\bar d\in\operatorname{cl}B(d^*;\varepsilon)$ and

$$\zeta^T\bar d=\zeta^Td^*+\varepsilon\zeta^Tz<0$$

for all $\zeta\in\partial g(x)$. Then

$$g^\circ(x;\bar d)=\max\left\{\zeta^T\bar d\mid\zeta\in\partial g(x)\right\}<0$$

contradicting inequality (29). Thus, $G^{\leq}(x)\subset T_S(x)$. □

There exist problems that satisfy the CQ1 constraint qualification, but does not satisfy the CQ2.

Example 2. Consider the problem (28) with $g(x) = |x|$. Then we have $G^{\leq}(0) = \{0\}$ and $T_S(0) = \{0\}$. Thus, $G^{\leq}(0) \subset T_S(0)$ and CQ1 holds at $x = 0$. However, $0 \in \partial g(0)$ and CQ2 does not hold.

Next we will consider the relations between CQ2 and CQ3. First we will show that CQ2 follows from CQ3.

Theorem 20. *If $I(x) \neq \emptyset$ and $G^{<}(x) \neq \emptyset$, then $0 \notin \partial g(x)$.*

Proof. It follows from the condition $G^{<}(x) \neq \emptyset$ that there exists d, such that $g_i^{\circ}(x;d) < 0$ for all $i \in I(x)$. In other words, $d^T \xi_i < 0$ for all $\xi_i \in \partial g_i(x)$ and $i \in I(x)$. Let $\lambda_i \geq 0$, $i \in I(x)$ be scalars such that $\sum_{i \in I(x)} \lambda_i = 1$. Then

$$d^T \sum_{i \in I(x)} \lambda_i \xi_i = \sum_{i \in I(x)} \lambda_i d^T \xi_i < 0.$$

Thus, $d^T \xi < 0$ for all $\xi \in \operatorname{conv} \bigcup_{i \in I(x)} \partial g_i(x)$. Since $\partial g(x) \subset \operatorname{conv} \bigcup_{i \in I(x)} \partial g_i(x)$, we have $g^{\circ}(x;d) < 0$ implying that $0 \notin \partial g(x)$. □

There exist problems for which CQ2 holds but CQ3 does not as the following example shows.

Example 3. Consider constraint functions

$$g_1(x) = x \quad \text{and} \quad g_2(x) = \begin{cases} x, & \text{if } x < 0 \\ 0, & \text{if } x \geq 0. \end{cases}$$

Then $g(x) = \max\{g_1(x), g_2(x)\} = g_1(x)$ and $0 \notin \partial g(0)$. However, $0 \in \partial g_2(0)$ which implies $G^{<}(0) = \emptyset$.

Despite Example 3 we can establish some conditions on constraint functions which guarantees that CQ2 implies CQ3. Namely, if all the constraint functions are subdifferentially regular or f°-pseudoconvex the CQ3 follows from CQ2.

Theorem 21. *Let $x \in \mathbb{R}^n$ and $I(x) \neq \emptyset$. If the functions g_i are subdifferentially regular for all $i \in M$ and $0 \notin \partial g(x)$, then $G^{<}(x) \neq \emptyset$.*

Proof. If $0 \notin \partial g(x)$, then there exists d, such that $g^{\circ}(x;d) < 0$. Due to regularity we have $\partial g(x) = \operatorname{conv} \bigcup_{i \in I(x)} \partial g_i(x)$. Hence,

$$d^T \sum_{i \in I(x)} \lambda_i \xi_i < 0, \text{ for all } \xi_i \in \partial g_i(x), \lambda_i \geq 0, \sum_{i \in I(x)} \lambda_i = 1,$$

implying $d^T \xi_i < 0$ for all $\xi_i \in \partial g_i(x)$. In other words $g_i^{\circ}(x;d) < 0$ for all $i \in I(x)$. Thus, we have $d \in G^{<} \neq \emptyset$. □

Theorem 22. *Let $x \in \mathbb{R}^n$ and $I(x) \neq \emptyset$. If the functions g_i are f°-pseudoconvex for all $i \in M$ and $0 \notin \partial g(x)$, then $G^<(x) \neq \emptyset$.*

Proof. On contrary, assume that $G^< = \emptyset$. Then for all $d \in \mathbb{R}^n$ there exists $i \in I(x)$, for which $g_i^\circ(x;d) \geq 0$. Due to f°-pseudoconvexity we have $g_i(x+td) \geq g_i(x)$ for all $t \geq 0$. Since $g(x) \geq g_i(x)$ for all $i \in M$ we have $g(x+td) \geq g(x)$ for all $d \in \mathbb{R}^n$. Thus, x is a global minimum and $0 \in g(x)$ by Theorem 5. In other words, if $0 \notin g(x)$ we will have $G^< \neq \emptyset$. □

Finally, we will show that constraint qualification CQ3 is equivalent to CQ4.

Theorem 23. *Suppose $I(x) \neq \emptyset$. Then $0 \notin \operatorname{conv} G(x)$ iff $G^<(x) \neq \emptyset$.*

Proof. The condition $0 \notin \operatorname{conv} G(x)$ is equivalent to condition $\operatorname{conv} G(x) \cap \{0\} = \emptyset$. By Lemma 2 $\operatorname{conv} G(x)$ is a closed convex set and trivially $\{0\}$ is a closed convex cone. Also, $\{0\}^\leq = \mathbb{R}^n = -\{0\}^\leq$. By Lemma 6 $\operatorname{conv} G(x) \cap \{0\} = \emptyset$ is equivalent to

$$(\operatorname{conv} G(x))^< \cap \mathbb{R}^n = (\operatorname{conv} G(x))^< \neq \emptyset.$$

Furthermore, $(\operatorname{conv} G(x))^< = G^<(x)$ according to Lemma 7. □

References

1. Avriel, M., Diewert, W.E., Schaible, S., Zang, I.: Generalized Concavity. Plenum Press, New York (1988)
2. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming Theory and Algorithms. Wiley, New York (1979)
3. Bhatia, D., Aggarwal, S.: Optimality and duality for multiobjective nonsmooth programming. Eur. J. Oper. Res. **57**, 360–367 (1992)
4. Bhatia, D., Jain, P.: Generalized (F, ρ)-convexity and duality for non smooth multi-objective programs. Optimization **31**, 153–164 (1994)
5. Bhatia, D., Mehra, A.: Optimality conditions and duality involving arcwise connected and generalized arcwise connected functions. J. Optim. Theory Appl. **100**, 181–194, (1999)
6. Brandão, A.J.V., Rojas-Medar, M.A., Silva, G.N.: Invex nonsmooth alternative theorem and applications. Optimization **48**, 239–253 (2000)
7. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley-Interscience, New York (1983)
8. Diewert, W.E.: Alternative characterizations of six kinds of quasiconcavity in the nondifferentiable case with applications to nonsmooth programming. In: Schaible, S., Ziemba, W.T. (eds.) Generalized Concavity in Optimization and Economics, pp. 51–95. Academic, New York (1981)
9. Hanson, M.A.: On sufficiency of the Kuhn-Tucker conditions. J. Math. Anal. Appl. **80**, 545–550 (1981)
10. Hiriart-Urruty, J.B.: New concepts in nondifferentiable programming. Bull. Soc. Math. Fr. **60**, 57–85 (1979)
11. Jourani, A.: Constraint qualifications and Lagrange multipliers in nondifferentiable programming problems. J. Optim. Theory Appl. **81**, 533–548 (1994)
12. Kim, D.S., Lee, H.J.: Optimality conditions and duality in nonsmooth multiobjective programs. J. Inequal. Appl. (2010). doi: 10.1155/2010/939537

13. Komlósi, S.: Generalized monotonicity and generalized convexity. J. Optim. Theory Appl. **84**, 361–376 (1995)
14. Li, X.F.: Constraint qualifications in nonsmooth multiobjective optimization. J. Optim. Theory Appl. **106**, 373–398 (2000)
15. Mäkelä, M.M., Neittaanmäki, P.: Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control. World Scientific, Singapore (1992)
16. Mäkelä, M.M., Karmitsa, N., Eronen, V-P.: On Generalized Pseudo- and Quasiconvexities for Nonsmooth Functions. TUCS Technical Report 989, Turku Centre for Computer Science, Turku (2010)
17. Mäkelä, M.M., Karmitsa, N., Eronen, V-P.: On Nonsmooth Optimality Conditions with Generalized Convexities. TUCS Technical Report 1056, Turku Centre for Computer Science, Turku (2012)
18. Mangasarian, O.L.: Pseudoconvex functions. SIAM J. Control **3**, 281–290 (1965)
19. Miettinen, K.: Nonlinear Multiobjective Optimization. Kluwer Academic, Boston (1999)
20. Miettinen, K., Mäkelä, M.M.: On cone characterizations of weak, proper and Pareto optimality in multiobjective optimization. Math. Methods Oper. Res. **53**, 233–245 (2001)
21. Migdalas, A., Pardalos, P.M., Värbrand, P. (eds.): From Local to Global Optimization. Nonconvex Optimization and Its Applications, vol. 53. Kluwer Academic, Dordrecht (2001)
22. Mishra, S.K.: On sufficiency and duality for generalized quasiconvex nonsmooth programs. Optimization **38**, 223–235 (1996)
23. Nobakhtian, S.: Infine functions and nonsmooth multiobjective optimization problems. Comput. Math. Appl. **51**, 1385–1394 (2006)
24. Nobakhtian, S.: Multiobjective problems with nonsmooth equality constraints. Numer. Funct. Anal. Optim. **30**, 337–351 (2009)
25. Osuna-Gómez, R., Beato-Moreno, A., Rufian-Lizana, A.: Generalized convexity in multiobjective programming. J. Math. Anal. Appl. **233**, 205–220 (1999)
26. Pini, R., Singh, C.: A survey of recent [1985–1995] advances in generalized convexity with applications to duality theory and optimality conditions. Optimization **39**, 311–360 (1997)
27. Preda, V.: On efficiency and duality for multiobjective programs. J. Math. Anal. Appl. **166**, 365–377 (1992)
28. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
29. Sach, P.H., Lee, G.M., Kim, D.S.: Efficiency and generalized convexity in vector optimisation problems. ANZIAM J. **45**, 523–546 (2004)
30. Schaible, S.: Generalized monotone maps. In: Giannessi, F. (ed.) Nonsmooth Optimization: Methods and Applications, pp. 392–408. Gordon and Breach Science Publishers, Amsterdam (1992)
31. Staib, T.: Necessary optimality conditions for nonsmooth multicriteria optimization problem. SIAM J. Optim. **2**, 153–171 (1992)
32. Yang, X.M., Liu, S.Y.: Three kinds of generalized convexity. J. Optim. Theory Appl. **86**, 501–513 (1995)

A Game Theoretical Model for Experiment Design Optimization

Lina Mallozzi, Egidio D'Amato, and Elia Daniele

1 Introduction

This work concerns the optimization of receiver location on ground, under uniform cosmic source distribution, on a bounded settlement area and constrained by a limited number of receivers due to a budget limitation. Assuming the capture range of each receiver (e.g. a radar) to be shaped as a circular area, this problem could be considered to have many points in common with classic sphere packing problem [4, 10, 17] that has been applied in several fields and faced with algorithmic optimization procedure [2, 9, 11, 15, 18, 19].

Here we present the experimental design problem as a Nash equilibrium problem as stated in Game Theory: the choice of the variables in n experiments is made by n players, each of them has to decide his location far as possible from the opponents and also from the border of the region. It turns out that the game has a peculiar structure, namely it is a potential game [12, 14] and the Nash equilibrium solutions will be the minimum points of a function that is called the potential function or potential in short. The numerical procedure to compute the maximum points of the potential is based on a genetic algorithm [3, 5, 6, 8, 13, 16, 21].

In Fig. 1 some sketches for the classic sphere packing problem are shown. In this case is exploited the greater difference between the typical location problem on a bounded domain and the specific location problem that we deal with in this work. In our case the nature of the domain's boundaries is such to act as a cut-off line on which the receiver lost its efficacy or any other measure of profit. In other words,

L. Mallozzi (✉)
Department of Mathematics and Applications "R. Caccioppoli", Università degli Studi di Napoli "Federico II", Via Claudio 21, 80125 Napoli, Italy
e-mail: mallozzi@unina.it

E. D'Amato • E. Daniele
Department of Industrial Engineering - Aerospace Section, Università degli Studi di Napoli "Federico II", Via Claudio 21, 80125 Napoli, Italy
e-mail: egidio.damato@unina.it; elia.daniele@unina.it

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_18,

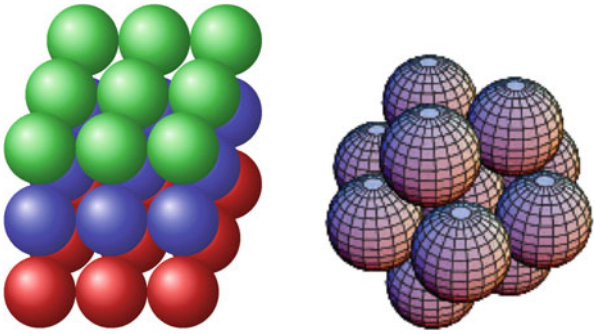


Fig. 1 Sketches for classic sphere packing problem

as in the sphere packing problem the spheres are forced to stay within a limited bounded volume avoiding to consider their elasticity to reduce their size and in our location problem the receiver or sensor (for an experiment) would lose a portion of its efficacy in collecting the signal (pressure, temperature, etc.) by allowing itself to be pushed on the boundary because the information is limited within the same boundary. This is the analogy that let us expose the location problem as explained in Sect. 2.

In literature, there are many different alternatives concerning location problem among which the ones for which:

- more than one facility has to be located (multifacility location);
- location on continuous regions or networks;
- case of absence of demand points (facility layout models);
- when facilities compete for costumers and their objective is to maximize the market share they capture (competitive models).

The problem could be stated in different ways, and one of the better known historical formalization are the Weber's problem or minisum, that minimizes the sum of weighted distances, and the *minimax* problem (von Neumann) that minimizes the maximal distance between facilities and demand points (Fig. 2).

In this paper a game theoretical model concerning experimental design optimization is illustrated, focusing, in Sect. 2 on the mathematical model, the facility location game, with and without constraints, explaining the procedure to recognize a potential problem inside a Nash equilibrium problem; in Sect. 3 numerical results are shown, using a genetic algorithm procedure to minimize the potential function just developed.

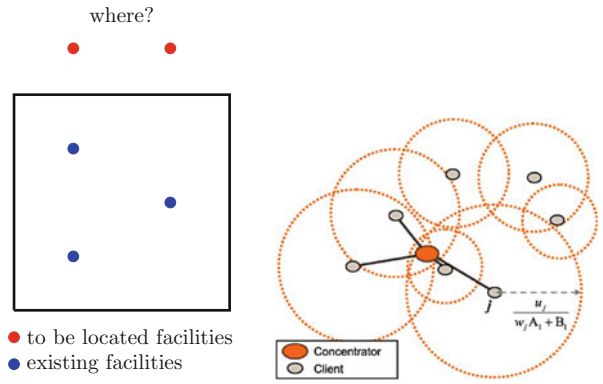


Fig. 2 Location problem statement as a sketch on the *left side*; Weber's problem or *minisum* on the *right side*

2 The Model

Let Ω be a rectangular region of $\mathcal{R}^2$. We restrict the model to the unit square $\Omega = [0,1]^2$ without leading the generalities (rescaling the variables our results hold). The problem is to decide for two variables x and y the values of n available experiments ($n \in \mathcal{N}$). So we want to settle n points $P_1, P_2, \ldots, P_n$ in the square in such a way that they are far as possible from the rest and from the boundary of the square. This implies to maximize the dispersion of the points. We assign each point to a virtual player, whose decision variables are the coordinates and whose payoff function translates the dispersion in terms of distances.

Problem 1. Experimental Design (ED)

The problem of deciding the values of two variables for n assigned experiments is to choose $P_1, \ldots, P_n \in \Omega$ maximizing the

$$\text{dispersion}(P_1, \ldots, P_n),$$

where the dispersion function is defined in a suitable way [7].

There is a competition between the points in the square, because the dispersion depends on the mutual position of all the points, also with respect to the boundary of Ω, so we use a game theoretical model.

The overview of the experimental design problem as a Nash equilibrium problem as stated in Game Theory could be given by Fig. 3 in which each player has to decide his location as far as possible from the other players and also from the border of the region.

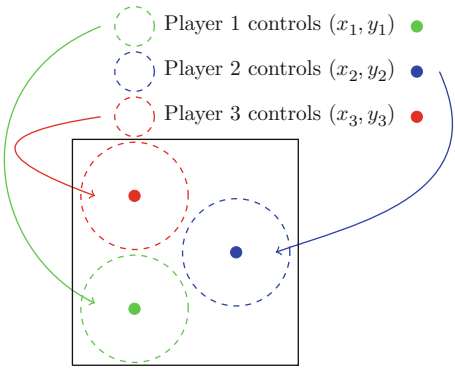


Fig. 3 Location problem as a game

2.1 Preliminaries

Let us consider an n-player normal form game Γ ($n \in \mathcal{N}$, where $\mathcal{N}$ is the set of natural numbers), that consists of a tuple

$$\Gamma = \langle N; X_1, \ldots, X_n; f_1, \ldots, f_n \rangle$$

where $N = \{1, 2, \ldots, n\}$ is the finite player set, for each $i \in N$ the set of player i's strategies is X_i (i.e. the set of player i's admissible choices) and $f_i : X_1 \times \cdots \times X_n \to \mathcal{R}$ is player i's payoff function ($\mathcal{R}$ is the set of real numbers). We suppose here that players are cost minimizing, so that player i has a cost $f_i(x_1, x_2, \ldots, x_n)$ when player 1 chooses $x_1 \in X_1$, player 2 chooses $x_2 \in X_2, \ldots$, player n chooses $x_n \in X_n$. We define $X = X_1 \times \cdots \times X_n$ and for $i \in N$: $X_{-i} = \Pi_{j \in N \setminus \{i\}} X_j$. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in X$ and $i \in N$. Sometimes we denote $\mathbf{x} = (x_i, \mathbf{x}_{-i})$, where $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Definition 1. A *Nash equilibrium* [1] for Γ is a strategy profile $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) \in X$ such that for any $i \in N$ and for any $x_i \in X_i$ we have that

$$f_i(\hat{\mathbf{x}}) \leq f_i(x_i, \hat{\mathbf{x}}_{-i}).$$

Such a solution is self-enforcing in the sense that once the players are playing such a solution, it is in every player's best interest to remain in his strategy. We denote by $\mathrm{NE}(\Gamma)$ the set of the Nash equilibrium strategy profiles.

Any $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n) \in \mathrm{NE}(\Gamma)$ is a vector such that for any $i \in N$, $\hat{x}_i$ is solution to the optimization problem

$$\min_{x_i \in X_i} f_i(x_i, \hat{\mathbf{x}}_{-i}).$$

Not always a game admits a Nash equilibrium solution. There are special situation in which this is true, for example in potential games.

Potential games have been introduced by Monderer and Shapley: the idea is that a game is said potential if the information that is sufficient to determine Nash equilibria can be summarized in a single function on the strategy space, the potential function [12, 14].

Definition 2. A game $\Gamma = \langle N; X_1, \ldots, X_n; f_1, \ldots, f_n \rangle$ is an *exact potential game* (or simply *potential game*) if there exists a function $V : \Pi_{i \in N} X_i \to \Re$ such that for each player $i \in N$, each strategy profile $x_{-i} \in \Pi_{j \in N \setminus \{i\}} X_j$ of i's opponents, and each pair $x_i, y_i \in X_i$ of strategies of player i:

$$f_i(y_i, \mathbf{x}_{-i}) - f_i(x_i, \mathbf{x}_{-i}) = V(y_i, \mathbf{x}_{-i}) - V(x_i, \mathbf{x}_{-i}).$$

The function V is called an exact potential (or, in short, a *potential*) of the game Γ. If V is a potential function of Γ, the difference induced by a single deviation is equal to that of the deviator's payoff function.

Clearly, by definition, the set of all strategy profiles that minimize V (called potential minimizers) is a subset of the Nash equilibrium set of the game Γ:

$$\operatorname{argmin}_{\mathbf{x} \in X} V(\mathbf{x}) \subseteq \mathrm{NE}(\Gamma).$$

This implies that for a potential game finding a Nash equilibrium means to solve an optimization problem.

The following theorem characterizes potential games [20].

Theorem 1. *$\Gamma = \langle N; X_1, \ldots, X_n; f_1, \ldots, f_n \rangle$ is a potential game with potential V iff*

$$f_i(x_1, \ldots, x_n) = V(x_1, \ldots, x_n) + d_i(x_{-i})$$

for some $d_i : X_{-i} \to \Re$ for any $i \in N$.

2.2 The Facility Location Game

We define the following n-player normal form game $\Gamma_n^{\mathrm{ED}} = \langle N; \Omega, \ldots, \Omega; f_1, \ldots, f_n \rangle$ where each player in $N = \{1, 2, \ldots, n\}$, for each $i \in N$, minimizes the cost $f_i : A \to \Re$ defined by

$$f_i(P_1, \ldots, P_n) = \sum_{1 \le j \le n, j \ne i} \frac{1}{d(P_i, P_j)} + \frac{1}{\sqrt{2d(P_i, \partial\Omega)}}$$

being $A = \left\{(P_1, \ldots, P_n) \in \Omega^n : P_i \in (]0,1[)^2,\ P_i \ne P_j\ \forall i, j = 1, \ldots, n, j \ne i\right\}$ and $d(x,y)$ is the Euclidean metric in $\Re^2$.

In terms of coordinates, if $P_i = (x_i, y_i), i \in N$ the distance of a point $P = (x,y)$ from the set $\partial\Omega$, the boundary of Ω, is

$$d(P, \partial\Omega) = \min_{Q \in \partial\Omega} d(P,Q) = \min\{x, y, 1-x, 1-y\}$$

and we have for $(x_1, y_1, \ldots, x_n, y_n) \in A$

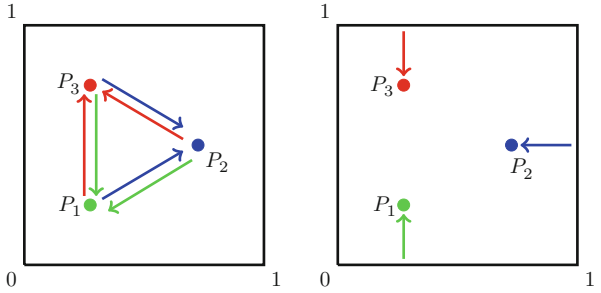


Fig. 4 Location problem with requirements of distance from both boundaries and each other player

$$f_i(x_1,y_1,\ldots,x_n,y_n) = \sum_{1\le j\le n, j\neq i} \frac{1}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}} + \frac{1}{\sqrt{2\min\{x_i,y_i,1-x_i,1-y_i\}}}.$$

The first $n-1$ terms in the definition of f_i give the distance between the point P_i and the rest of the points, the last term an increasing function of the distance of P_i from the boundary of the square. In Fig. 4 a graphical interpretation of both requirements is shown.

Definition 3. Any $(\hat{x}_1,\hat{y}_1,\ldots,\hat{x}_n,\hat{y}_n)\in A$ Nash equilibrium solution of the game Γ_n^{ED} is an optimal solution of the problem (ED). For any $i\in N$, $(\hat{x}_i,\hat{y}_i)$ is solution to the optimization problem

$$\min_{(x_i,y_i)\in\Omega} f_i(\hat{x}_1,\hat{y}_1,\ldots,\hat{x}_{i-1},\hat{y}_{i-1},x_i,y_i,\hat{x}_{i+1},\hat{y}_{i+1},\ldots,\hat{x}_n,\hat{y}_n)$$

with $(x_1,y_1,\ldots,x_n,y_n)\in A$.

The following theorem gives the existence of a solution to the problem (ED), namely of a Nash equilibrium solution of the game Γ_n^{ED}.

Theorem 2. *Γ_n^{ED} is a potential game and has at least a Nash equilibrium solution.*

Proof. By using Theorem 1, the function $V: A\to\Re$ defined by

$$V(x_1,y_1,\ldots,x_n,y_n) = \sum_{1\le i<j\le n} \frac{1}{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}} + \sum_{1\le i\le n} \frac{1}{\sqrt{2\min\{x_i,y_i,1-x_i,1-y_i\}}}$$

where $A=\left\{(P_1,\ldots,P_n)\in\Omega^n : P_i\in(]0,1[)^2,\ P_i\neq P_j\ \forall i,j=1,\ldots,n, j\neq i\right\}$ is a potential for the game Γ_n^{ED} and by using the Tonelli–Weierstrass theorem (Appendix)

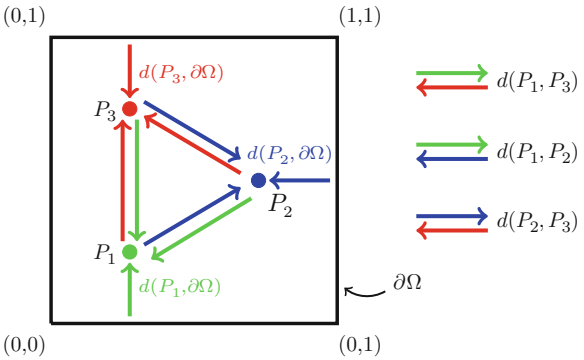


Fig. 5 Sketch for the game of location problem (ED)

admits at least a minimizer that is a Nash equilibrium solution of the game Γ_n^{ED} in the set A.

Any Nash equilibrium solution of the game Γ_n^{ED} is an optimal solution of the problem (ED), for which a sketch in Fig. 5 is illustrated.

2.3 The Constrained Facility Location Game

When there are logistic and economic limitation there is a constrained location case, such as that in which radar are to be located along defined grid lines, such as those of electricity feeding. We want to settle n points $P_1, P_2, \ldots, P_n$ in the square in such a way that they are far as possible from the rest and from the boundary of the square with an additional task. Now we have a constraint in the possible choices: the points $P_1, P_2, \ldots, P_n$ can be located only along prescribed lines $y = y_1, \ldots, y = y_k$ ($k \in \mathcal{N}$) of the region Ω.

As before, given the set $\{y_1, \ldots, y_k\}$ ($y_i \in]0,1[$, $i = 1, \ldots, k$) we define the following n-player normal form game $\Gamma_{n,k}^{\mathrm{ED}} = \langle N; \Omega, \ldots, \Omega; f_1, \ldots, f_n \rangle$ where each player in $N = \{1, 2, \ldots, n\}$, for each $i \in N$, minimizes the cost $f_i : B \to \Re$ defined by

$$f_i(P_1, \ldots, P_n) = \sum_{1 \le j \le n, j \ne i} \frac{1}{d(P_i, P_j)} + \frac{1}{\sqrt{2d(P_i, \partial\Omega)}}$$

being $B = \left\{(P_1, \ldots, P_n) \in A : \forall i \in \{1, \ldots, n\}\, \exists j \in \{1, \ldots, k\} \text{ s.t. } y_i = \tilde{y_j}\right\}$.

The following theorem gives the existence of a solution to the constrained problem (ED), namely of a Nash equilibrium solution of the game $\Gamma_{n,k}^{\mathrm{ED}}$.

Theorem 3. *$\Gamma_{n,k}^{ED}$ is a potential game and has at least a Nash equilibrium solution.*

Proof. The proof is similar as in Theorem 2 by considering the potential function V in the set B.

3 Results

The results summarized in this section are pure numerical, and they have been computed by a genetic algorithm for several cases, increasing the number of sensor devices to be located.

A genetic algorithm (GA) is an optimization technique based on the principles of genetics and natural selection. This technique is based on a population of individuals improved during generations by using several genetic operators, such as crossover and mutation, that combine the good features of each individual (crossover) to explore the search domain and to evolve the population to a state that minimizes the cost function. Each individual (or chromosome) represents a feasible solution in the search space. It's made by a string of bits, called chromosome, that may be divided in several genes, one for each problem variable or property.

A finite set of chromosomes make up a population. It can be viewed as a sampling of the problem domain that generation by generation maps zones with a higher probability of presence of the optimum.
A typical genetic algorithm consists of several steps:

- Population initialization: the algorithm in the first step randomly generates a set of solutions in the search space.
- Fitness computation: each individual is analysed to evaluate objective function and constraints. This procedure permits to sort the population for the following step.
- Selection: a probabilistic based selection of parents allows coupling of best individuals without wasting worst chromosomes, useful to move towards unexplored zones of search space.
- Crossover: on selected parents, a binary crossover operator is applied to create two new individuals.
- Mutation: to avoid premature stagnation of the algorithm a mutation operator is used, randomly changing a bit of the just created chromosomes.

The characteristics of the genetic algorithm employed for the solution of the location problem are summarized in Table 1. Each of the following results are intended to represent only one of the several solutions that differs only for the permutation of sensor locations. This reduces the number of evaluation of the location problem solutions proportional to the factorial of the number of sensors to be located.

Table 1 Genetic algorithm characteristics

Parameter	Value or type
Chromosome	Binary string
Crossover	Multi-cat
Mutation probability	0.01 %
Population size	100
Mating-pool	50

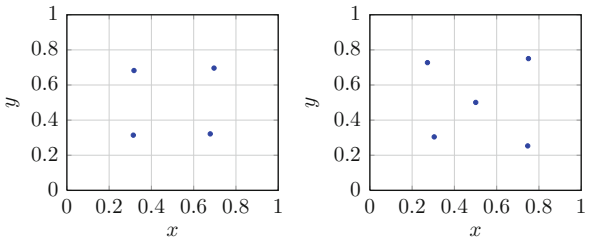


Fig. 6 Cases for $n = 4, 5$

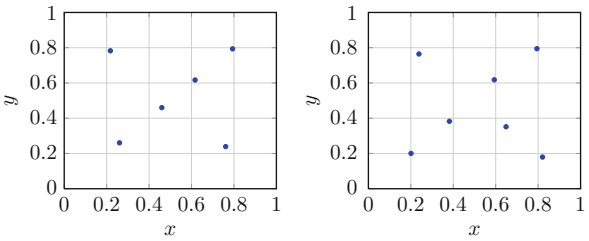


Fig. 7 Cases for $n = 6, 7$

In Figs. 6, 7, 8, 9, and 10 the results for unconstrained cases are shown.

In Figs. 11, 12, and 13 the comparison for unconstrained and constrained cases is shown, changing the number of rows in which the sensor is constrained case by case, depending on the results of the unconstrained cases.

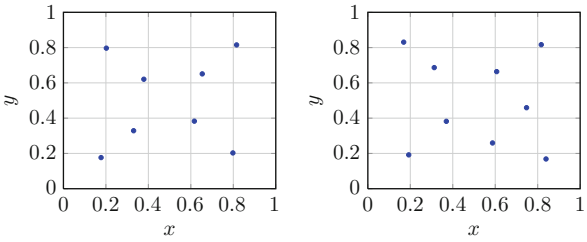


Fig. 8 Cases for $n = 8, 9$

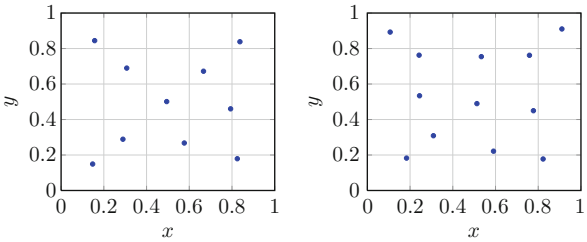


Fig. 9 Cases for $n = 10, 12$

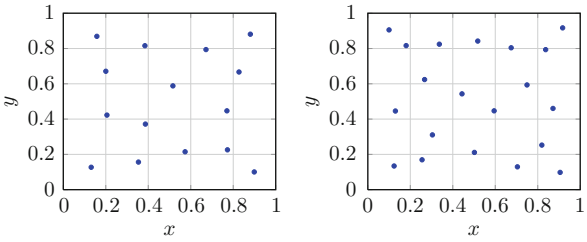


Fig. 10 Cases for $n = 15, 20$

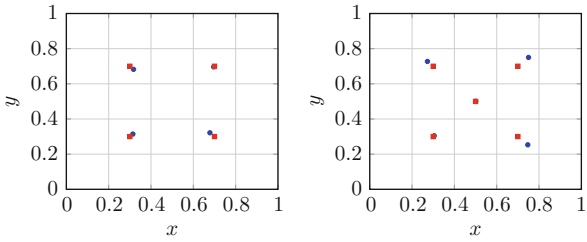


Fig. 11 Cases for $n = 4, 5$: *blue circles* unconstrained case, *red squares* constrained case

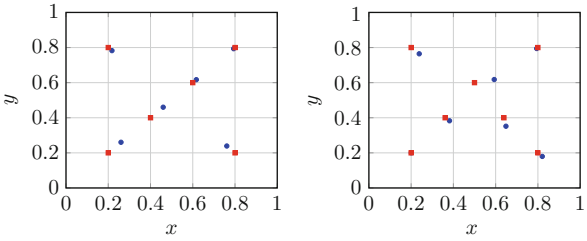


Fig. 12 Cases for $n = 6, 7$: *blue circles* unconstrained case, *red squares* constrained case

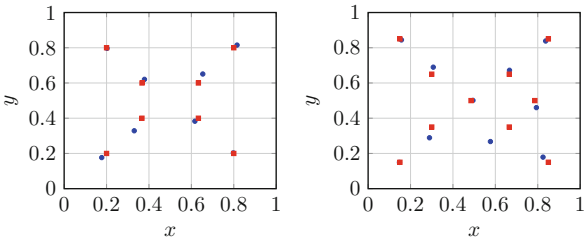


Fig. 13 Cases for $n = 8, 10$: *blue circles* unconstrained case, *red squares* constrained case

4 Conclusions

In this paper a game theoretical model to the experimental design is presented. This potential approach permits to avoid the computational difficulties due to the numerical evaluation of the Nash equilibria, reducing the problem to the research of the optimal point of a suitable objective function [12]. A numerical procedure based on a genetic algorithm has been used to compute results for several test cases using different number of sensors.

In results section, two models were considered: the constrained and the unconstrained one. The constrained model is a special case where the admissible region is made by a set of parallel segments due to operative constraints (for example, electricity lines).

An important improvement to this work could be using a domain with obstacles situation closer to real life application, extending the constrained model to handle convex obstacles. In a future paper this innovation will be investigated, also considering the generalization of the model to a 3D case.

Appendix

Definition 4. X a metric space, $F : X \mapsto [-\infty, +\infty]$. F is *coercive* if $\forall t \in \mathcal{R}\ \exists\, K(t) \subset X$, $K(t)$ compact s.t

$$\{x \in X : F(x) \leq t\} \subseteq K(t)$$

Theorem 4. *Tonelli–Weierstrass Theorem. X a metric space, $F : X \mapsto [-\infty, +\infty]$. Supposing F lower semi-continuous and coercive. Then F has a minimum in X.*

Proof. If $F(x) = +\infty\ \forall x \in X$ nothing is to be proved cause every $x \in X$ is a minimum point.

Supposing that $\inf_{x\in X} F(x) = m < +\infty$. Let $t > m$ and $K(t)$ compact. Than:

$$\inf_{x\in X} F(x) = \inf_{x\in K(t)} F(x)$$

Now, let $\{x_n\}_{n\in N}$ a minimizing succession. For n big enough we have $F(x_n) < t$ thus $x_n \in K(t)$ compact. Thus, it exists a sub-succession $\{x_{n_k}\}_{k\in N}$ that converges to a point $\bar{x} \in K(t)$.

Standing the lower semi-continuity:

$$F(\bar{x}) \leq \lim_{k\to\infty} \inf F(x_{n_k}) = \lim_{n\to\infty} F(x_n) = m$$

but then $F(\bar{x}) = m$ and $\bar{x}$ is a minimum point in $K(t)$ and thus in X.

References

1. Başar, T., Olsder, G.J.: Dynamic Noncooperative Game Theory. In: Classics in Applied Mathematics, **23**. Society for Industrial and Applied Mathematics, Philadelphia (1999)
2. Benabbou, A., Borouchaki, H., Laug, P., Lu, J.: Sphere packing and applications to granular structure modeling. In: Garimella, R.V. (ed.) Proceedings of the 17th International Meshing Roundtable, 12–15 October 2008. Springer, Berlin (2008)
3. Clarich, A., Periaux, J., Poloni, C.: Combining game strategies and evolutionary algorithms for CAD parametrization and multi-point optimization of complex aeronautic systems. EUROGEN, Barcelona (2003)
4. Conway, J.H., Sloane, N.J.A.: Sphere Packings, Lattices and Groups. Springer, Berlin (1998) [ISBN-13 978-0387985855]
5. D'Amato, E., Daniele, E., Mallozzi, L., Petrone, G.: Equilibrium strategies via GA to Stackelberg games under multiple follower best reply. Int. J. Intell. Syst. **27**, 74–85 (2012)
6. D'Amato, E., Daniele, E., Mallozzi, L., Petrone, G., Tancredi, S.: A hierarchical multi-modal hybrid Stackelberg-Nash GA for a leader with multiple followers game. In: Sorokin, A., Murphey, R., Thai, M.T., Pardalos, P.M. (eds.) Dynamics of Information Systems: Mathematical Foundations. Springer Proceedings in Mathematics & Statistics, **20**, pp. 267–280. Springer, New York (2012)
7. Dean, A., Voss, D.: Design and Analysis of Experiments, In: Springer Texts in Statistics, Springer Science+Business Media, New York, USA (1998) [ISBN 978-0387985619]
8. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 181–197 (2002)

9. Donev, A., Torquato, S., Stillinger, F.H., Connelly, R.: A linear programming algorithm to test for jamming in hard-sphere packings. J. Comput. Phys. **197**(1), 139–166 (2004). doi:10.1016/j.jcp.2003.11.022 [ISSN 0021-9991]
10. Hales, T.C.: The sphere packing problem. J. Comput. Appl. Math. **42**, 41–76 (1992)
11. Mallozzi, L.: Noncooperative facility location games. Oper. Res. Lett. **35**, 151–154 (2007)
12. Mallozzi, L.: An application of optimization theory to the study of equilibria for games: a survey. Cent. Eur. J. Oper. Res. **21**(3), 523–539 (2013)
13. Mallozzi, L., D'Amato, E., Daniele, E., Petrone, G.: N leader - M follower coalition games with genetic algorithms and applications. In: Poloni, C., Quagliarella, D., Periaux, J., Gauger, N., Giannakoglou, K. (eds.) Evolutionary and Deterministic Methods for Design, Optimization and Control. CIRA, Capua (2011)
14. Monderer, D., Shapley, L.S.: Potential Games. Games Econ. Behav. **14**, 124–143 (1996)
15. Nurmela, K.J.: Stochastic optimization methods in sphere packing and covering problems in discrete geometry and coding theory. Ph.D. Thesis, Helsinki University of Technology, printed by Picaset Oy (1997) [ISBN 952-90-9460-4]
16. Periaux, J., Chen, H.Q., Mantel, B., Sefrioui, M., Sui, H.T.: Combining game theory and genetic algorithms with application to DDM-nozzle optimization problems. Finite Elem. Anal. Des. **37**, 417–429 (2001)
17. Sloane, N.J.A.: The Sphere Packing Problem. 1998 Shannon Lecture. AT&T Shannon Lab, Florham Park (1998)
18. Sorokin, A., Pardalos, P. (eds.): Dynamic of Information Systems: Algorithmics Approaches. In: Springer Proceedings in Mathematics & Statistics, **51**. Springer Science+Business Media, New York, USA (2013)
19. Sutou, A., Dai, Y.: Global optimization approach to unequal sphere packing problems in 3D. J. Optim. Theory Appl. **114**(3), 671–694 (2002)
20. Voorneveld, M., Borm, P., van Megen, F., Tijs, S., Facchini, G.: Congestion games and potentials reconsidered. Int. Game Theory Rev. **1**(3–4), 283–299 (1999)
21. Wang, J.F., Periaux, J.: Multi-Point optimization using GAS and Nash/Stackelberg games for high lift multi-airfoil design in aerodynamics. In: Proceedings of the 2001 Congress on Evolutionary Computation, CEC2001 (May 2001), pp. 552–559 (2001)

A Supply Chain Network Game Theoretic Framework for Time-Based Competition with Transportation Costs and Product Differentiation

Anna Nagurney and Min Yu

1 Introduction

Supply chains today span the globe and provide the infrastructure for the production and delivery of goods and services, with more knowledgeable consumers demanding timely deliveries, despite, paradoxically, the great distances that may be involved. Indeed, delivery times are becoming a strategy, as important as productivity, quality, and even innovation (see, e.g., [10, 19, 28, 38, 52]). As noted by Ray and Jewkes [42], practitioners have realized that speed of product delivery is a competitive advantage [4, 47].

It is now well recognized (cf. [5, 15, 22]) that, whether in manufacturing (especially in build-to-order and made-on-demand industries such as certain computers, electronic equipment, specific cars, airplanes, and furniture) or in digitally based production and delivery (DVDs, online shopping, online content distribution, etc.) speed and consistency of delivery time are two essential components of customer satisfaction, along with price (cf. [1, 20]). Stalk, Jr., in his seminal *Harvard Business Review* 1988 article [46], "Time - The next source of competitive advantage," utilized the term *time-based competition*, to single out time as the major factor for sustained competitive advantage. Today, time-based competition has emerged as a paradigm for strategizing about and operationalizing supply chain networks in which efficiency and timeliness matter (see [8, 9, 25, 28, 49]).

A. Nagurney (✉)
Department of Operations and Information Management, Isenberg School of Management, University of Massachusetts, Amherst, MA 01003, USA

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden
e-mail: nagurney@isenberg.umass.edu

M. Yu
Pamplin School of Business Administration, University of Portland, Portland, OR 97203, USA
e-mail: yu@up.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_19,

Advances in production and operations management thought and practice, including such revolutionary concepts as time-based competition, have, in turn, provided a rich platform for the accompanying research investigations. The extensive literature review of Hum and Sim [22] of time-based competition emphasized both its intellectual history as well as the associated mathematical modeling, thereby, constructing a bridge between practice and scholarship on this important topic. They, nevertheless, concluded that much of the time-based focus in modeling was limited to the areas of transportation modeling, lead time and inventory modeling, and set-up time reduction analysis. Moreover, they argued that the literature emphasized cost minimization but what was needed was the explicit incorporation of time as a significant variable in modeling. The complexity of the production and operations management landscape in the real world could not be adequately captured through an objective function representing simply cost minimization. Gunasekaran and Ngai [18] further emphasized this shortcoming and the relevance of analyzing the trade-offs between operational costs and delivery time in supply chain management.

Hence, in order to rigorously capture time-based competition within an analytical, computable supply chain framework, one needs to utilize game theory and the appropriate strategic variables with the explicit recognition of time.

Li and Whang [24] developed an elegant game theory model for time-based competition in which firms choose, as their strategic variables, both prices and production rates and discussed several special cases. Their approach was a generalization of the contributions of Lederer and Li [23], who, in turn, built on some of the prior research in queuing and delays. However, since the focus in those papers was on operations management, and not on supply chain management, Li and Whang [24] did not consider the time component associated with the transportation of the products, which is a central issue in increasingly globalized supply chains (cf. [28]). In addition, the underlying functions were assumed to have an explicit structure. Moreover, they assumed that the firms were price-takers. In various industries, as noted above, in which made-to-order and build-to-order strategies are relevant, the underlying industrial organization is that of oligopolies and, imperfect, rather than perfect competition (cf. [48, 50]). Shang and Liu [43], in turn, investigated the impacts of promised delivery time and on-time delivery rates under oligopolistic competition. Blackburn [3] discussed some of the limits of time-based competition quantitatively through the introduction of the marginal value of time derived from a total cost objective function. However, he exclusively focused on inventory costs and did not include transportation costs which are fundamental to global supply chains. Moreover, a single cost-minimizing decision-maker was assumed, whereas in order to appropriately address time-based *competition*, a framework that captures the interactions among decision-makers, that is, firms, in a supply chains, along with the reality of product differentiation, is needed.

In this paper, hence, we develop a game theoretical framework for supply chain network time-based competition, which has the following major, novel features:

1. firms are assumed to be spatially separated and can compete both on the production side and on the demand side;

2. firms compete in an oligopolistic manner and, hence, influence the prices;
3. the time consumption of both production and transportation/shipment supply chain activities is made explicit;
4. the strategic variables of the firms are quantity variables and guaranteed delivery time variables;
5. consumers at the demand markets for the substitutable, but differentiated, products respond to both the quantities of the products and to their guaranteed delivery times, as reflected in the prices of the products.

In addition, by capturing the total cost associated with delivery times of each firm, along with their production costs and their transportation costs in their respective objective functions, the marginal cost of time can be quantified in this more general competitive network framework.

The intellectual platform upon which our model is based has several foundational supports. First, it builds upon the existing literature on oligopolistic competition and network equilibria (cf. [12, 27, 28]), coupled with the recent modeling advances that incorporate brand/product differentiation and supply chain network competition (see [26, 30, 32, 38]). However, unlike Nagurney and Yu [32], where the goal was to minimize total cost and total time in the supply chain network for a time-sensitive product, which, in that case, was fast fashion apparel, here we focus on the delivery times to the consumers at the demand market. In addition, in contrast to work noted above, in this paper, the consumers reflect their preferences for the different products through both the prices and the guaranteed delivery times, where the guaranteed delivery time here includes the time required for production and for transportation/shipment, with the understanding that different products will be distributed in an appropriate manner (digital products, e.g., are distributed via the web).

It is also important to recognize the literature on time-sensitive products from food to the, already noted, fashion apparel, to even perishable products in healthcare, as well as critical needs products in humanitarian operations; see Nagurney et al. [38] for such a survey. Finally, we note that although the book by Nagurney [28] contains a spectrum of dynamic supply chain network models the dynamics therein are modeled using projected dynamical systems (cf. [34]) without delivery times being explicit strategic variables.

To the best of our knowledge, this is the first paper to synthesize oligopolistic competition, product differentiation, and time-based competition, with guaranteed delivery times as strategic variables, in a computable supply chain network game theoretic model under Nash [39, 40] equilibrium.

For the reader, we also highlight the paper by Geunes and Pardalos [16] and the edited volume of theirs—Geunes and Pardalos [17], which provide excellent literature overviews of supply chain optimization with the latter also focusing on networks.

This paper is organized as follows. In Sect. 2 we develop the supply chain network game theoretic model with differentiated products and time-based competition by describing each firm's individual profit-maximizing behavior and the underlying cost functions and demand price functions, with an emphasis on the time element

and the network structure. We then define the governing supply chain Nash equilibrium and establish alternative variational inequality formulations. We emphasize that variational inequalities for supply chain network equilibrium problems were first utilized by Nagurney et al. [36] and initiated a rich literature. Recent applications have included also supply chain disruptions; see [31, 41, 51].

In Sect. 3, we focus on the variational inequality formulation that has elegant features for computations for which we propose an algorithm that yields, at each iteration, closed form expressions for the product shipments, the delivery times, and the associated Lagrange multipliers with the constraints for the latter. In Sect. 4, we illustrate the model through a series of numerical examples, which are solved using the algorithm. In Sect. 5, we conclude the paper with a summary of results and discussion.

2 The Supply Chain Network Game Theoretic Model with Product Differentiation and Guaranteed Delivery Times

In this section, we develop a supply chain network model with product differentiation in which the firms have as strategic variables their product shipments to the demand markets and the guaranteed times of the deliveries of the products. The firms compete under the Cournot-Nash equilibrium concept of non-cooperative behavior. The consumers, in turn, signal their preferences for the products through the demand price functions associated with the demand markets, which are, in general, functions of the demands for the products at all the demand markets as well as the guaranteed delivery times of the products, from the manufacturing/production stage to demand market delivery. We assume that there are m firms and n demand markets that can be located in different physical locations. There is a distinct (but substitutable) product produced by each of the m firms and is consumed at the n demand markets. Please refer to Fig. 1 for the underlying structure of the supply chain network problem under consideration here. The notation for the model is given in Table 1. The vectors are assumed to be column vectors. The equilibrium solution is denoted by "*."

The model is a strategic model rather than an operational-level model. Hence, we do not consider possible sequencing of jobs for specific demand markets. Such an extension may be considered in future research.

The following conservation of flow equations must hold:

$$s_i = \sum_{j=1}^{n} Q_{ij}, \quad i = 1, \ldots, m, \tag{1}$$

$$d_{ij} = Q_{ij}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n, \tag{2}$$

$$Q_{ij} \geq 0, \quad i = 1, \ldots, m; \; j = 1, \ldots, n. \tag{3}$$

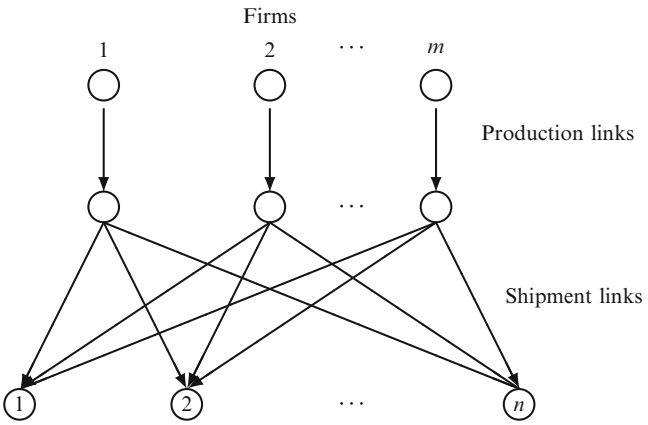


Fig. 1 The network structure of the supply chain problem

Table 1 Notation for the game theoretic supply chain network model with product differentiation and guaranteed delivery times

Notation	Definition
Q_{ij}	The nonnegative shipment of firm i's product to demand market j; $i = 1,\ldots,m$; $j = 1,\ldots,n$. We group the $\{Q_{ij}\}$ elements for firm i into the vector $Q_i \in R^n_+$ and all the firms' product shipments into the vector $Q \in R^{mn}_+$.
s_i	The nonnegative output produced by firm i; $i = 1,\ldots,m$. We group the firm production outputs into the vector $s \in R^m_+$.
d_{ij}	The demand for the product produced by firm i at demand market j; $i = 1,\ldots,m$; $j = 1,\ldots,n$. We group the demands into the vector $d \in R^{mn}_+$.
T_{ij}	The guaranteed delivery time of product i, which is produced by firm i, at demand market j; $i = 1,\ldots,m$; $j = 1,\ldots,n$. We group the delivery times of firm i into the vector $T_i \in R^n_+$ and then group all these vectors of all firms into the vector $T \in R^{mn}_+$.
$f_i(s)$	The production cost of firm i; $i = 1,\ldots,m$.
$g_i(T_i)$	The total cost associated with the delivery time of firm i; $i = 1,\ldots,m$.
$p_{ij}(d,T)$	The demand price of the product produced by firm i at demand market j; $i = 1,\ldots,m$; $j = 1,\ldots,n$.
$\hat{c}_{ij}(Q)$	The total transportation cost associated with shipping firm i's product to demand market j; $i = 1,\ldots,m$; $j = 1,\ldots,n$.

Consequently, the quantity of the product produced by each firm is equal to the sum of the amounts shipped to all the demand markets; the quantity of a firm's product consumed at a demand market is equal to the amount shipped from the firm to that demand market, and the product shipments must be nonnegative.

As noted in Sect. 1, the firms are also competing with time, that is, the guaranteed delivery times are strategic variables. Since each product must be manufactured and then delivered, as depicted in Fig. 1, we need to account for the time consumption associated with these supply chain network activities. Hence, associated with each firm and demand market pair, we also have the following constraint:

$$t_i s_i + h_i + t_{ij} Q_{ij} + h_{ij} \leq T_{ij}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n, \tag{4a}$$

where t_i, h_i, t_{ij}, and h_{ij} are all positive parameters. The first two terms in (4a) reflect the actual time consumption associated with producing product i and the second two terms reflect the actual time consumption associated with delivering product i to demand market j. Constraint (4a), thus, guarantees that the product of each firm i will be produced and shipped to demand market j within the guaranteed delivery time T_{ij} determined by firm i.

Note that, according to (4a), the supply chain network activities of production/manufacturing and transportation are functions, respectively, of how much is produced and of how much is transported. Indeed, it may take longer to produce a larger quantity of product and also (since the product may need to be loaded/unloaded) to deliver a larger volume of product to a demand point. The fixed terms h_i and h_{ij} denote the physical lower bounds of the time needed to produce and to transport product i to demand market j, respectively. Even in the case of digital products there will be a lower bound, albeit, small, in size. In light of (1), (4a) also ensures that the guaranteed delivery time strategic variables will be nonnegative. Furthermore, the total transportation cost functions $\hat{c}_{ij}$; $i = 1, \ldots, m$; $j = 1, \ldots, n$ since they, for the sake of generality, are functions of the product shipment pattern, capture possible congestion or competition for shipment resources (see also [28] and the references therein). Of course, a special case of (4a) and (4b) is when some (or all) of the parameters t_i; $i = 1, \ldots, m$ and t_{ij}; $i = 1, \ldots, m$; $j = 1, \ldots, n$ are identically equal to zero. The transportation costs that we consider, as a special case, capture the possibility of fixed transportation costs between firm and demand market pairs.

In view of (1), we may rewrite (4a) in product shipment variables only, that is,

$$t_i \sum_{j=1}^{n} Q_{ij} + h_i + t_{ij} Q_{ij} + h_{ij} \leq T_{ij}, \quad i = 1, \ldots, m; \; j = 1, \ldots, n. \tag{4b}$$

In our numerical examples, we illustrate different realizations of constraint (4b) in which we show that sometimes there may be a slack associated with (4b) in the equilibrium solution and sometimes not.

A firm's production cost may depend not only on its production output but also on that of the other firms. This is reasonable since firms which produce substitutable products may compete for the resources needed to produce their products. Also, in lieu of the time consumption [cf. (4a), (4b)] associated with producing a product the production costs $f_i(s)$; $i = 1, \ldots, m$, also capture the cost associated with the timely production of different levels of output. Due to the conservation of flow equation (1), we can define the production cost functions $\hat{f}_i$; $i = 1, \ldots, m$, in quantity shipments only, that is

$$\hat{f}_i = \hat{f}_i(Q) \equiv f_i(s), \quad i = 1, \ldots, m. \tag{5}$$

The production cost functions (5) are assumed to be convex and continuously differentiable.

It is important to emphasize that faster guaranteed delivery may be more costly, since it may require additional capacity and may be dependent on the operational

efficiency (cf. [5, 7, 37, 42, 44, 52]). For example, shipping costs of Amazon.com were doubled when the guaranteed delivery time was decreased from 1 week to 2 days [45]. This is captured in our functions g_i; $i = 1, \ldots, m$, which are also assumed to be convex and continuously differentiable.

In view of (2), we may define demand price functions, $\hat{p}_{ij}$, for all (i, j), in terms of the product shipments, that is:

$$\hat{p}_{ij} = \hat{p}_{ij}(Q, T) \equiv p_{ij}(d, T), \quad i = 1, \ldots, m; \; j = 1, \ldots, n. \tag{6}$$

We note that including both product quantities and guaranteed delivery time into demand functions has a tradition in economics as well as in operations research and marketing (cf. [5, 6, 21, 23, 42, 43, 45]) and the references therein). The demand price functions (6) and the total transportation cost functions $\hat{c}_{ij}$; $i = 1, \ldots, m$ and $j = 1, \ldots, n$, are assumed to be continuous and continuously differentiable.

Representing both the production cost (5) and the demand price functions (6) as functions of the product shipments, along with the time delivery constraints (4b) and the total cost function associated with the guaranteed delivery times, yields an elegant formulation of the supply chain network game with strategic variables being the product shipments and the delivery times, as we shall establish in Theorem 1.

The strategic variables of firm i are its product shipments $\{Q_i\}$ where $Q_i = (Q_{i1}, \ldots, Q_{in})$ and its guaranteed delivery times $\{T_i\}$, note that $T_i = (T_{i1}, \ldots, T_{in})$.

The profit or utility U_i of firm i; $i = 1, \ldots, m$, is, hence, given by the expression

$$U_i = \sum_{j=1}^{n} \hat{p}_{ij} Q_{ij} - \hat{f}_i - g_i - \sum_{j=1}^{n} \hat{c}_{ij}, \tag{7}$$

which is the difference between its total revenue and its total costs.

In view of (1)–(7), one may write the profit as a function solely of the shipment pattern and delivery times, that is,

$$U = U(Q, T), \tag{8}$$

where U is the m-dimensional vector with components: $\{U_1, \ldots, U_m\}$.

Let K^i denote the feasible set corresponding to firm i, where $K^i \equiv \{(Q_i, T_i) | Q_i \geq 0, \;$ and (4b) is satisfied for $i\}$ and $K \equiv \prod_{i=1}^{m} K^i$.

In the oligopolistic market mechanism, the m firms supply their products in a non-cooperative fashion, each one trying to maximize its own profit. We seek to determine an equilibrium product shipment and delivery time pattern (Q^*, T^*), according to the definition below (see also [11, 39, 40]).

Definition 1. A Supply Chain Network Equilibrium with Product Differentiation and Delivery Times

A product shipment and delivery time pattern $(Q^*, T^*) \in K$ is said to constitute a network equilibrium if for each firm $i; i = 1, \ldots, m$,

$$U_i(Q_i^*, T_i^*, \hat{Q}_i^*, \hat{T}_i^*) \geq U_i(Q_i, T_i, \hat{Q}_i^*, \hat{T}_i^*), \quad \forall (Q_i, T_i) \in K^i, \tag{9}$$

where

$$\hat{Q}_i^* \equiv (Q_1^*, \ldots, Q_{i-1}^*, Q_{i+1}^*, \ldots, Q_m^*); \quad \text{and} \quad \hat{T}_i^* \equiv (T_1^*, \ldots, T_{i-1}^*, T_{i+1}^*, \ldots, T_m^*). \tag{10}$$

According to (9), an equilibrium is established if no firm can unilaterally improve its profits by selecting an alternative vector of product shipments and delivery times of its product, given the decisions of the other firms.

2.1 Variational Inequality Formulations

We now derive alternative variational inequality formulations of the above supply chain network equilibrium with product differentiation in the following theorem.

Theorem 1. *Assume that for each firm i the profit function $U_i(Q,T)$ is concave with respect to the variables $\{Q_{i1}, \ldots, Q_{in}\}$, and $\{T_{i1}, \ldots, T_{in}\}$, and is continuous and continuously differentiable. Then $(Q^*, T^*) \in K$ is a supply chain network equilibrium according to Definition 1 if and only if it satisfies the variational inequality*

$$-\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial U_i(Q^*,T^*)}{\partial Q_{ij}} \times (Q_{ij} - Q_{ij}^*) - \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial U_i(Q^*,T^*)}{\partial T_{ij}} \times (T_{ij} - T_{ij}^*) \geq 0, \quad \forall (Q,T) \in K, \tag{11}$$

or, equivalently, $(Q^*, T^*, \gamma^*) \in K^1$ is an equilibrium product shipment and guaranteed delivery time pattern if and only if it satisfies the variational inequality

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial \hat{f}_i(Q^*)}{\partial Q_{ij}} + \sum_{l=1}^{n}\frac{\partial \hat{c}_{il}(Q^*)}{\partial Q_{ij}} - \sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q^*,T^*)}{\partial Q_{ij}} Q_{il}^* - \hat{p}_{ij}(Q^*,T^*) + \sum_{l=1}^{n}\gamma_{il}^* t_i + \gamma_{ij}^* t_{ij}\right] \\
&\times (Q_{ij} - Q_{ij}^*) + \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{\partial g_i(T_i^*)}{\partial T_{ij}} - \sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q^*,T^*)}{\partial T_{ij}} Q_{il}^* - \gamma_{ij}^*\right] \times (T_{ij} - T_{ij}^*) \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n}\left[T_{ij}^* - t_i\sum_{l=1}^{n} Q_{il}^* - t_{ij}Q_{ij}^* - h_i - h_{ij}\right] \times \left[\gamma_{ij} - \gamma_{ij}^*\right] \geq 0, \quad \forall (Q,T,\gamma) \in K^1,
\end{aligned} \tag{12}$$

where $K^1 \equiv \{(Q,T,\gamma) | Q \geq 0, T \geq 0, \gamma \geq 0\}$ with γ being the mn-dimensional vector with component (i,j) consisting of the element γ_{ij} corresponding to the Lagrange multiplier associated with the (i,j)-th constraint (4b).

Proof. Equation (11) follows directly from Gabay and Moulin [12, 14].

In order to obtain variational inequality (12), we note that, for a given firm i, under the imposed assumptions, (11) holds if and only if (see, e.g., [2]) the following holds:

$$\sum_{j=1}^{n}\left[\frac{\partial \hat{f}_i(Q_i^*)}{\partial Q_{ij}}+\sum_{l=1}^{n}\frac{\partial \hat{c}_{il}(Q^*)}{\partial Q_{ij}}-\sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q^*,T^*)}{\partial Q_{ij}}Q_{il}^*-\hat{p}_{ij}(Q^*,T^*)+\sum_{l=1}^{n}\gamma_{il}^*t_i+\gamma_{ij}^*t_{ij}\right]$$
$$\times(Q_{ij}-Q_{ij}^*)+\sum_{j=1}^{n}\left[\frac{\partial g_i(T_i^*)}{\partial T_{ij}}-\sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q^*,T^*)}{\partial T_{ij}}Q_{il}^*-\gamma_{ij}^*\right]\times(T_{ij}-T_{ij}^*)$$
$$+\sum_{j=1}^{n}\left[T_{ij}^*-t_i\sum_{l=1}^{n}Q_{il}^*-t_{ij}Q_{ij}^*-h_i-h_{ij}\right]\times\left[\gamma_{ij}-\gamma_{ij}^*\right]\geq 0,\quad \forall(Q_i,T_i,\gamma_i)\in K_i^1, \tag{13}$$

where $K_i^1\equiv\{(Q_i,T_i,\gamma_i)|(Q_i,T_i,\gamma_i)\in R_+^{3n}\}$, with $\{\gamma_i\}$=$(\gamma_{i1},\ldots,\gamma_{in})$.

But (13) holds for each firm i; $i=1,\ldots,m$, and, hence, the summation of (13) yields variational inequality (12). The conclusion follows. □

We now put variational inequality (12) into standard form (cf. [27]): determine $X^*\in\mathcal{K}\subset R^N$, such that

$$\langle F(X^*)^T,X-X^*\rangle\geq 0,\quad \forall X\in\mathcal{K}, \tag{14}$$

where F is a given continuous function from $\mathcal{K}$ to R^N, and $\mathcal{K}$ is a closed and convex set.

We define the $3mn$-dimensional vector $X\equiv(Q,T,\gamma)$ and the $3mn$-dimensional row vector $F(X)=(F^1(X),F^2(X),F^3(X))$ with the (i,j)-th component, F_{ij}^1, of $F^1(X)$ given by

$$F_{ij}^1(X)\equiv\frac{\partial \hat{f}_i(Q)}{\partial Q_{ij}}+\sum_{l=1}^{n}\frac{\partial \hat{c}_{il}(Q)}{\partial Q_{ij}}-\sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q,T)}{\partial Q_{ij}}\times Q_{il}-\hat{p}_{ij}(Q,T)+\sum_{l=1}^{n}\gamma_{il}t_i+\gamma_{ij}t_{ij}, \tag{15}$$

the (i,j)-th component, F_{ij}^2, of $F^2(X)$ given by

$$F_{ij}^2(X)\equiv\frac{\partial g_i(T_i)}{\partial T_{ij}}-\sum_{l=1}^{n}\frac{\partial \hat{p}_{il}(Q,T)}{\partial T_{ij}}\times Q_{il}-\gamma_{ij}, \tag{16}$$

and the (i,j)-th component, F_{ij}^3, of $F^3(X)$ given by

$$F_{ij}^3(X)=T_{ij}-t_i\sum_{l=1}^{n}Q_{il}-t_{ij}Q_{ij}-h_i-h_{ij}, \tag{17}$$

and with the feasible set $\mathcal{K}\equiv K$. Then, clearly, variational inequality (12) can be put into standard form (14).

We now present two examples in order to illustrate some of the above concepts and results.

2.2 Illustrative Examples

Consider a supply chain network oligopoly problem consisting of two firms and one demand market, as depicted in Fig. 2.

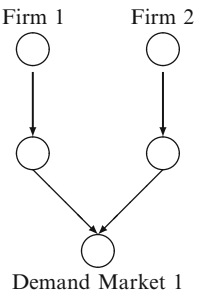


Fig. 2 The network structure for the illustrative examples

Example 1 We assume that these two firms are located in the same area. Both of them adopt similar technologies for the production and delivery of their highly substitutable products. Therefore, the production and transportation cost functions of Firms 1 and 2 are identical. Meanwhile, consumers at the demand market are indifferent between the products of Firms 1 and 2. The production cost functions are:

$$f_1(s) = 2s_1^2 + 3s_1, \quad f_2(s) = 2s_2^2 + 3s_2,$$

so that (cf. (5)):

$$\hat{f}_1(Q) = 2Q_{11}^2 + 3Q_{11}, \quad \hat{f}_2(Q) = 2Q_{21}^2 + 3Q_{21}.$$

The total transportation cost functions are:

$$\hat{c}_{11}(Q_{11}) = Q_{11}^2 + Q_{11}, \quad \hat{c}_{21}(Q_{21}) = Q_{21}^2 + Q_{21},$$

the total cost functions associated with delivery times are:

$$g_1(T_1) = T_{11}^2 - 30T_{11} + 400, \quad g_2(T_2) = T_{21}^2 - 40T_{21} + 450,$$

and the demand price functions are assumed to be:

$$\begin{aligned} p_{11}(d,T) =& 300 - 2d_{11} - 0.5d_{21} - T_{11} + 0.2T_{21}, \\ p_{21}(d,T) =& 300 - 2d_{21} - 0.5d_{11} - T_{21} + .2T_{11} \end{aligned}$$

so that [cf. (6)]:

$$\hat{p}_{11}(Q,T) = 300 - 2Q_{11} - 0.5Q_{21} - T_{11} + 0.2T_{21},$$
$$\hat{p}_{21}(Q,T) = 300 - 2Q_{21} - 0.5Q_{11} - T_{21} + .2T_{11}.$$

The above nonlinear cost functions, although hypothetical, were constructed to capture the potential resource competition and congestion in the production and delivery activities. Moreover, the total cost associated with delivery times decreases if the delivery time is increased in a certain range. However, the slower delivery may also be costly since resources could be used elsewhere.

The parameters associated with the production time consumption are:

$$t_1 = 0, \quad h_1 = 1, \quad t_2 = 0, \quad h_2 = 1,$$

and the parameters associated with the transportation time consumption are:

$$t_{11} = 0, \quad h_{11} = 1, \quad t_{21} = 0, \quad h_{21} = 1,$$

which means that the actual production times and the actual transportation times of these two firms are fixed.

Hence, for Firm 1, the following guaranteed delivery time constraint must be satisfied:

$$1 + 1 \leq T_{11},$$

and for Firm 2, the corresponding guaranteed delivery time constraint is:

$$1 + 1 \leq T_{21}.$$

The equilibrium product shipment and guaranteed delivery time pattern is:

$$Q_{11}^* = 28.14, \quad Q_{21}^* = 27.61, \quad T_{11}^* = 2.00, \quad T_{21}^* = 6.19,$$

and the corresponding Lagrange multipliers are:

$$\gamma_{11}^* = 2.14, \quad \gamma_{21}^* = 0.00.$$

Furthermore, the equilibrium prices associated with these two products are:

$$p_{11} = 229.15, \quad p_{21} = 224.91,$$

and the profits of the two firms are:

$$U_1 = 3{,}616.20, \quad U_2 = 3{,}571.90.$$

In this example, Firm 2's guaranteed delivery time, which is 6.19, is longer than the actual delivery time, which is 2, mainly because the total cost associated with delivery time would increase notably if Firm 2 were to reduce its guaranteed delivery time.

Example 2 This has the same data as Example 1 except that now the actual production times and the actual transportation times of Firms 1 and 2 depend on how much is produced and how much is shipped, respectively, that is,

$$t_1 = 0.2, \quad t_2 = 0.3, \quad t_{11} = 0.1, \quad t_{21} = 0.2.$$

The new equilibrium product shipment and guaranteed delivery time pattern are:

$$Q_{11}^* = 27.06, \quad Q_{21}^* = 26.13, \quad T_{11}^* = 10.12, \quad T_{21}^* = 15.07,$$

and the corresponding Lagrange multipliers are:

$$\gamma_{11}^* = 17.30, \quad \gamma_{21}^* = 16.26.$$

The equilibrium prices associated with these two products are:

$$p_{11} = 225.70, \quad p_{21} = 221.17,$$

and the profits of the two firms are:

$$U_1 = 3{,}603.89, \quad U_2 = 3{,}551.89.$$

This example shows that Firm 1 attracts more consumers with a notably shorter guaranteed delivery time, although the price of its product is higher than that of Firm 2's product. Due to its competitive advantage in delivery time performance, Firm 1 achieves a relatively higher profit.

Example 3 This has the same data as Example 2 except that now Firm 2 has reduced its production cost by improving its operational efficiency. The production cost function of Firm 2 is now given by:

$$f_2(s) = s_2^2 + 2s_2,$$

so that [cf. (5)]:

$$\hat{f}_2(Q_{21}) = Q_{21}^2 + 2Q_{21}.$$

The equilibrium product shipment and guaranteed delivery time pattern is:

$$Q_{11}^* = 26.86, \quad Q_{21}^* = 31.75, \quad T_{11}^* = 10.06, \quad T_{21}^* = 17.87,$$

and the corresponding Lagrange multipliers are:

$$\gamma_{11}^* = 16.97, \quad \gamma_{21}^* = 27.49.$$

The equilibrium prices associated with these two products are:

$$p_{11} = 223.93, \quad p_{21} = 207.22,$$

and the profits of the two firms are:

$$U_1 = 3{,}543.33, \quad U_2 = 4{,}413.00.$$

As a result of its lower production cost, Firm 2 is able to provide consumers with its product at an appealing price. Hence, the demand for Firm 2's product increases remarkably, even with a longer guaranteed delivery time, while there is a slight decrease in the demand for Firm 1's product. Therefore, in this example, Firm 2's profit improves significantly.

3 The Algorithm

The feasible set underlying variational inequality (12) is the nonnegative orthant, a feature that we will exploit for computational purposes. Specifically, we will apply the Euler-type method, which is induced by the general iterative scheme of Dupuis and Nagurney [13], where, at iteration τ of the Euler method (see also [34]) one must solve the following problem:

$$X^{\tau+1} = P_{\mathcal{K}}(X^{\tau} - a_{\tau}F(X^{\tau})), \tag{18}$$

where $P_{\mathcal{K}}$ is the projection on the feasible set $\mathcal{K}$ and F is the function that enters the variational inequality problem (12).

As demonstrated in Dupuis and Nagurney [13] and in Nagurney and Zhang [34], for convergence of the general iterative scheme, which induces this algorithmic scheme, the sequence $\{a_\tau\}$ must satisfy: $\sum_{\tau=0}^{\infty} a_\tau = \infty$, $a_\tau > 0$, $a_\tau \to 0$, as $\tau \to \infty$. Specific conditions for convergence of this scheme as well as various applications to the solutions of other supply chain and network oligopoly models can be found in Nagurney and Zhang [34], Nagurney et al. [35], Nagurney [29], Nagurney and Yu [33], and Nagurney et al. [37].

3.1 Explicit Formulae for the Euler Method Applied to the Supply Chain Network Model

The elegance of this procedure for the computation of solutions to our model with product differentiation and time deliveries can be seen in the following explicit formulae. In particular, we have the following closed form expression for all the product shipments $i = 1, \ldots, m; j = 1, \ldots, n$:

$$Q_{ij}^{\tau+1} = \max\{0, Q_{ij}^{\tau} + a_\tau(-F_{ij}^1(X^\tau))\}, \tag{19}$$

and the following closed form expression for all the guaranteed delivery time values $i = 1, \ldots, m$; $j = 1, \ldots, n$:

$$T_{ij}^{\tau+1} = \max\{0, T_{ij}^{\tau} + a_\tau(-F_{ij}^2(X^\tau))\}, \tag{20}$$

with the Lagrange multipliers being computed for all $i = 1, \ldots, m$; $j = 1, \ldots, n$ according to:

$$\gamma_{ij}^{\tau+1} = \max\{0, \gamma_{ij}^{\tau} + a_\tau(-F_{ij}^3(X^\tau))\}; \quad i = 1, \ldots, m; j = 1, \ldots, n. \tag{21}$$

In the next section, we apply the Euler method to compute solutions to additional numerical supply chain network problems.

4 Numerical Examples

We implemented the Euler method, as described in Sect. 3, using Matlab. The convergence criterion was $\varepsilon = 10^{-6}$; that is, the Euler method was considered to have converged if, at a given iteration, the absolute value of the difference of each product shipment, each guaranteed delivery time value, and each Lagrange multiplier differed from its respective value at the preceding iteration by no more than ε. We set the sequence $a_\tau = .1(1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots)$.

In this section, we considered a supply chain network consisting of three firms and two demand markets, which are geographically separated (as depicted in Fig. 3). Consumers at Demand Market 2 are more sensitive with respect to guaranteed delivery times than consumers at Demand Market 1.

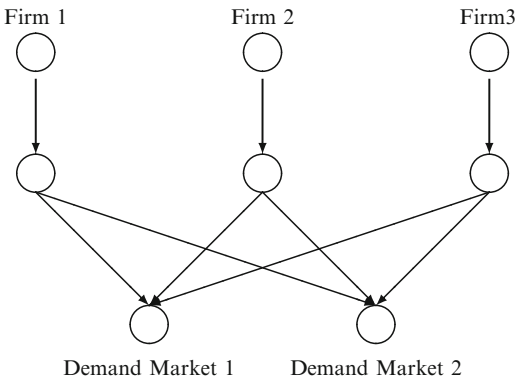


Fig. 3 The network structure for the numerical examples

Example 4 The cost functions, demand price functions, and parameters associated with time consumption are as follows:

Firm 1:

$$f_1(s) = s_1^2 + 0.5s_1s_2 + 0.5s_1s_3, \quad g_1(T_1) = T_{11}^2 + T_{12}^2 - 30T_{11} - 40T_{12} + 650,$$
$$\hat{c}_{11}(Q_{11}) = Q_{11}^2 + 0.5Q_{11}, \quad \hat{c}_{12}(Q_{12}) = Q_{12}^2 + Q_{12},$$
$$p_{11}(d,T) = 400 - 2d_{11} - d_{21} - 0.8d_{31} - 1.2T_{11} + 0.3T_{21} + 0.2T_{31},$$
$$p_{12}(d,T) = 400 - 1.5d_{12} - 0.5d_{22} - 0.8d_{32} - 2T_{12} + 0.2T_{22} + 0.3T_{32},$$
$$t_1 = 0.8, \quad h_1 = 1.5, \quad t_{11} = 0.4, \quad h_{11} = 1.5, \quad t_{12} = 0.5, \quad h_{12} = 1.5;$$

Firm 2:

$$f_2(s) = 1.5s_2^2 + 0.8s_1s_2 + 0.8s_2s_3, \quad g_2(T_2) = T_{21}^2 + T_{22}^2 - 30T_{21} - 30T_{22} + 480,$$
$$\hat{c}_{21}(Q_{21}) = Q_{21}^2 + Q_{21}, \quad \hat{c}_{22}(Q_{22}) = Q_{22}^2 + Q_{22},$$
$$p_{21}(d,T) = 400 - 2d_{21} - d_{11} - d_{31} - 1.2T_{21} + 0.2T_{11} + 0.2T_{31},$$
$$p_{22}(d,T) = 400 - 1.5d_{22} - 0.5d_{12} - 0.5d_{32} - 2T_{22} + 0.3T_{12} + 0.3T_{32},$$
$$t_2 = 0.6, \quad h_2 = 1.5, \quad t_{21} = 0.4, \quad h_{21} = 1.3, \quad t_{22} = 0.4, \quad h_{22} = 1.3;$$

Firm 3:

$$f_3(s) = 2s_3^2 + 0.8s_1s_3 + 0.8s_2s_3, \quad g_3(T_3) = 0.8T_{31}^2 + 0.8T_{32}^2 - 25T_{31} - 20T_{32} + 400,$$
$$\hat{c}_{31}(Q_{31}) = 1.5Q_{31}^2 + Q_{31}, \quad \hat{c}_{32}(Q_{32}) = Q_{32}^2 + 1.5Q_{32},$$
$$p_{31}(d,T) = 400 - 2d_{31} - 0.8d_{11} - d_{21} - 1.2T_{31} + 0.2T_{11} + 0.3T_{21},$$
$$p_{32}(d,T) = 400 - 1.5d_{32} - 0.8d_{12} - 0.5d_{22} - 2T_{32} + 0.3T_{12} + 0.2T_{22},$$
$$t_3 = 0.3, \quad h_3 = 1, \quad t_{31} = 0.2, \quad h_{31} = 1, \quad t_{32} = 0.1, \quad h_{32} = 1.$$

We utilized (5) and (6) to construct the production cost functions and the demand price functions, respectively, in shipment variables, for all examples in this section.

The equilibrium product shipment and guaranteed delivery time pattern, the Lagrange multipliers, and the prices are reported in Tables 2 and 3.

Note that, in Example 4, Firm 1 has a slight advantage over its competitors in Demand Market 1, despite the longer guaranteed delivery time, perhaps as a consequence of the lower price. Firm 3 captures the majority of the market share at Demand Market 2, due to consumers' preference for timely delivery. However, Firm 2 attains the lowest profit, as compared to its rivals, since Firm 2 is neither cost-effective enough nor sufficiently time-efficient.

Example 5 This has the identical data to that in Example 4, except that consumers at Demand Market 2 are becoming even more time-sensitive. The new demand price functions are now given by:

$$p_{12}(d,T) = 400 - 1.5d_{12} - 0.5d_{22} - 0.8d_{32} - 3T_{12} + 0.2T_{22} + 0.3T_{32},$$
$$p_{22}(d,T) = 400 - 1.5d_{22} - 0.5d_{12} - 0.5d_{32} - 3T_{22} + 0.3T_{12} + 0.3T_{32},$$
$$p_{32}(d,T) = 400 - 1.5d_{32} - 0.8d_{12} - 0.5d_{22} - 3T_{32} + 0.3T_{12} + 0.2T_{22}.$$

We also provided the solutions to Example 5 in Tables 2 and 3.

Table 2 The equilibrium product shipment and guaranteed delivery time patterns, the Lagrange multipliers, and the prices for Examples 4 and 5

		Example 4				Example 5			
Firm	Demand market	Q^*	T^*	γ^*	p	Q^*	T^*	γ^*	p
1	1	18.05	36.44	64.54	302.76	19.30	35.34	63.84	299.96
	2	14.73	36.59	62.65	288.15	11.48	33.36	61.15	268.47
2	1	15.96	29.10	47.36	308.78	17.01	28.32	47.04	305.76
	2	17.23	29.61	63.67	311.74	14.19	27.19	66.94	295.34
3	1	17.14	17.63	23.78	330.18	17.47	17.24	23.55	327.49
	2	23.55	16.56	53.60	328.05	21.69	15.91	70.54	318.91

Table 3 The profits of Firms 1, 2, and 3 in Examples 4 and 5

	Firm 1	Firm 2	Firm 3
Example 4	6,097.14	5,669.63	6,782.11
Example 5	5,697.97	5,3072.64	6,560.58

In Example 5, Firms 1 and 3 still dominate Demand Markets 1 and 2, respectively. Consumers' increasing time sensitivity at Demand Market 2 has forced all these three firms to shorten their guaranteed delivery times. The decrease in Firm 3's profit is negligible, while the profits of Firms 1 and 2 shrink notably. The results in Examples 4 and 5 suggest that delivery times, as a strategy, are particularly influential in time-based competition.

5 Summary and Conclusions

In this paper, we developed a rigorous modeling and computational framework for time-based competition in supply chain networks using game theory and variational inequality theory.

Specifically, the firms are assumed to compete in an oligopolistic manner using as strategic variables not only their product shipments to the various demand markets, under brand differentiation, but also their guaranteed delivery times. Here the guaranteed delivery times provide upper bounds on the sum of the production time and the transportation time between the firm and demand market pairs. All firms are assumed to be profit-maximizers and subject to production and transportation costs. The consumers, in turn, reflect their preferences for the firms' brands or products through the demand price functions which are functions of not only the demands for the firms' products at the different demand markets but also their guaranteed delivery times.

Numerical supply chain network examples were presented to illustrate the generality of the proposed model with a complete reporting of the input data and the computed equilibrium product shipments and guaranteed delivery times, along with the Lagrange multipliers associated with the delivery time constraints.

The modeling and analytical framework can be used as the foundation for the investigation of supply chain networks in the case of build to order and made on demand products. It can also be extended in several directions through the inclusion of multiple options of transportation and multiple technologies for production. One may also incorporate additional tiers of suppliers. Nevertheless, we have laid out the foundations for time-based competition in supply chain networks with this study that enables numerous explorations both theoretical and empirical with a focus on particular industrial sectors.

Acknowledgements The first author's research was supported, in part, by the School of Business, Economics and Law at the University of Gothenburg through its Visiting Professor Program.

This research was also supported, in part, by the National Science Foundation (NSF) grant CISE #1111276, for the NeTS: Large: Collaborative Research: Network Innovation Through Choice project awarded to the University of Massachusetts Amherst.

The above support is gratefully acknowledged.

The authors dedicate this paper to Professor Panos M. Pardalos whose scholarship, leadership, and friendship have made our community richer in numerous ways.

References

1. Ballou, R.H.: Business Logistics Management. Prentice Hall, Upper Saddle River (1998)
2. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. Prentice Hall, Englewood Cliffs (1989)
3. Blackburn, J.D.: Valuing time in supply chains: establishing limits of time-based competition. J. Oper. Manag. **30**, 396–405 (2012)
4. Blackburn, J.D., Elrod, T., Lindsley, W.B., Zahorik, A.J.: The strategic value of response time and product variety. In: Voss, C.A. (ed.) Manufacturing Strategy: Process and Content, pp. 261–281. Chapman and Hall, London (1992)
5. Boyaci, T., Ray, S.: Product differentiation and capacity cost interaction in time and price sensitive markets. Manuf. Serv. Oper. Manag. **5**(1), 18–36 (2003)
6. Boyaci, T., Ray, S.: The impact of capacity costs on product differentiation in delivery time, delivery reliability, and price. Prod. Oper. Manag. **15**(2), 179–197 (2006)

7. Cachon, G.P., Zhang, F.: Procuring fast delivery: sole sourcing with information asymmetry. Manag. Sci. **52**(6), 881–896 (2006)
8. Carter, P.L., Melnyk, S.A., Handfield, R.B.: Identifying the basic process strategies for time-based competition. Prod. Inv. Manag. J. **36**(1), 65–70 (1995)
9. Ceglarek, D., Huang, W., Zhou, S., Ding, Y., Kuma, R., Zhou, Y.: Time-based competition in multistage manufacturing: stream-of-variation analysis (SOVA) methodology – review. Int. J. Flex. Manuf. Syst. **16**, 11–44 (2004)
10. Christopher, M.: Logistics and supply chain management: creating value-adding networks, 3rd edn. Prentice-Hall/Financial Times, London (2005)
11. Cournot, A.A.: Researches into the Mathematical Principles of the Theory of Wealth (English translation). MacMillan, London (1838)
12. Dafermos, S., Nagurney, A.: Oligopolistic and competitive behavior of spatially separated markets. Reg. Sci. Urban Econ. **17**, 245–254 (1987)
13. Dupuis, P., Nagurney, A.: Dynamical systems and variational inequalities. Ann. Oper. Res. **44**, 9–42 (1993)
14. Gabay, D., Moulin, H.: On the uniqueness and stability of Nash equilibria in noncooperative games. In: Bensoussan, A., Kleindorfer, P., Tapiero, C.S. (eds.) Applied Stochastic Control of Econometrics and Management Science, pp. 271–294. North-Holland, Amsterdam (1980)
15. Geary, S., Zonnenberg, J.P.: What it means to be best in class. Supply Chain Manag. Rev. **4**, 42–49 (2000)
16. Geunes, J., Pardalos, P.M.: Network optimization in supply chain management and financial engineering: an annotated bibliography. Networks **42**(2), 66–84 (2003)
17. Geunes, J., Pardalos, P.M. (eds.): Supply Chain Optimization. Springer, New York (2005)
18. Gunasekaran, A., Ngai, E.W.T.: Build-to-order supply chain management: a literature review and framework for development. J. Oper. Manag. **23**, 423–451 (2005)
19. Gunasekaran, A., Patel, C., McGaughey, R.E.: A framework for supply chain performance measurement. Int. J. Prod. Econ. **87**, 333–347 (2004)
20. Handfield, R.B., Pannesi, R.T.: Antecedents of lead-time competitiveness in make-to-order manufacturing firms. Int. J. Prod. Res. **33**(2), 511–537 (1995)
21. Hill, A.V., Khosla, I.S.: Models for optimal lead time reduction. Prod. Oper. Manag. **1**(2), 185–197 (1992)
22. Hum, S.H., Sim, H.H.: Time-based competition: literature review and implications for modelling. Int. J. Oper. Prod. Manag. **16**, 75–90 (1996)
23. Lederer, P.J., Li, L.: Pricing, production, scheduling and delivery-time competition. Oper. Res. **45**(3), 407–420 (1997)
24. Li, L., Whang, S.: Game theory models in operations management and information systems. In: Chatterjee, K., Samuelson, W.F. (eds.) Game Theory and Business Applications, pp. 95–131. Kluwer Academic, Dordrecht (2001)
25. Li, S., Ragu-Nathan, B., Ragu-Nathan, T.S., Subba Rao, S.: The impact of supply chain management practices on competitive advantage and organizational performance. Omega-Int. J. Manag. Sci. **34**(2), 107–124 (2006)
26. Masoumi, A.H., Yu, M., Nagurney, A.: A supply chain generalized network oligopoly model for pharmaceuticals under brand differentiation and perishability. Transport. Res. E **48**, 762–780 (2012)
27. Nagurney, A.: Network Economics: A Variational Inequality Approach, 2nd and rev. edn. Kluwer Academic, Boston (1999)
28. Nagurney, A.: Supply Chain Network Economics: Dynamics of Prices, Flows, and Profits. Edward Elgar Publishing, Cheltenham (2006)
29. Nagurney, A.: Supply chain network design under profit maximization and oligopolistic competition. Transport. Res. E **46**, 281–294 (2010)
30. Nagurney, A., Li, D.: A dynamic spatial oligopoly model with transportation costs, product differentiation, and quality competition. Comput. Econ. (2014, in press)
31. Nagurney, A., Qiang, Q.: Fragile Networks: Identifying Vulnerabilities and Synergies in an Uncertain World. Wiley, Hoboken (2009)

32. Nagurney, A., Yu, M.: Fashion supply chain management through cost and time minimization from a network perspective. In: Choi, T.M. (ed.) Fashion Supply Chain Management: Industry and Business Analysis, pp. 1–20. IGI Global, Hershey (2011)
33. Nagurney, A., Yu, M.: Sustainable fashion supply chain management under oligopolistic competition and brand differentiation. Int. J. Prod. Econ. **135**, 532–540 (2012)
34. Nagurney, A., Zhang, D.: Projected dynamical systems and variational inequalities with applications. Kluwer Academic, Boston (1996)
35. Nagurney, A., Dupuis, P., Zhang, D.: A dynamical systems approach for network oligopolies and variational inequalities. Ann. Reg. Sci. **28**, 263–283 (1994)
36. Nagurney, A., Dong, J., Zhang, D.: A supply chain network equilibrium model. Transport. Res. E **38**, 281–303 (2002)
37. Nagurney, A., Yu, M., Qiang, Q.: Supply chain network design for critical needs with outsourcing. Pap. Reg. Sci. **90**, 123–142 (2011)
38. Nagurney, A., Yu, M., Masoumi, A.H., Nagurney, L.S.: Networks Against Time: Supply Chain Analytics for Perishable Products. Springer Science + Business Media, New York (2013)
39. Nash, J.F.: Equilibrium points in n-person games. Proc. Natl. Acad. Sci. USA **36**, 48–49 (1950)
40. Nash, J.F.: Noncooperative games. Ann. Math. **54**, 286–298 (1951)
41. Qiang, Q., Nagurney, A., Dong, J.: Modeling of supply chain risk under disruptions with performance measurement and robustness analysis. In: Wu, T., Blackhurst, J. (eds.) Managing Supply Chain Risk and Vulnerability: Tools and Methods for Supply Chain Decision Makers, pp. 91–111. Springer, Berlin (2009)
42. Ray, S., Jewkes, E.M.: Customer lead time management when both demand and price are lead time sensitive. Eur. J. Oper. Res. **153**, 769–781 (2004)
43. Shang, W., Liu, L.: Promised delivery time and capacity games in time-based competition. Manag. Sci. **57**(3), 599–610 (2011)
44. So, K.C.: Price and time competition for service delivery. Manuf. Serv. Oper. Manag. **2**, 392–409 (2000)
45. So, K.C., Song, J.S.: Price, delivery time guarantees and capacity selection. Eur. J. Oper. Res. **111**, 28–49 (1998)
46. Stalk, G., Jr.: Time – the next source of competitive advantage. Harv. Bus. Rev. **66**(4), 41–51 (1988)
47. Stalk, G., Jr., Hout, T.M.: Competing Against Time: How Time-Based Competition is Reshaping Global Markets. Free Press, New York (1990)
48. Tirole, J.: The Theory of Industrial Organization. MIT Press, Cambridge (1988)
49. Vickery, S.K., Droge, C.L.M., Yeomans, J.M., Markland, R.E.: Time-based competition in the furniture industry. Prod. Inv. Manag. J. **36**(4), 14–21 (1995)
50. Vives, X.: Oligopoly Pricing: Old Ideas and New Tools. MIT Press, Cambridge (1999)
51. Wakolbinger, T., Cruz, J.: Supply chain disruption risk management through strategic information acquisition and sharing and risk sharing contracts. Int. J. Prod. Res. **49**(13), 4063–4084 (2011)
52. Yu, M.: Analysis, design, and management of supply chain networks with applications to time-sensitive products. Dissertation, University of Massachusetts, Amherst (2012)

On the Discretization of Pseudomonotone Variational Inequalities with an Application to the Numerical Solution of the Nonmonotone Delamination Problem

Nina Ovcharova and Joachim Gwinner

1 Introduction

While the existence of solutions of pseudomonotone variational inequalities under appropriate coerciveness conditions is well known [7, 14], the discretization theory is more challenging since it bridges the gap to the appropriate numerical analysis. In other words, the more interesting question is how to discretize such problems in order to find their solutions by using efficient numerical methods. In this paper we present a novel approximation method for pseudomonotone variational inequalities applicable to regularization of nonsmooth functionals and further finite element discretization. For this approximation procedure a convergence analysis is established. We also turn our attention back to the concept of pseudomonotonicity due to Brézis and show that this property is a weaker condition than the pseudomonotonicity of the functional that is defined by means of set-valued operators with nonvoid convex closed bounded values. However, by both conditions the existence of solution can be guaranteed.

We continue with hemivariational inequalities in linear elasticity involving nonmonotone contact of elastic bodies. Such problems arise in unilateral contact of an elastic body with a rigid foundation or in bilateral contact between elastic bodies, when friction cannot be neglected and is modelled by a nonmonotone friction law. We point out also delamination problems for laminated composite structures under loading where two bodies are in adhesive contact, i.e., they are glued on a surface by an adhesive material. We show that such nonmonotone laws can be modelled with pseudomonotone functionals for those our approximation method can be applied. As an application we study a two-dimensional delamination problem. We illustrate our theoretical results and present some numerical results. More detailed, we first

N. Ovcharova (✉) • J. Gwinner
Department of Aerospace Engineering, Institute of Mathematics,
Universität der Bundeswehr München, Munich, Germany
e-mail: nina.ovcharova@unibw.de; joachim.gwinner@unibw.de

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_20,

regularize the nonmonotone law describing the behaviour of the binding interlayer material in the normal direction on the contact boundary and then use a finite element discretization of the regularized problem [19, 20]. Altogether, we apply our approximation scheme with two parameters (ε, h), where $\varepsilon > 0$ is a regularization parameter and h is a mesh parameter of discretization.

Concerning the mathematical study of hemivariational inequalities and their discretization by finite elemente methods we refer the reader, respectively, to the monographs [8, 13, 18, 20, 21, 23] and [16]. For discretization and numerical realization of contact problems with nonmonotone friction and delamination, see e.g. [4, 5].

2 An Approximation Scheme for Pseudomonotone Variational Inequalities

Let $(V, \|\cdot\|_V)$ be a real reflexive Banach space and $K \subseteq V$ a closed, convex nonempty set. Let $\psi : K \times K \to \mathbb{R}$ be a given functional such that $\psi(\cdot, v)$ is upper semicontinuous on each finite dimensional part of K. For a fixed linear form $g \in V^*$ we consider the variational inequality $(\mathcal{P})$: Find $u \in K$ such that

$$\psi(u, v) \geq \langle g, v - u \rangle \quad \forall v \in K. \tag{1}$$

We assume that the functional ψ is pseudomonotone in the sense that

(PM) for any sequence $\{u_n\}$ in K,

$$u_n \rightharpoonup u \quad \text{and} \quad \liminf_{n\to\infty} \psi(u_n, u) \geq 0$$

implies that for any $v \in K$

$$\psi(u, v) \geq \limsup_{n\to\infty} \psi(u_n, v)$$

holds.

A simple example of pseudomonotone function is $\psi(u, v) = f(v) - f(u)$, where f is a weakly lower semicontinuous function.

Further, we assume also *asymptotic coercivity* in the sense that there exist $v_0 \in K$ such that

$$\lim_{\|u\|_V \to \infty,\, u \in K} \frac{-\psi(u, v_0)}{\|u - v_0\|_V} = \infty.$$

Then, the existence of a solution to (1) can be guaranteed by the existence theory in [14].

Remark 1. The definition (PM) of pseudomonotonicity is motivated by a topologically pseudomonotone operators $T : V \to V^*$ in the sense of Brézis [7]. Indeed, $\Psi : V \times V \to \mathbb{R}$ defined by

$$\Psi(u,v) = \langle Tu, v-u\rangle$$

is pseudomonotone if and only if T is pseudomonotone operator.

Remark 2. Let now $T : V \Rightarrow V^*$ be a set-valued mapping with nonempty convex closed bounded values. Then T is said to be Brézis-pseudomonotone if the following conditions hold:

(a) T is upper semicontinuous from each finite dimensional subspace of V to the weak topology on V^*;
(b) for each sequence $\{u_n\}$ in V and $u_n^* \in T(u_n)$ such that $u_n \rightharpoonup u$ and

$$\limsup \langle u_n^*, u_n - u\rangle \le 0$$

it follows that for any $v \in V$ there exists $u^*(v) \in T(u)$ such that

$$\liminf \langle u_n^*, u_n - v\rangle \ge \langle u^*(v), u - v\rangle.$$

As we see below, the Brézis-pseudomonotonicity of T is a weaker condition than the pseudomonotonicity of the functional $\Psi : V \times V \to \mathbb{R}$ that is defined by

$$\Psi(u,v) := \max\{\langle w, v-u\rangle : w \in T(u)\}. \tag{2}$$

Indeed, let $\{u_n\}$ in V be such that $u_n \rightharpoonup u$ and $\liminf \Psi(u_n,u) \ge 0$. Since V is a reflexive Banach space and $T(u_n)$ is a nonempty, convex, closed and bounded subset of V^* it follows that $T(u_n)$ is weakly compact and therefore there exists $u_n^* \in T(u_n)$ such that

$$\max_{w\in T(u_n)} \langle w, u-u_n\rangle = \langle u_n^*, u-u_n\rangle.$$

Hence, for any $u_n \in V$ there exists $u_n^* \in T(u_n)$ such that

$$\liminf \Psi(u_n,u) = \liminf\{\max_{w\in T(u_n)} \langle w, u-u_n\rangle\} = \liminf \langle u_n^*, u-u_n\rangle \ge 0.$$

Since T is pseudomonotone then for each $v \in V$ there exists $u^*(v) \in T(u)$ such that

$$\liminf \langle u_n^*, u_n - v\rangle \ge \langle u^*(v), u-v\rangle,$$

which is equivalent to

$$\limsup \langle u_n^*, v-u_n\rangle \le \langle u^*(v), v-u\rangle. \tag{3}$$

To establish the pseudomonotonicity of Ψ we have to verify

$$\Psi(u,v) \ge \limsup_{n\to\infty} \Psi(u_n,v) \quad \forall v \in V.$$

More detailed, we have to check if there exists $u^* \in T(u)$ such that there holds

$$\langle u^*, v-u\rangle \ge \limsup \langle w_n, v-u_n\rangle \quad \forall v \in V, \quad \forall w_n \in T(u_n).$$

But this inequality cannot be guaranteed for all $w_n \in T(u_n)$ as inequality (3) shows. Nevertheless, by both conditions one can guarantee the existence of solution, see [14].

Now we investigate the approximation of (1) by a family of finite-dimensional variational inequalities. Let T be a directed set, $\{V_t\}$ a family of finite-dimensional subspaces of V and $\{K_t\}$ a family of closed convex nonempty subsets of V_t such that $K_t \subset K$. Without the loss of generality we can always assume that $0 \in K$. Since $0 \in K$, we assume also that $0 \in K_t$ for all $t \in T$. We require the following hypotheses:

(H0) for any $v \in K$ there exists a net $\{v_t\}$ such that $v_t \in K_t$ and $v_t \to v$ in V;
(H1) ψ_t is pseudomonotone;
(H2) for any nets $\{u_t\}$ and $\{v_t\}$ from K_t such that $u_t \rightharpoonup u$ and $v_t \to v$ in V it follows that

$$\limsup_{t \in T} \psi_t(u_t, v_t) \leq \psi(u, v);$$

(H3) the family $\{-\psi_t(\cdot, 0)\}$ is uniformly bounded from below in the sense that there exist constants $c > 0$, d, $d_0 \in \mathbb{R}$ and $\alpha > 1$ (independent of t) such that

$$-\psi_t(u_t, 0) \geq c\|u_t\|_V^{\alpha} - d\|u_t\|_V + d_0 \quad \forall u_t \in K_t, \ \forall t.$$

Now, the discrete approximate problem $(\mathcal{P}_t)$, $t \in T$, of the problem $(\mathcal{P})$ reads: Find $u_t \in K_t$ such that

$$\psi_t(u_t, v_t) \geq \langle g, v_t - u_t \rangle \quad \forall v_t \in K_t. \tag{4}$$

Further, we study the behaviour of u_t and present the following basic convergence result.

Theorem 1 (*General Approximation Result*). *Under conditions (H0)–(H3), the family $\{u_t\}$ of solutions to the problem $(\mathcal{P}_t)$ is uniformly bounded in V. Moreover, there exists a subnet of $\{u_t\}$ that converges weakly in V to a solution of the problem $(\mathcal{P})$. Furthermore, any weak accumulation point of $\{u_t\}$ is a solution to $(\mathcal{P})$.*

Proof. Setting $v_t = 0$ in (4) and using (H2) we obtain

$$c\|u_t\|_V^{\alpha} - d\|u_t\|_V + d_0 \leq -\psi_t(u_t, 0) \leq \|g\|_{V^*}\|u_t\|,$$

which proves the norm boundness of $\{u_t\}$. So, we can extract a subnet of $\{u_t\}$ denoted by $\{u_{t'}\}_{t' \in T'}$ such that $u_{t'}$ converges weakly to u in V. By $K_t \subset K$ and the closedness of K, $u \in K$.

Now, we take an arbitrary $v \in K$. By (H0), there exist a net $\{v_t\}$ such that $v_t \in K_t$ and $v_t \to v$ in V. Using (H2) and definition of (4), we get for any $v \in K$ that

$$\psi(u, v) \geq \limsup_{t' \in T'} \psi_{t'}(u_{t'}, v_{t'}) \geq \liminf_{t' \in T'} \psi_{t'}(u_{t'}, v_{t'}) \geq \lim_{t' \in T'} \langle g, v_{t'} - u_{t'} \rangle \geq \langle g, v - u \rangle$$

and consequently u is a solution to $(\mathcal{P})$. At the same time we have proved that any weak accumulation point of $\{u_t\}$ is a solution to the problem $(\mathcal{P})$. This should be

understood in the sense that every weak limit of any subnet of $\{u_t\}$ is a solution to the problem $(\mathcal{P})$. □

Remark 3. Without coercivity we get a stability result using Kuratowski set convergence [3] in the form that

$$\limsup_{t\in T} \text{ solution}\,(\mathcal{P}_t) \subset \text{ solution}\,(\mathcal{P}).$$

3 A Hemivariational Inequality in Linear Elasticity as a Pseudomonotone Variational Inequality

Let V be the classical Sobolev space $H^1(\Omega;\mathbb{R}^m)$, where $\Omega \subset \mathbb{R}^2$ is a domain with Lipschitz boundary $\partial\Omega$, and let $K \subseteq V$ be a nonempty closed, convex set specified later. Further, let the boundary $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_c \cup \bar{\Gamma}_F$ be composed of three mutually disjoint parts: a Dirichlet boundary Γ_D, a contact boundary Γ_c and a part Γ_F, where given external forces are applied. We also assume that the measure of Γ_D and Γ_c is strictly positive.

With γ we denote the trace operator from V into $L^2(\Gamma_c;\mathbb{R}^m)$. As known the trace operator is a linear continuous mapping. Therefore, there exists a constant c_0 depending on Ω, Γ_D and Γ_c such that

$$\|\gamma v\|_{L^2(\Gamma_c;\mathbb{R}^m)} \le c_0 \|v\|_V \quad \forall v \in V. \tag{5}$$

Moreover, the trace $V \hookrightarrow L^2(\Omega;\mathbb{R}^m)$ is compact [1, 2], and γ is compact too.

We adopt the standard notations from linear elasticity [17] and introduce the linear elastic operator $A : V \to V^*$ by

$$\langle Au, v\rangle = \int_\Omega \varepsilon(u) : \sigma(v)\, dx, \tag{6}$$

where $\varepsilon(u) = \frac{1}{2}(\nabla u + (\nabla u)^T)$ is the linearized strain tensor and $\sigma(v) = C : \varepsilon(v)$ is the stress tensor. Here, C is the elasticity tensor with symmetric positive L^∞ coefficients. Hence, the linear operator A is continuous, symmetric and due to the first Korn's inequality coercive.

The linear form $g : V \to \mathbb{R}$ is defined by

$$\langle g, v\rangle = \int_\Omega f_0 v\, dx + \int_{\Gamma_F} f_1 v\, ds$$

where $f_0 \in L^2(\Omega;\mathbb{R}^m)$ are the body forces and $f_1 \in L^2(\Gamma;\mathbb{R}^m)$ are the prescribed surface tractions on Γ_F.

In what follows we consider a function $f : \Gamma_c \times \mathbb{R}^m \to \mathbb{R}$ such that $f(\cdot,\xi) : \Gamma_c \to \mathbb{R}$ is measurable on Γ_c for all $\xi \in \mathbb{R}^m$ and $f(s,\cdot) : \mathbb{R}^m \to \mathbb{R}$ is locally Lipschitz on $\mathbb{R}^m$ for almost all $s \in \Gamma_c$.

The hemivariational inequality under consideration reads as follows:

Problem (P): Find $u \in K$ such that

$$\langle Au - g, v - u\rangle + \int_{\Gamma_c} f^0(s, \gamma u(s); \gamma v(s) - \gamma u(s))\, ds \geq 0 \quad \forall v \in K. \tag{7}$$

where $f^0(s, \cdot\,; \cdot)$ is the generalized Clarke directional derivative [11] of $f(s, \cdot)$.

In the study of *Problem (P)* we need the following growth condition on $\partial f(s, \cdot)$:

(H_f) there exist positive constants c and d such that for a.e. $s \in \Gamma_c$, all $\xi \in \mathbb{R}^m$ and for all $\eta \in \partial f(s, \xi)$ it holds

(i) $|\eta| \leq c(1 + |\xi|)$;

(ii) $\eta^T \xi \geq -d|\xi|$.

It follows from (i) and (ii) that for a.e. $s \in \Gamma_c$

$$\left|f^0(s, \xi; \varsigma)\right| = \left|\max_{\eta \in \partial f(s,\xi)} \eta^T \varsigma\right| \leq \max_{\eta \in \partial f(s,\xi)} |\eta|\,|\varsigma| \leq c(1 + |\xi|)|\varsigma| \quad \forall \xi, \varsigma \in \mathbb{R}^m \tag{8}$$

and

$$f^0(s, \xi; -\xi) = \max_{\eta \in \partial f(s,\xi)} \eta^T(-\xi) \leq d|\xi| \quad \forall \xi \in \mathbb{R}^m. \tag{9}$$

By virtue of (8) the integral in (7) is well defined.

Next, we define the functional $\varphi : V \times V \to \mathbb{R}$ by

$$\varphi(u, v) = \int_{\Gamma_c} f^0(s, \gamma u(s); \gamma v(s) - \gamma u(s))\, ds \quad \forall u, v \in V. \tag{10}$$

The main properties of φ are given in the following lemma.

Lemma 1. *The functional φ is pseudomonotone and satisfies*

$$\varphi(u, 0) \leq C\|u\|_V \quad \forall u \in V \tag{11}$$

for some positive constant C.

Proof. Let $\{u_n\}$ be a sequence in V such that

$$u_n \rightharpoonup u \text{ in } V \text{ as } n \to \infty.$$

Since γ is compact, it follows that

$$\gamma u_n \to \gamma u \text{ in } L^2(\Gamma_c; \mathbb{R}^m) \text{ as } n \to \infty. \tag{12}$$

Now, we fix $v \in V$ and show that

$$\limsup_{n \to \infty} \varphi(u_n, v) \leq \varphi(u, v). \tag{13}$$

We first observe that by (12) there exists a subsequence of $\{\gamma u_n\}$, which we denote again by $\{\gamma u_n\}$, such that

$$\gamma u_n(s) \to \gamma u(s) \quad \text{for a.e. } s \in \Gamma_c \tag{14}$$

and

$$|\gamma u_n(s)| \le \kappa_0(s) \quad \text{for some function } \kappa_0 \in L^2(\Gamma_c;\mathbb{R}_+). \tag{15}$$

Using (8) and (15), it follows that

$$\begin{aligned} f^0(s,\gamma u_n(s);\gamma v(s)-\gamma u_n(s)) &\le c(1+|\gamma u_n(s)|)|\gamma v(s)-\gamma u_n(s)| \\ &\le c(1+\kappa_0(s))\big(|\gamma v(s)|+\kappa_0(s)\big) \in L^1(\Gamma_c). \end{aligned}$$

From (14) and the upper semicontinuity of $f^0(s;\cdot,\cdot)$, we conclude by applying the Fatou lemma that

$$\begin{aligned} \limsup_{n\to\infty} \varphi(u_n,v) &= \limsup_{n\to\infty} \int_{\Gamma_c} f^0(s,\gamma u_n(s);\gamma v(s)-\gamma u_n(s))\,ds \\ &\le \int_{\Gamma_c} \limsup_{n\to\infty} f^0(s,\gamma u_n(s);\gamma v(s)-\gamma u_n(s))\,ds \\ &\le \int_{\Gamma_c} f^0(s,\gamma u(s);\gamma v(s)-\gamma u(s))\,ds = \varphi(u,v) \end{aligned} \tag{16}$$

and thus, (13) is shown. Hence, the functional φ is pseudomonotone.

Furthermore, by (9) for any $u \in V$ we can estimate

$$\begin{aligned} \varphi(u,0) &= \int_{\Gamma_c} f^0(s,\gamma u(s);-\gamma u(s))\,ds \le d\int_{\Gamma_c} |\gamma u(s)|\,ds \\ &\le d((\mathrm{meas}\,(\Gamma_c))^{1/2}\|\gamma u\|_{L^2(\Gamma_c;\mathbb{R}^m)} \overset{(5)}{\le} d((\mathrm{meas}\,(\Gamma_c))^{1/2}c_0\|u\|_V, \end{aligned}$$

which implies (11). The proof of the lemma is thus complete. □

Remark 4. Arguments similar to those used to derive (13) show that φ is weakly upper semicontinuous with respect to the both arguments.

Define now

$$\psi(u,v) = \langle Au, v-u\rangle + \varphi(u,v).$$

Since the linear continuous operator $A : V \to V^*$ gives rise to a pseudomonotone function $\langle Au, v-u\rangle$, the functional ψ is pseudomonotone as a sum of two pseudomonotone functions, see [7, 15], and the existence of a solution u to problem $(\mathcal{P})$ is due to the existence theorem in [14]. Moreover, taking into account Remark 4, we conclude that ψ satisfies the hypothesis (H2). Finally, it follows from Lemma 1 and the coercivity of the operator A that (H3) with $\alpha = 2$ holds too.

In what follows, in view of our applications, we consider the maximum function

$$f : \Gamma_c \times \mathbb{R}^m \to \mathbb{R}, \quad f(s,\xi) = \max\{g_1(s,\xi), g_2(s,\xi), \ldots, g_p(s,\xi)\},$$

where $g_i(s,\cdot):\mathbb{R}^m\to\mathbb{R}$ are continuously differentiable. We introduce the following growth conditions on the gradient $\nabla_\xi g_i(s,\cdot)$:

(H_{g_i}) there exist positive constants c_i, d_i such that for a.e. $s\in\Gamma_c$ and all $\xi\in\mathbb{R}^m$ it holds

(i) $|\nabla_\xi g_i(s,\xi)|\le c_i(1+|\xi|)$;

(ii) $\nabla_\xi g_i(s,\xi)^T\xi\ge -d_i|\xi|$.

Note that (H_g)(ii) holds for any continuously differentiable function g satisfying

$$\begin{array}{l} g'(x)\le d^0 \quad \text{for } x<0 \\ g'(x)\ge -d^0 \text{ for } x\ge 0 \end{array}, \quad \text{for some} \quad d^0=\text{const}\ge 0.$$

Let now $g:\mathbb{R}\to\mathbb{R}$ be a continuously differentiable function with ultimately increasing derivative g'; that is

$$\sup_{x\in(-\infty,-\xi)} g'(x)\le c^0\le \inf_{x\in(\xi,+\infty)} g'(x) \quad \text{for some} \quad \xi\ge 0 \ \text{ and } c^0\in\mathbb{R}.$$

Such a function fulfills the directional growth condition (H_g)(ii) as well. In this case, the constant d^0 is defined by

$$d^0=|c^0|+\sup_{x\in[-\xi,\xi]}|g'(x)|.$$

Moreover, the condition (H_{g_i}) implies that the maximum function f belongs to the class of functions for which (H_f) is satisfied.

4 A Delamination 2D Benchmark Problem

In this section to illustrate our theoretical results we consider a delamination problem for a laminated composite structure and present some numerical results. As a model example we consider a symmetric two-dimensional laminated structure depicted in Fig. 1 with the modulus of elasticity $E=210\,\text{GPa}$ and Poisson's ratio $\nu=0.3$ (steel). Because of the symmetry of the structure we consider only the upper half ($100\,\text{mm}\times 10\,\text{mm}$). The body is fixed on Γ_u, i.e.

$$u_i=0 \text{ on } \Gamma_u, \quad i=1,2.$$

the part Γ_{F^2} is load-free. On Γ_{F^1} the boundary forces F is prescribed

$$\mathbf{F}=(0,P) \quad \text{on } \Gamma_{F^1}.$$

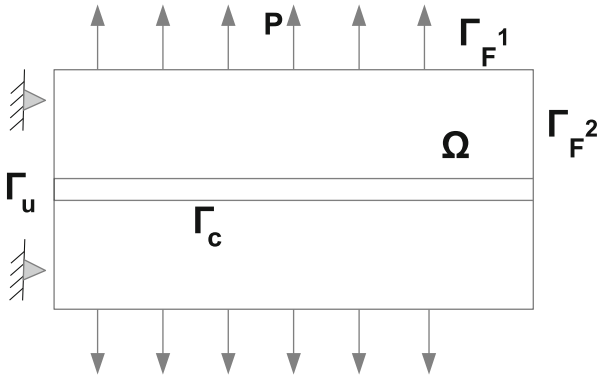


Fig. 1 A 2D benchmark with force distribution and boundary decomposition

The linear form $\langle \mathbf{g}, \cdot \rangle$ is defined by

$$\langle \mathbf{g}, \mathbf{v} \rangle = P \int_{\Gamma_{F1}} v_2 \, ds.$$

Further,

$$u_2 \geq 0 \quad \text{a.e. on } \Gamma_c$$

and

$$-S_N(s) \in \partial j(s, u_N(s)) \quad \text{for a.a. } s \in \Gamma_c.$$

Note that S_N denotes the normal component of the boundary stress vector. A typical nonmonotone law $\partial j(s, \cdot)$ describing delamination is shown in Fig. 2. This law is derived from a nonconvex and a nonsmooth superpotential j expressed in terms of a minimum function. In particular, $j(s, \cdot)$ is a minimum of four convex quadratic and one linear function.

We assume also that no tangential traction is given, i.e. $S_T(s) = 0$. The weak formulation of the delamination problem is given now by the following hemivariational inequality: Find $\mathbf{u} \in K$ such that

$$\langle A\mathbf{u}, \mathbf{v} - \mathbf{u} \rangle + \int_{\Gamma_c} j^0(s, u_N(s); v_N(s) - u_N(s)) \, ds \geq \langle \mathbf{g}, \mathbf{v} - \mathbf{u} \rangle \tag{17}$$

for all $\mathbf{v} \in K$, where $A : V \to V^*$ is the linear elastic operator defined by (6),

$$V = \{\mathbf{v} \in H^1(\Omega; \mathbb{R}^2) \, : \, \mathbf{v} = 0 \text{ on } \Gamma_u\},$$

and

$$K = \{\mathbf{v} \in V \, : \, v_2 \geq 0 \text{ on } \Gamma_c\}.$$

We solve (17) numerically by first using a regularization of the nonsmooth functional and then by applying a finite element scheme for the regularized problem.

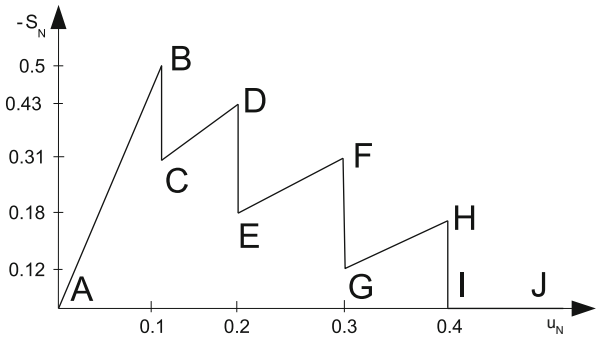


Fig. 2 A nonmonotone delamination law

This approach allows us to replace (17) by a smooth optimization problem that can be solved by using global minimization algorithms like trust region methods.

The regularized problem of (17) is given now by: find $\mathbf{u}_\varepsilon \in K$ such that

$$\langle A\mathbf{u}_\varepsilon, \mathbf{v}-\mathbf{u}_\varepsilon\rangle + \langle -DJ_\varepsilon(\mathbf{u}_\varepsilon), \mathbf{v}-\mathbf{u}_\varepsilon\rangle \geq \langle g, \mathbf{v}-\mathbf{u}_\varepsilon\rangle \quad \forall \mathbf{v} \in V, \tag{18}$$

where $DJ_\varepsilon : V \to V^*$ is the Gâteaux derivative of

$$J_\varepsilon(\mathbf{v}) = \int_{\Gamma_c} S(s, v_N(s), \varepsilon)\, ds$$

defined by

$$\langle DJ_\varepsilon(\mathbf{u}), \mathbf{v}\rangle = \int_{\Gamma_c} S'_\xi(s, u_N(s), \varepsilon) v_N(s)\, ds.$$

Here $S : \Gamma_c \times \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$ is a smoothing approximation of the maximum function $-j$ based on the Bertsekas representation formula for the maximum function using the plus function [6] and the smoothing approximation $P : \mathbb{R}_{++} \times \mathbb{R} \to \mathbb{R}$ of the plus function due to Zang [24] defined by

$$P(\varepsilon, t) = \begin{cases} 0 & \text{if } t < -\frac{\varepsilon}{2} \\ \frac{1}{2\varepsilon}(t+\frac{\varepsilon}{2})^2 & \text{if } -\frac{\varepsilon}{2} \leq t \leq \frac{\varepsilon}{2} \\ t & \text{if } t > \frac{\varepsilon}{2}. \end{cases}$$

For more details concerning this regularization technique, we refer the reader to [9, 19, 20].

Next, we briefly describe the discretization of (18). Let $\{\mathcal{T}_h\}$ be a regular triangulation of Ω. By V_h we denote the space of all continuous, piecewise linear functions on $\mathcal{T}_h$ vanishing on Γ_u. Thus $V_h \subset V$. Further, K is discretized by

$$K_h = \{v_h \in V_h : v_{h2}(P_i) \geq 0 \quad \forall P_i \in \overline{\Gamma}_c \backslash \overline{\Gamma}_u\}.$$

With piecewise linear approximation, $K_h \subset K$. Moreover, according to [10], we have $K \subset \liminf_h K_h$.
The approximation of (17) now reads as follows:
Find $u_h \in K_h$ such that for all $v_h \in K_h$

$$\langle Au_h, v_h - u_h\rangle + \langle -DJ_{\varepsilon,h}(u_h), v_h - u_h\rangle \geq P \int_{\Gamma_{F1}} (v_{h2} - u_{h2})\, dx_1, \tag{19}$$

where we use the trapezoidal quadrature rule

$$\langle DJ_{\varepsilon,h}(u_h), v_h\rangle = \frac{1}{2} \sum_{\overline{P_iP_{i+1}} \in \overline{\Gamma}_c \setminus \overline{\Gamma}_u} |P_iP_{i+1}| \left[\frac{\partial S}{\partial \xi}(P_i, u_{h2}(P_i), \varepsilon) v_{h2}(P_i) + \frac{\partial S}{\partial \xi}(P_{i+1}, u_{h2}(P_{i+1}), \varepsilon) v_{h2}(P_{i+1}) \right].$$

Altogether, we have $t = (\varepsilon, h)$ in our approximation scheme of Sect. 2. Next, we use the condensation technique based on the Schur complement to reduce the total number of unknowns in (19). As a result we receive a reduced finite-dimensional variational inequality problem formulated only in terms of the displacements at the free nodes on the boundary Γ_c. The latter problem is rewritten into a mixed complementarity problem which by means of the Fischer–Burmeister Function [12] $\phi(a,b) = \sqrt{a^2+b^2} - (a+b)$ is further reformulated as a system of nonlinear equations. Finally, by using an appropriate merit function we receive an equivalent smooth, unconstrained minimization problem which is numerically solved by applying an algorithm based on trust region methods. For more details we refer the reader to [19].

The numerical results are shown in Figs. 3 and 4. They illustrate the behaviour of the normal displacements and the distribution of the normal stresses along the contact boundary Γ_c for the grid 40×4 and for different values of P. From Fig. 4 it is easy to see that for a load $0.4\,\mathrm{N/mm^2}$ no delamination occurs. A partial one takes place for a load $0.6\,\mathrm{N/mm^2}$. In this case some of computed normal displacements are larger than 0.2 mm and the first jump occurs (see Fig. 4). Finally, with a load $1.0\,\mathrm{N/mm^2}$ we have a complete damage of the adhesive material. In this case some of the computed normal displacements are larger than 0.4 mm and the computed normal stresses jump down to zero as described by the nonmonotone delamination law presented in Fig. 2.

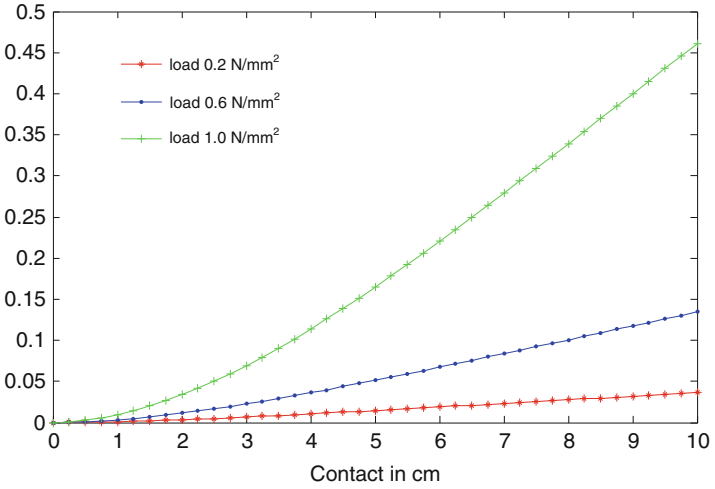


Fig. 3 The vertical displacements on Γ_c for the grid 40×4

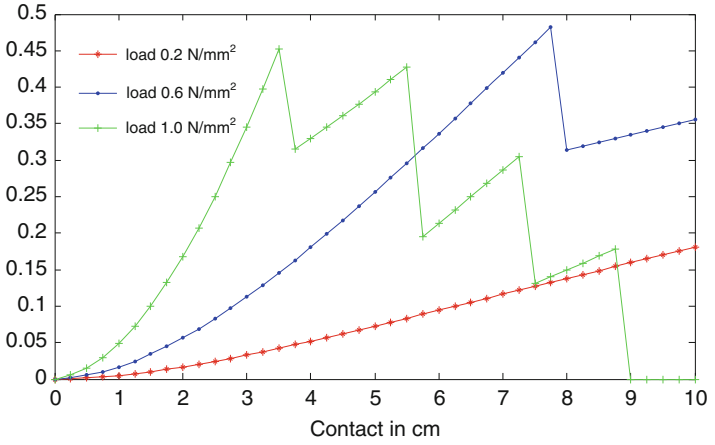


Fig. 4 The normal stresses on Γ_c for the grid 40×4

References

1. Adams, Robert A., Fournier, John J. F.: Sobolev spaces, Elsevier Ltd. (2003)
2. Alt, H.W.: Lineare Funktional-Analysis. Springer, Berlin (1999)
3. Aubin, J.-P., Frankowska, H.: Set-Valued Analysis. Birkhäuser, Boston (2008)
4. Baniotopoulos, C.C., Haslinger, J., Morávková, Z.: Mathematical modeling of delamination and nonmonotone friction problems by hemivariational inequalities. Appl. Math. **50**(1), 1–25 (2005)
5. Baniotopoulos, C.C., Haslinger, J., Morávková, Z.: Contact problems with nonmonotone friction: discretization and numerical realization. Comput. Mech. **40**, 157–165 (2007)

6. Bertsekas, D.P.: Nondifferentiable optimization via approximation. Math. Program. Stud. vol. 3, pp. 1–25 (1975)
7. Brézis, H.: Equations et inequations non linéaires dans les espaces vectoriels en dualité. Ann. Inst. Fourier **18**, 115–175 (1968)
8. Carl, S., Le, V.K., Motreanu, D.: Nonsmooth Variational Problems and Their Inequalities. Springer, New York (2007)
9. Chen, X., Qi, L., Sun, D.: Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities. Math. Comput. **67**, 519–540 (1998)
10. Ciarlet, P.G., Lions, J.L.: Handbook of Numerical Analysis, vol. 2, Elsevier, Amsterdam (1991)
11. Clarke, F.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
12. Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, vols. 1 and 2. Springer, New York (2003)
13. Goeleven, D., Motreanu, D., Dumont, Y., Rochdi, M.: Variational and Hemivariational Inequalities: Theory, Methods and Applications, Vol. I: Unilateral Analysis and Unilateral Mechanics, Vol. II: Unilateral problems. Kluwer, Boston (2003)
14. Gwinner, J.: Nichtlineare Variationsungleichungen mit Anwendungen. Ph.D. thesis, Universität Mannheim (1978)
15. Gwinner, J.: A note on pseudomonotone functions, regularization, and relaxed coerciveness. J. Nonlinear Anal. Theory Methods Appl. **30**(7), 4217–4227 (1997)
16. Haslinger, J., Miettinen, M., Panagiotopoulos, P.D.: Finite Element Methods for Hemivariational Inequalities. Kluwer Academic, Boston (1999)
17. Kikuchi, N., Oden, J.T.: Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods. SIAM, Philadelphia (1988)
18. Naniewicz, Z., Panagiotopoulos, P. D.: Mathematical Theory of Hemivariational Inequalities and Applications. Dekker, New York (1995)
19. Ovcharova, N.: Regularization Methods and Finite Element Approximation of Hemivariational Inequalities with Applications to Nonmonotone Contact Problems. Ph.D. thesis, Universität der Bundeswehr München, Cuvillier Verlag, Göttingen (2012)
20. Ovcharova, N., Gwinner, J.: On the regularization method in nondifferentiable optimization applied to hemivariational inequalities. In: Constructive Nonsmooth Analysis and Related Topics. Springer Optimization and Its Application, vol. 87, pp. 59–70. Springer, New York (2014)
21. Panagiotopoulos, P.D.: Hemivariational Inequalities. Applications in Mechanics and Engineering. Springer, Berlin (1993)
22. Panagiotopoulos, P.D.: Inequality Problems in Mechanics and Application. Convex and Nonconvex Energy Functions. Birkhäuser, Basel (1998)
23. Sofonea, M., Matei, A.: Variational Inequalities with Applications. Springer, New York (2009)
24. Zang, I.: A smoothing-out technique for min-max optimization. Math. Program. **19**, 61–77 (1980)

Designing Groundwater Supply Systems Using the Mesh Adaptive Basin Hopping Algorithm

Elisa Pappalardo and Giovanni Stracquadanio

1 Introduction

Groundwater supply systems are an important field of energy engineering, characterized by several challenging computational problems, ranging from modeling water flows to finding optimal pump location. Particularly interesting is the task of minimizing the cost of providing a specific quantity of water, subject to constraints on the net extraction rate, pumping rates, and hydraulic head location. The problem is extremely complex from a mathematical point of view, since it takes into account parameters that are stochastic in nature, leading to objective functions that are often discontinuous, nonlinear, non-convex, and with a large number of local minima. Classical gradient-based methods produce low quality solutions in this scenario and tend to be computationally expensive.

It has been showed that *derivative-free optimization (DFO) algorithms* represent an effective approach to solve water supply problems [11]. These methods rely on a minimization scheme that takes into account only the objective function value, making them robust to noise, discontinuity and nonlinearity of the objective function: this is a classical scenario when the objective function value is the output of a simulator [7].

Recently, the *Mesh Adaptive Basin Hopping* (MABH) method has been applied to solve industrial problems [23]; the algorithm combines a heuristic search with a local optimization step, which relies on the *Mesh Adaptive Direct Search* (MADS, [2]) algorithm. In this work, we extend the MABH algorithm with a new heuristic for the groundwater supply problem, which is able to explore the landscape of mixed-integer problems.

We test our algorithm on four different problems, which take into account different number of extraction wells, and both confined and unconfined aquifers.

E. Pappalardo • G. Stracquadanio (✉)
Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
e-mail: pappalardo@jhu.edu; stracquadanio@jhu.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_21,

The performance of MABH is compared with the state-of-the-art deterministic and stochastic methods, such as *Implicit Filtering for Constrained Optimization* (IFFCO) [10], *Generalized Pattern Search* (GPS) [18], MADS [3], *Differential Evolution* (DE) [20, 21], *classical Genetic (GA) Algorithm* [10], and the *Covariance Adaptation Matrix Evolution Strategy* (CMA-ES) [13]. The experimental results show that MABH outperforms these methods in terms of quality of solutions and number of function evaluations; moreover, we show that the use of the monotonic basin hopping scheme improves the performance of GPS and MADS algorithms.

2 The Groundwater Supply Model

Aquifers are underground layers of water-bearing regions from which groundwater can be pumped out through wells; *confined* aquifers are bounded on the top and on the bottom by impermeable materials, whereas there are no confining layers between an *unconfined* aquifer and the surface.

In this work, we focus on a well-field design problem, where the objective is the minimization of the cost of providing a specific quantity of water, subject to a set of constraints, such as the net extraction and pumping rates, and the hydraulic head location. The hydrological settings considered are based on the model proposed by Fowler et al. [11], characterized by homogeneous confined and unconfined aquifers, in which wells can inject or extract water. System costs include the installation and maintenance of wells, along with the extraction costs. The latter takes into account the costs for lifting the water from the aquifer to the discharge point, and those for supplying sufficient discharge pressure to achieve the desired flow [11].

In this design problem, the decision variables are the number of the wells n, their $\{(x_i, y_i)\}_{i=1}^{n}$ locations, and the pumping rates $\{Q_i\}_{i=1}^{n}$ (m^3/s) [11].

The physical domain is defined as $\Omega = [0, 1000] \times [0, 1000] \times [0, 30]$ m, with ground elevation $z_{gs} = 60$ m for the confined aquifer, and $z_{gs} = 30$ m for the unconfined aquifer [11]; however, for consistency with Fowler et al. [11], wells are required to be at least 200 m from the Dirichlet boundaries defined by $0 \leq x_i, y_i \leq 800$ m.

The objective function consists of two components: the capital cost of installing a well, defined as:

$$f^c = \sum_{i=1}^{n} c_0 d_i^{b0} + \sum_{i, Q_i < 0} c_1 |Q_i^m|^{b1} (z_{gs} - h^{\min})^{b2} \tag{1}$$

which accounts for the drilling and installation costs; c and b represent the coefficient costs (Table 1).

The operational cost for a well is defined as

$$f^0 = \int_0^{t_f} \left[\sum_{i, Q_i < 0} c_2 Q_i (h_i - z_{gs}) + \sum_{i, Q_i > 0} c_3 Q_i \right] dt \tag{2}$$

Table 1 Parameters of the objective function

	Parameter	Value	Units
Cost coefficients	c_0	5.5×10^3	$\$/\mathrm{m}^{b0}$
	c_1	5.75×10^3	$\$/[(\mathrm{m}^3/\mathrm{s})^{b1} \cdot \mathrm{m}^{b2}]$
	c_2	2.90×10^{-4}	$\$/\mathrm{m}^4$
	c_3	1.45×10^{-4}	$\$/\mathrm{m}^3$
	b_0	0.3	–
	b_1	0.45	–
	b_2	0.64	–
Well depth	z_{gs}	60 confined	m
	z_{gs}	30 unconfined	m
	d_i	z_{gs}	m
Pumping rate	Q_i^m	$1.5Q_i$	m^3/s

where the term for the extraction wells includes a lifting cost to raise the water to the ground; t_f is the simulation time and is set to 5 years. A negative pumping rate Q_i means that a well is extracting groundwater, while a positive pumping rate Q_i indicates water injection; $d_i = z_{gs}$ is the depth of well i, Q_i^m is the designed pumping rate, which represents the maximum rate at which a given well can pump water; $h^{\min}$ is the minimum allowable head, and h_i is the hydraulic head in well i.

The objective function for the problem takes into account both the installation and operational costs and is defined as the minimization of the function:

$$F = f^c + f^0 \tag{3}$$

To evaluate the objective function, the computation of the hydraulic heads in wells h_i, for a set $\{Q_i\}_{i=1}^n$ of pumping rates at the given locations $\{(x_i, y_i)\}_{i=1}^n$, is required.

Obtaining the head values requires a call to a groundwater flow simulator, MODFLOW-96 [14] in our simulations. MODFLOW takes in input the locations and pumping rates of the wells and returns the h_i values used to evaluate the objective function and the constraints.

To ensure that wells are located appropriately in the physical domain and operate at reasonable levels, two constraints are enforced on the hydraulic head locations and pumping rates. Each hydraulic head must verify the following constraints:

$$h^{\min} \leq h_i \leq h^{\max}, \qquad i = 1, \ldots, n \tag{4}$$

where the upper bound allows to maintain the hydraulic head below the surface, while the lower bound limits its drawdown. Pumping rates are constrained as follows:

$$Q^{e\max} \leq Q_i \leq Q^{i\max}, \qquad i = 1, \ldots, n \tag{5}$$

where $Q^{e\max}$ and $Q^{i\max}$ represent the maximum extraction rate and the maximum injection rate, respectively.

The maximum and minimum values for head location and pumping rates are shown in Table 2. Finally, the total amount of water to supply is given by:

$$Q_T = \sum_{i=1}^{n} Q_i \leq Q_T^{\min}, \tag{6}$$

where $Q_T^{\min}$ is the minimum allowable total extraction rate.

Table 2 Bounds for the constraints of the objective function

	Parameter	Value	Units
Maximum extraction rate	$Q^{e\max}$	-6.4×10^{-3}	m^3/s
Maximum injection rate	$Q^{i\max}$	6.4×10^{-3}	m^3/s
Minimum total extraction rate	$Q_T^{\min}$	-3.2×10^{-2}	m^3/s
Minimum allowable head	$h^{\min}$	40 confined	m
		10 unconfined	m
Maximum allowable head	$h^{\max}$	60 confined	m
		30 confined	m

3 Derivative-Free Optimization Methods

The inherent complexity of the groundwater supply problem requires the introduction of ad hoc optimization methods; in particular, due to the extreme roughness of the search landscape, classical gradient-based algorithms have been proved to perform poorly [10]. For this reason, several derivative-free optimization methods have been proposed in literature; in general, DFOs aim to find a minimizer of an objective function by using only the objective function value. This approach does not require any derivate or gradient information and, in general, performs effectively on problems where the objective function is computed by a simulator.

Two subclasses of DFO methods can be identified: deterministic and stochastic. While deterministic algorithms provide a theoretical proof of convergence to a minimizer, stochastic methods strongly rely on randomized sampling procedures, without any convergence guarantee. According to the above classification, in the next paragraphs we introduce a brief review of the most effective methods for designing groundwater supply systems.

3.1 Deterministic Methods

Several deterministic methods have been proposed and compared in [10, 11]. The term "deterministic" does not imply that the algorithm is able to rigorously find a global minimum but, conversely, the method assures some convergence results, when the function has specific mathematical properties. In particular, the IFFCO [6] algorithm and *Pattern Search* (PS) methods have been successfully applied to the groundwater supply problem.

Implicit filtering is a projected quasi-Newton method that uses a sequence of finite difference steps to approximate the gradients. The difference increment is reduced as the optimization progresses, allowing to avoid local minima, discontinuities, or non-smooth regions [5, 10, 12]. IFFCO evaluates the objective function in all the points required for the poll step, in order to provide a gradient estimate. The Hessian estimate is computed by a quasi-Newton update; this provides a quadratic surrogate used to explore the search space [11].

Conversely, GPS [18] and MADS, [3] rely on sampling of a finite number of points, selected according to a set of directions; typically, search directions are chosen such that they form a positive spanning set or positive bases. Additionally, the sampled points are constrained to lie on a mesh, which fineness is coarsened based on the outcome of the current iteration. This mechanism, called *polling*, tends to favor large steps when a minimizing direction is found, otherwise restricts the sampling to a smaller basin. The strategy ensures the convergence of the method to a secondary stationary point using the *Clarke directional derivatives* [1]. IFFCO and GPS algorithms provide the current putative optimal solutions for the groundwater supply problem; they achieve such results using a tight budget of simulator calls, which makes them suitable in industrial environments.

3.2 Stochastic Methods

Stochastic algorithms represent an effective alternative when no boundary conditions are available to tackle the problem. These algorithms rely on two main components: a sampling procedure to generate candidate solutions, and a selection scheme to assure asymptotic convergence to a minimizer.

Evolutionary Algorithms (EA) are a class of methods inspired by the Darwinian process of natural selection [17]: they generate a population of candidate solutions, which are recombined and mutated to explore and exploit the search space.

Genetic Algorithms (GA) are among the most used evolutionary algorithms; they mimic the natural evolutionary process by evolving a population of solutions. GAs are based on natural selection and sexual reproduction processes; the first mechanism determines which members of a population survive and are able to reproduce, the second one assures genetic recombination among individuals of the same population. EAs do not use any mathematical information and do not provide any convergence property; from a theoretical point of view, there is no proof of EAs

convergence but in practice they have been shown to be effective even for complex problems. A single-objective variant of the *Non-dominated Sorting Genetic Algorithm* (NSGA-II) [8] has been proposed for the groundwater supply problem [10, 11]. This algorithm deals with continuous and discrete variables using ad hoc evolutionary operators; for real-coded variables, the simulated binary crossover (SBX) operator with polynomial mutation is used, while the single-point crossover with bitwise mutation is used for binary-coded variables.

In this work, we compare our method to two state-of-the-art evolutionary optimizers: the *Differential Evolution* (DE) [21] and *Covariance Matrix Adaptation Evolution Strategy* (CMA-ES) [13] algorithms. DE generates candidate solutions by simply adding a weighted difference of two individuals to a third; this scheme does not need any separate probability distribution, which makes it a robust self-organizing method. The classical selection scheme is based on a $(1,1)$-replacement, where the children substitute the parent according to the objective function value. Starting from its first application, the Chebyshev polynomial fitting problem [20, 21], DE has been successfully applied to a number of real world applications [22].

CMA-ES is an evolutionary algorithm that adopts a multivariate Gaussian distribution, such that the likelihood of successful steps is maximized. Currently, CMA-ES is the best evolutionary algorithm for derivative-free optimization and one of the most effective derivative-free optimizers in literature [4].

4 Mesh Adaptive Basin Hopping

The MABH [23] algorithm combines the MADS [2] algorithm with a local heuristic step, which is responsible for performing a sampling in the neighborhood of the current minimizer (*incumbent*). This approach can be viewed as a local descent method combined with a stochastic heuristic [19], where the Metropolis acceptance criterion is discarded in favor of a monotonic one.

Although no restriction is enforced on the landscape of the objective function, MABH is supposed to be more effective on funneled landscapes, as already shown for other basin hopping methods; recently, Stracquadanio et al. have shown that MABH is the most effective optimizer for the class of the antenna design problems [23], outperforming state-of-the-art methods in terms of efficiency and quality of solutions. Moreover, the use of a basin hopping strategy improves the robustness of MADS, since it overcomes the problem of poor initial iterates.

We extend the MABH strategy with heuristics to tackle the groundwater supply problem, aiming at improving the putative global optimal solutions with a tight budget of function evaluations; the algorithm is presented in Algorithm 1. MABH alternates a heuristic step (line 4) with a truncated local optimization step (line 5), in order to find a new incumbent. Local optimization can perform λ objective function evaluations at most; this strategy has the benefit of limiting the chance of being trapped into a local minimum. Finally, the current iterate is accepted if it improves

the incumbent objective function value. The stopping criterion can accommodate any strategy, such as the maximum number of function evaluations, maximum running time, or a predefined objective function value.

Algorithm 1 Mesh Adaptive Basin Hopping

```
1: procedure MABH(f, μ, ρ, λ)
2:     x* ← random_solution                          ▷ Best found
3:     while ¬ StopCondition do
4:         x̂ ← HeuristicSearch(x̂, μ, ρ)              ▷ Heuristic perturbation
5:         x̂ ← MADS(x̂, λ)                            ▷ Local optimization
6:         if f(x̂) < f(x*) then
7:             x* ← x̂
8:         end if
9:     end while
10: end procedure
```

In our implementation, MABH uses variable scaling; since the range of each variable could be extremely different, stagnation issues might happen. MABH overcomes this problem by projecting the variables on the $[-0.5, 0.5]^n \subseteq \mathbb{R}^n$ space, where n is the dimension of the problem.

The heuristic perturbation procedure performs a local search around the incumbent solution, by applying some stochastic noise; in particular, let x_i be the i-th variable of the problem, the following perturbation is applied:

$$x_i' = x_i + (-\rho + (\texttt{rand} \times 2 \times \rho))$$

where $\rho \in [0,1]$ is a parameter of the algorithm and `rand` is a function that generates a uniform random number in $[0,1)$. The number of perturbed variables is determined by the parameter $\mu \in [0,1]$, where 0 means no perturbation and 1 perturbs all; large μ values are not recommended since the perturbation operator tends to destroy suboptimal solutions.

5 Experimental Results

The experiments are focused on two instances of the groundwater supply problem, considering five and six wells. By using $n = 5$ wells extracting at $Q^{e\max} = -0.0064$ (m^3/s), the water supply demand is satisfied; since the installation cost is a fixed amount in this case ($\approx$ \$10,000 per well), it is interesting to find the best well position to minimize the operation cost (f^o).

For the six-well case design, both the installation and operational costs are considered; since the water supply demand can be satisfied by a five-well installation pumping at maximum rate, it is expected that an optimizer is able to find this solution even starting from a six-well design, since the installation cost of a well is greater than the operation cost.

5.1 Five-Well Design

The five-well design problem requires to find the spatial coordinates $\{x_i, y_i\}_{i=1}^{5}$ that minimize the operational costs. The position of the wells is subject to box constraints, such as $\{x_i, y_i\} \in \{[20{,}800] \times [20{,}800]\}, \forall i = 1, \ldots, 5$. The optimization has been conducted on both the *confined* and *unconfined* cases, in order to estimate how the performances of the algorithm change. An initial feasible solution has been determined for both the problems, and each optimization algorithm starts from this iterate to maintain an experimental coherence with the results presented in [10]. The initial solutions have an operation cost of \$23,204 and \$127,421 for the confined and unconfined problems, respectively; for the unconfined case, the installation cost of the system is considered for consistency with the most recent results available [11].

To have comparable results with the existing state-of-the-art methods, the number of objective function evaluations has been fixed to 500; MABH performs 10 iterations allowing at most 50 simulator calls to the local optimizer, and setting $\mu = 0.1$ and $\rho = 0.05$. Local optimization has been performed by using MADS with coordinate directions (GPS) and orthogonal directions (ORTHOMADS) [2]; default parameters have been used for both instances. CMA-ES and DE have been used with default parameters (Fig. 1).

By inspecting the results for the confined aquifer in Table 3, we assess that the two instances of MABH are able to find a new putative global optimum with at most 400 simulator calls (Fig. 2).

The deterministic methods are all ranked within the first four positions, remarking their ability of rapidly converging to a satisfactory solution; however, the results

Table 3 Five-well design results

	Confined five-wells		Unconfined five-wells	
Algorithm	Cost (\$)	Sim calls	Cost (\$)	Sim calls
MABH + MADS	**21,778**	400	**124,386**	300
MABH + GPS	**21,778**	300	**124,386**	250
GPS	**21,778**	340	**124,386**	321
IFFCO [11]	*21,830*	275	*125,129*	275
MADS	21,903	461	124,389	461
DE/RAND/1/EXP	21,933	500	124,573	500
DE/BEST/1/EXP	22,035	475	125,485	475
DE/RAND-TO-BEST/1/EXP	22,316	475	124,626	500
GA	22,822	330	**124,386**	930
CMA-ES	22,978	500	125,672	500

We report the cost of the best solution found and the number of simulator calls required. The new putative optimal solution is reported in boldface, in italic the previous putative global minimum

obtained by MABH show that random perturbations can prevent a premature stagnation and improve the convergence speed. Additionally, MABH variants reach the previous putative global optimum after ≈220 simulator calls versus the 275 of IFFCO, confirming the efficiency of the monotonic basin hopping strategy. DE performs sufficiently well, although the best solution is obtained by using all the budget of function evaluations; conversely, CMA-ES is not able to find a satisfactory solution, which might be apportioned to a difficult adaptation of the covariance matrix, mainly because of the high number of simulator failures.

Since a new best solution has been found, in Fig. 3 and in Table 4, we report the 2D visualization of the new optimal system and the coordinates of its wells, respectively. From a comparison with the previous best found solution, it is possible to note that some coordinates differ substantially, which means that despite the limited decrease of the objective function, the algorithm is able to locate a new basin.

For the unconfined five-well design problem, the putative global optimum is found by MABH with the smallest number of function evaluations. For this instance, deterministic methods are the most effective in terms of quality of solutions and computational efficiency, even if the GA offers comparable results at the price of an increased computational effort.

The analysis of the convergence plot (Fig. 1) shows that MABH is extremely efficient in reaching a nearly optimal solution and improves the performances of the local solver. It is also interesting to note that DE performs well on this instance, despite the limited budget of simulator calls; it clearly outperforms CMA-ES, which could be an evidence of the former algorithm to work better with noisy objective functions.

5.2 Six-Well Design

Starting from the consideration that the five-well solution is the best to supply the water demand, the effectiveness of the algorithm has been tested on a six-well system. This formulation considers both the installation and the operational costs; since the installation cost of each well is much higher than the operation cost, a satisfactory solution is the one that turns off a well [10].

A common approach to handle this situation is to reformulate the problem as a mixed-integer optimization problem [16], where an integer variable represents the well to remove. To eliminate the integer decision variable, we set an inactive well-threshold [9, 15], so that if $|Q_i| < 10^{-6}\,\mathrm{m}^3/\mathrm{s}$, the well i is removed from the design space, and not included in the cost calculation.

Two initial feasible solutions have been found for the six-well design, both for the confined and unconfined cases, having operation cost \$170,972 and \$152,878, respectively. The six-well design requires the optimization of the spatial coordinates $\{x_i, y_i\}_{i=1}^{6}$ of the wells in order to minimize the operational and installation cost; the well positions are subject to box constraints such that $\{x_i, y_i\} \in \{[20{,}800] \times [20{,}800]\}, \forall i = 1, \ldots, 6$.

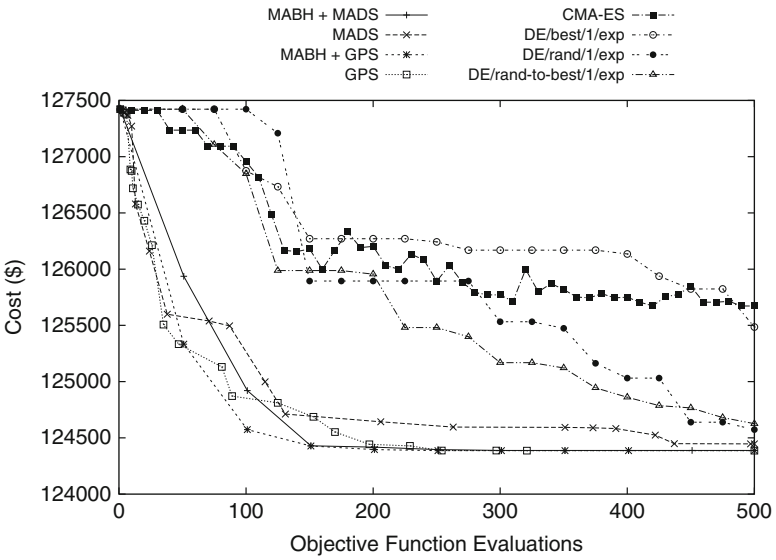


Fig. 1 Convergence plot for the five-well unconfined groundwater supply problem

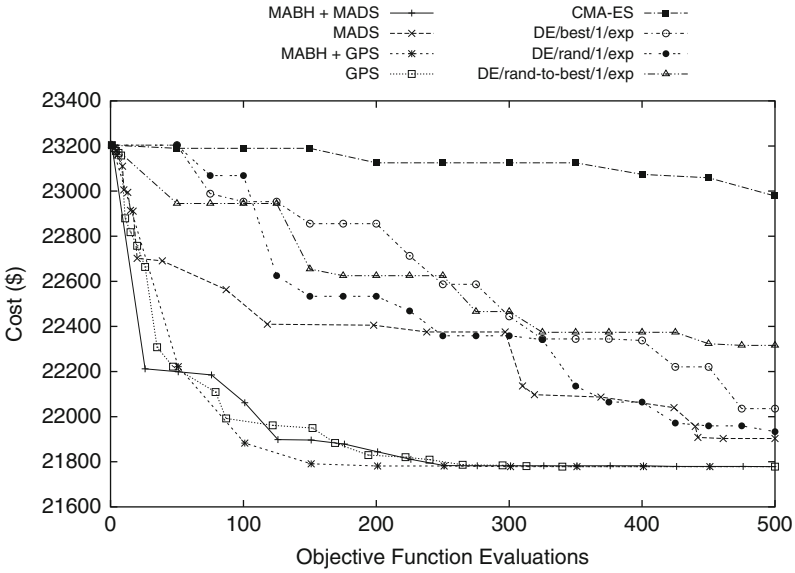


Fig. 2 Convergence plot for the five-well confined groundwater supply problem

For each well, the initial extraction rate $Q_i = -0.0064, \forall i = 1, \ldots, 6$ is considered; if Q_i falls below the installation threshold, the corresponding well is removed from the designed system.

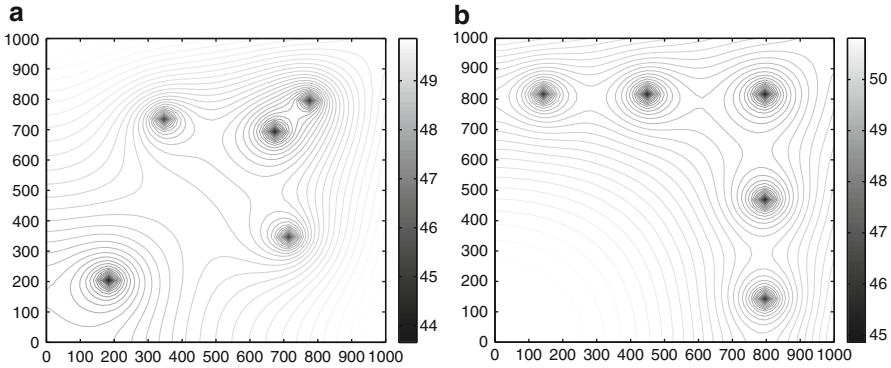


Fig. 3 Five-well confined groundwater supply problem. (**a**) Wells location as defined by the initial solution provided to the optimizers. (**b**) Wells location for the new putative global optimum found by MABH

Table 4 Five-well confined aquifer wells location

	Initial	IFFCO	MABH
X_1	350	401.7	151.6
Y_1	725	800.0	795.1
X_2	775	800.0	793.0
Y_2	775	800.0	794.7
X_3	675	776.9	452.1
Y_3	675	481.1	795.1
X_4	200	138.2	792.7
Y_4	200	800.0	460.1
X_5	725	798.4	795.1
Y_5	350	168.9	144.3

We report coordinates of the initial solution, the previous best known design found by IFFCO and the new optimal solutions found by MABH

In order to let the MABH algorithm explore the six-well design, the perturbation operator is designed such that it randomly chooses whether perturbing a location or turning off a well; this choice provides a natural way of handling the mixed-integer nature of the problem.

The stopping criterion for all the adopted algorithms is the attainment of 500 simulator calls; MABH uses the same settings adopted for the five-well case.

The results in Table 5 show that MABH finds new putative global minima for the confined and unconfined case; these solutions are found within the prefixed budget of simulator calls and at the beginning of the convergence.

Table 5 Six-well design results

Algorithm	Confined six-wells		Unconfined six-wells	
	Cost ($)	Simulator calls	Cost ($)	Simulator calls
MABH + MADS	**139,878**	150	**124,414**	250
MABH + GPS	140,147	250	**124,414**	301
IFFCO [11]	*140,237*	346	*124,582*	327
GA	140,628	464	127,069	161
MADS	143,639	500	160,981	500
GPS	161,143	500	161,143	500

We report the cost of the best solution found and the number of simulator calls required. The new putative optimal solution is reported in boldface, in italic the previous best found solution

It is interesting to observe that GPS and ORTHOMADS are not able to find satisfactory solutions, confirming that using the basin hopping scheme improves their performances.

The plots in Figs. 4 and 5 show that the solutions attained for the six-well design are very similar to the five-well solutions, which means that MABH is able to provide solutions close to the known optimal. Moreover, by comparing the solutions with the previous best found by IFFCO (Tables 6 and 7), it is possible to note that at least two wells are located in very different locations, confirming that MABH explores different basins of the search space.

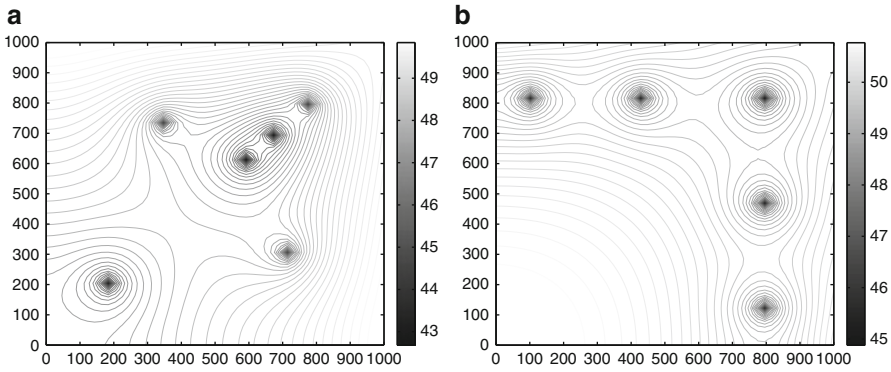


Fig. 4 Six-well confined groundwater supply problem. (**a**) Wells location as defined by the initial solution provided to the optimizers. (**b**) Wells location for the new putative global optimum found by MABH

Table 6 Six-well confined aquifer wells location

	Initial	IFFCO	MABH
X_1	350	350	116
Y_1	725	798.1	800
X_2	775	799.4	800
Y_2	775	775.0	463
X_3	675	626.2	–
Y_3	675	772.5	–
X_4	200	–	800
Y_4	200	–	122
X_5	725	737.2	800
Y_5	350	287.8	800
X_6	725	112.5	444
Y_6	350	795.0	800

We report coordinates of the initial solution, the previous best known design found by IFFCO and the new optimal solutions found by MABH

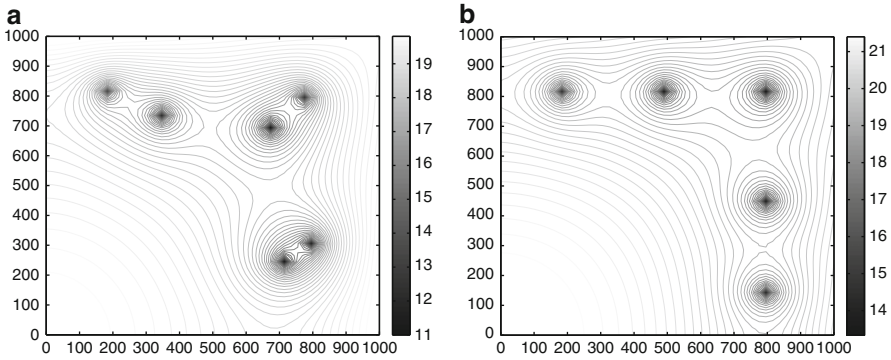


Fig. 5 Six-well unconfined groundwater supply problem. (**a**) Wells location as defined by the initial solution provided to the optimizers. (**b**) Wells location for the new putative global optimum found by MABH

6 Conclusions

Designing efficient systems for water supply is a challenging optimization task; the output of the simulator is extremely noisy and sensitive to the well locations and pumping rates, hence, finding a cost-effective design is not trivial.

Due to this complex scenario, derivative-free algorithms have been applied with satisfactory results. We propose a new Monotonic Basin Hopping algorithm (MABH), which combines a deterministic local optimization step with a heuristic

Table 7 Six-well unconfined aquifer wells location

	Initial	IFFCO	MABH
X_1	350	410.9	504.7
Y_1	725	798	800
X_2	775	715	800
Y_2	775	799	800
X_3	675	772.5	800
Y_3	675	473.9	441
X_4	200	187.8	195
Y_4	200	800	800
X_5	725	–	–
Y_5	350	–	–
X_6	725	800.0	800
Y_6	350	202.5	144

We report coordinates of the initial solution, the previous best known design found by IFFCO and the new optimal solutions found by MABH

perturbation method. This approach aims at preventing stagnation and improves the exploring ability of the method. The experimental results on the design of five- and six-well systems show that our approach is suitable to tackle this class of problems; on three out of four design problems, we were able to find new optimal solutions that minimize the costs of the system and identify new designs with substantially different wells locations.

Finally, we proved that using the basin hopping strategy improves the efficiency and effectiveness of the pattern search methods; these results confirm that hybrid optimization methods represent a viable approach for computationally expensive black-box optimization problems.

References

1. Abramson, M.A., Audet, C.: Convergence of mesh adaptive direct search to second-order stationary points. SIAM J. Optim. **17**(2), 606–619 (2006)
2. Abramson, M.A., Audet, C., Dennis, J.E., Jr., Le Digabel, S.: Orthomads: a deterministic mads instance with orthogonal directions. SIAM J. Optim. **20**(2), 948–966 (2009)
3. Audet, C., Dennis, J.E., Jr.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(1), 188–217 (2006)
4. Auger, A., Hansen, N.: A restart cma evolution strategy with increasing population size. In: Proceedings of the IEEE Congress on Evolutionary Computation, 2005, vol. 2, pp. 1769–1776. IEEE, Piscataway (2005)
5. Bertsekas, D.: On the Goldstein-Levitin-Polyak gradient projection method. IEEE Trans. Autom. Control **21**(2), 174–184 (2002)

6. Choi, T.D., Eslinger, O.J., Gilmore, P., Patrick, A., Kelley, C.T., Gablonsky, J.M.: IFFCO: implicit filtering for constrained optimization, version 2. Technical Report, Center for Research in Scientific Computation, North Carolina State University, Raleigh (1999)
7. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. Society for Industrial Mathematics, Philadelphia (2009)
8. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
9. Fowler, K.R., Kelley, C.T., Kees, C.E., Miller, C.T.: A hydraulic capture application for optimal remediation design. Dev. Water Sci. **55**, 1149–1157 (2004)
10. Fowler, K.R., Kelley, C.T., Miller, C.T., Kees, C.E., Darwin, R.W., Reese, J.P., Farthing, M.W., Reed, M.S.C.: Solution of a well-field design problem with implicit filtering. Optim. Eng. **5**(2), 207–234 (2004)
11. Fowler, K.R., Reese, J.P., Kees, C.E., Dennis, J.E., Jr., Kelley, C.T., Miller, C.T., Audet, C., Booker, A.J., Couture, G., Darwin, R.W.: Comparison of derivative-free optimization methods for groundwater supply and hydraulic capture community problems. Adv. Water Resour. **31**(5), 743–757 (2008)
12. Gilmore, P., Kelley, C.T.: An implicit filtering algorithm for optimization of functions with many local minima. SIAM J. Optim. **5**, 269 (1995)
13. Hansen, N., Müller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). Evol. Comput. **11**(1), 1–18 (2003)
14. Harbaugh, A.W., McDonald, M.G.: User's documentation for MODFLOW-96, an update to the US Geological Survey modular finite-difference ground-water flow model. US Department of the Interior, US Geological Survey (1996)
15. Hemker, T., Fowler, K.R., von Stryk, O.: Derivative-free optimization methods for handling fixed costs in optimal groundwater remediation design. In: Proceedings of the CMWR XVI-Computational Methods in Water Resources, pp. 19–22. Citeseer (2006)
16. Hemker, T., Fowler, K.R., Farthing, M.W., von Stryk, O.: A mixed-integer simulation-based optimization approach with surrogate functions in water resources management. Optim. Eng. **9**(4), 341–360 (2008)
17. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press, Cambridge (1992)
18. Lewis, R.M., Torczon, V.: Pattern search algorithms for linearly constrained minimization. SIAM J. Optim. **10**(3), 917–941 (2000)
19. Pardalos, P.M., Schoen, F.: Recent advances and trends in global optimization: deterministic and stochastic methods. In: Proceedings of the Sixth International Conference on Foundations of Computer-Aided Process Design (2004)
20. Price, K.V.: Differential evolution. In: Handbook of Optimization, pp. 187–214. Springer, New York (2013)
21. Storn, R., Price., K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. **11**(4), 341–359 (1997)
22. Stracquadanio, G., La Ferla, A., De Felice, M., Nicosia, G.: Design of robust space trajectories. In: Research and Development in Intelligent Systems XXVIII, pp. 341–354. Springer, New York (2011)
23. Stracquadanio, G., Pappalardo, E., Pardalos, P.M.: A mesh adaptive basin hopping method for the design of circular antenna arrays. J. Optim. Theory Appl. **155**(3), 1008–1024 (2012)

Regularity of a Kind of Marginal Functions in Hilbert Spaces

Fátima F. Pereira and Vladimir V. Goncharov

1 Introduction

Given a real-valued function $f : X \times Y \to \mathbb{R}$ and a multivalued mapping $C : X \rightrightarrows Y$ (X and Y are Banach spaces) the general *marginal function* is defined as

$$T(x) := \inf_{y \in C(x)} f(x,y) \quad , \tag{1}$$

where inf can be certainly substituted by sup. The *marginal mapping* instead associates with each $x \in X$ the set of minimizers (or maximizers):

$$\Pi(x) := \{y \in C(x) : T(x) = f(x,y)\} \quad . \tag{2}$$

Regularity properties of marginal functions and mappings are important due to numerous applications in control theory, theory of games, mathematical economics, stochastic analysis, etc. We refer to [1, 2] for their general topological properties regarding the continuity and the lipschitzeanity.

A lot of works (see, e.g., [12, 24, 28, 29, 31, 32, 38] and the bibliography therein) was devoted to the representation of various kinds of subdifferentials of the marginal function through the respective subdifferentials of the function $f(\cdot,\cdot)$. The authors studied also subdifferential regularity of (1) in the sense of coincidence of differ-

F.F. Pereira
CIMA-UE, Rua Romão Ramalho 59, 7000-671, Évora, Portugal
e-mail: fmfp@uevora.pt

V.V. Goncharov (✉)
CIMA-UE, Rua Romão Ramalho 59, 7000-671, Évora, Portugal

Institute of Systems Dynamics and Control Theory of Siberian Branch of RAS,
ul. Lermontov 134, 664033 Irkutsk, Russia
e-mail: goncha@uevora.pt

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_22,

ent subdifferentials, and other properties such as the *approximate convexity* [33] or *generic differentiability* [22, 24, 41].

As about the *differentiability* of $T(\cdot)$ at a given point, notice that it can be treated in different ways. Namely, the *Fréchet differentiability* means the possibility to reduce the (*Fréchet*) *subdifferential* to the (eventually continuous) singleton $\nabla T(x)$ called the *Fréchet derivative*, or *gradient*. On the other hand, one may take an interest in both existence and uniqueness of the *proximal subgradient* that is a stronger property.

The Fréchet differentiability of the marginal function was particularly well studied when $X = Y = H$ is a Hilbert space with the norm $\|\cdot\|$; $C(x) = C \subset H$ is a closed set and $f(x,y) = \|x-y\|$. In this case $T(\cdot)$ is nothing else than the distance of the point to C, denoted further by $\mathrm{d}_C(\cdot)$, whereas $\Pi(x) = \pi_C(x)$ is the (multivalued) *metric projection* of x onto C. In general, the set $\pi_C(x)$ can be certainly empty that does not occur when C is convex. Moreover, in the convex case $\pi_C(x)$ is a singleton, which is Lipschitz continuous w.r.t. x (with the Lipschitz constant 1) on the whole space H, and the gradient $\nabla \mathrm{d}_C(x)$ is continuous (and locally lipschitzean) out of C.

If, instead, C is no longer convex, then the latter property, in general, fails. However, there is a class of so named φ-*convex* sets (called also *prox-regular*, *proximally smooth*, sets *with positive reach*, etc.), for which the projection $\pi_C(x)$ is well defined and continuous (in fact, locally lipschitzean) on some (open) neighbourhood $\mathcal{U}$ of C (equivalently, the distance $\mathrm{d}_C(\cdot)$ is of class $\mathcal{C}^{1,1}_{\mathrm{loc}}$ on $\mathcal{U}\backslash C$). For the first time such sets were considered in the pioneer work [23] by Federer in finite dimensions, while afterwards various characterizations of them were given in Hilbert and Banach setting (see, e.g., [4, 5, 8, 13, 15, 34, 37] and the bibliography therein). This class of sets is well studied up to now, and we refer to the nice survey [17] for their basic properties.

The next step is to minimize the function $f(x,y) = \rho_F(x-y)$ in the place of the norm, where $F \subset H$ is a closed convex bounded set such that the origin belongs to the interior $\operatorname{int} F$, and $\rho_F(\cdot)$ is the *Minkowski functional* (*gauge function*),

$$\rho_F(\xi) = \inf\{\lambda > 0 : \xi \in \lambda F\} \ . \tag{3}$$

In the latter case the value function (denoted by $\mathfrak{T}^F_C(x)$) of the respective minimization problem can be seen as the minimal time, which is necessary to achieve the *target set* $C \subset H$ from a point x by trajectories of the differential inclusion (with a constant convex right-hand side)

$$\dot{x}(t) \in -F \ . \tag{4}$$

Observe that another interpretation via viscosity theory for Hamilton–Jacobi equations can be given. Namely, $\mathfrak{T}^F_C(\cdot)$ is nothing else than the (unique) *viscosity solution* to the equation

$$\rho_{F^0}(\nabla u(x)) - 1 = 0 \ , \qquad x \in H\backslash C \ , \tag{5}$$

with $u(x) = 0$ on C (here F^0 is the *polar set*). This is a natural generalization of the so-called *eikonal equation* arising from the geometric optics. For instance, if $H = \mathbb{R}^3$ and

$$F^0 = \left\{ \xi \in H : \sum_{i=1}^{3} c_i^2 \xi_i^2 \le 1 \right\},$$

then (5) describes the propagation of a light wave from a point source placed at the origin in (anisotropic) medium with the constant coefficients of refraction of light rays parallel to coordinate axes (denoted by c_i).

Notice that in the past many authors studied such *best approximation problem*

$$\min \{\rho_F (x-y) : \ y \in C\} \ . \tag{6}$$

For example, in [10, 20] the generic properties of (6) were established, while in the works [11, 18, 39, 40] the directional derivatives as well as various subdifferentials of the value (time-minimum) function $\mathfrak{T}_C^F(\cdot)$ were computed. Furthermore, Colombo and Wolenski gave in [18] the first sufficient condition guaranteeing the local well-posedness of the problem (6) as well as the regularity of $\mathfrak{T}_C^F(\cdot)$. Afterwards, in [25, 26] the authors essentially sharpened this condition and represented it as a certain balance between curvatures of the dynamics F and the target set C. Toward this end some quantitative results on convex duality in a Hilbert space were obtained in [25]. Besides that another (independent) "*first order*" hypothesis ensuring the well-posedness was proposed. It is written in terms of the balance between (external) normals to the sets F and C.

It turned out that the latter hypothesis can be easily adapted to a more general problem where a supplementary additive term appears. Namely, given a sufficiently regular function $\theta : C \to \mathbb{R}$ we are led to consider the mathematical programming problem

$$\min \{\rho_F (x-y) + \theta (y) : \ y \in C\} \ , \tag{7}$$

whose value function under an additional "slope assumption" is nothing else than the *viscosity solution* to the same *Hamilton–Jacobi equation* (5) but with the (general) boundary condition $u(x) = \theta(x)$, $x \in C$. Although the latter fact is well known (see, e.g., [9, 30]), for the sake of completeness we give in Sect. 3 its detailed proof emphasizing thereby one of the crucial interpretations of the problem (7). Section 4 instead is devoted to another interpretation in terms of an optimal time control problem, which somehow extends the problem with the constant dynamics (4) mentioned above.

In Sect. 5 we introduce basic assumptions, under which the well-posedness and regularity results are obtained. Notice that the main hypothesis is, roughly speaking, a sort of (Lipschitz) compatibility of the normal vectors to F, on the one hand, and of the (proximal) subdifferential to the restriction $\theta|_C$, on the other. Moreover, an auxiliary statement similar to Lemma 5.1 [25] is placed here. The (local) well-posedness of the problem (7) is proved in Sect. 6, while Sect. 7 is devoted to the regularity of the value function that includes its Fréchet differentiability and the (Hölder) continuity of the gradient near the target. We conclude in Sect. 8 with two examples, which illustrate the applicability and the novelty of obtained results even in finite dimensions.

The main results of the paper (without detailed proofs) were announced earlier in [27].

2 Preliminaries

Let us emphasize first the setting of the problem. Everywhere in our considerations we assume that H is a Hilbert space with the inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$, that $F \subset H$ is a convex closed bounded set containing the origin in the interior and that $C \subset H$ is an arbitrary nonempty closed subset. Given a sufficiently regular (e.g., Lipschitz continuous) real-valued function $\theta : C \to \mathbb{R}$ we are interested in the well-posedness of the problem (7), i.e., in the existence, uniqueness and stability of its minimizers as well as in some kind of regularity of the value function (denoted further by $\hat{u}(\cdot)$).

We define the *support function* $\sigma_F : H \to \mathbb{R}^+$,

$$\sigma_F(\xi^*) := \sup\{\langle \xi, \xi^* \rangle : \xi \in F\} \ ,$$

and recall the well-known identity

$$\rho_F(\xi) = \sigma_{F^0}(\xi) \ , \qquad \xi \in H \ , \tag{8}$$

where F^0 is the *polar set* of F. Hence

$$\frac{1}{\|F\|}\|\xi\| \le \rho_F(\xi) \le \left\|F^0\right\| \|\xi\| \ , \qquad \xi \in H \ , \tag{9}$$

where $\|F\| := \sup\{\|\xi\| : \xi \in F\}$, and $\rho_F(\cdot)$ is lipschitzean with the Lipschitz constant $\left\|F^0\right\|$.

In what follows we use also the so-called *duality mapping* $\mathfrak{J}_F : \partial F^0 \to \partial F$ that associates with each $\xi^* \in \partial F^0$ the set of all linear functionals $\langle \xi, \cdot \rangle$ with $\xi \in \partial F$ supporting F^0 in ξ^*. In other words,

$$\mathfrak{J}_F(\xi^*) := \{\xi \in \partial F : \langle \xi, \xi^* \rangle = 1\} \ .$$

Denoting by $\mathbf{N}_F(\xi)$ the *normal cone* to F at the point $\xi \in \partial F$ and by $\partial \rho_F(\xi)$ the *subdifferential* of the function $\rho_F(\cdot)$ in the sense of Convex Analysis we have other characterizations of the duality mapping:

$$\mathfrak{J}_F(\xi^*) = \partial \rho_{F^0}(\xi^*) \ ; \tag{10}$$

$$\mathfrak{J}_{F^0}(\xi) = \mathfrak{J}_F^{-1}(\xi) = \mathbf{N}_F(\xi) \cap \partial F^0 \ . \tag{11}$$

In particular, using positive homogeneity of the gauge function one easily deduces from (10) and (11) that

$$\partial \rho_F(\xi) = \mathbf{N}_F\left(\frac{\xi}{\rho_F(\xi)}\right) \cap \partial F^0 \ , \qquad \xi \neq 0 \ . \tag{12}$$

Following [25, Definition 3.2], for each dual pair (ξ, ξ^*) (i.e., $\xi^* \in \partial F^0$ and $\xi \in \mathfrak{J}_F(\xi^*)$) let us define the *modulus of rotundity*

$$\hat{\mathfrak{C}}_F(r,\xi,\xi^*) := \inf\{\langle \xi-\eta,\xi^*\rangle : \eta\in F,\ \|\xi-\eta\|\ge r\}\ ,\qquad r>0\ .$$

The set F is said to be *strictly convex* (or *rotund*) at ξ w.r.t. ξ^* if

$$\hat{\mathfrak{C}}_F(r,\xi,\xi^*)>0\quad \text{for all}\quad r>0\ . \tag{13}$$

If (13) is fulfilled, then ξ is an *exposed point* of F and, in particular, ξ is the unique element of $\mathfrak{J}_F(\xi^*)$. So, in this case ξ is well defined whenever ξ^* is fixed. Observe that there is a strong connection between the rotundity of F and the smoothness of F^0. Namely (see [25, Proposition 3.3 (iii)]), F is strictly convex at ξ w.r.t. ξ^* iff $\rho_{F^0}(\cdot)$ is Fréchet differentiable at ξ^* with $\nabla\rho_{F^0}(\xi^*)=\xi$. In this case we say also that F^0 is *smooth* at ξ^* (w.r.t. ξ).

Given a set $U\subset\partial F^0$ we say that F is *uniformly rotund* w.r.t. U if

$$\beta_U(r) := \inf\left\{\hat{\mathfrak{C}}_F(r,\xi,\xi^*) : \xi^*\in U\right\}>0\quad \text{for all}\quad r>0\ .$$

In [26, Proposition 2.1] the dual version of the latter property was given: the gauge F is uniformly rotund w.r.t. U if and only if the duality mapping $\mathfrak{J}_F(\cdot)$ is single-valued on U and uniformly continuous in the following sense

$$\sup_{\eta\in\mathfrak{J}_F(\eta^*)}\|\mathfrak{J}_F(\xi^*)-\eta\|\to 0\quad\text{as}\quad \|\xi^*-\eta^*\|\to 0\ ,\quad \xi^*\in U\ ,\quad \eta^*\in\partial F^0 \tag{14}$$

(we clearly identify $\mathfrak{J}_F(\xi^*)$ with its element whenever it is a singleton). Recalling the characterization of the duality mapping through the subdifferential of the Minkowski functional (see (10)) we derive from (14) that the uniform rotundity of F w.r.t. U implies, in particular, the uniform continuity of the *Fréchet gradient* $\nabla\rho_{F^0}(\cdot)$ on the set U.

In Sects. 4 and 7 we will use also the distance between sets $A,B\subset H$. So, let us remind the *Pompeiu–Hausdorff metric*:

$$\begin{aligned}\mathcal{D}(A,B) &:= \max\left\{\sup_{x\in A}\mathrm{d}_B(x),\sup_{y\in B}\mathrm{d}_A(y)\right\}\\ &= \inf\left\{r>0 : A\subset B+r\overline{\mathbf{B}}\ \text{ and }\ B\subset A+r\overline{\mathbf{B}}\right\}\ .\end{aligned} \tag{15}$$

Here and in what follows $\overline{\mathbf{B}}$ means the closed unit ball in H. It is well known that the family $\operatorname{conv}H$ of all nonempty convex closed bounded subsets of H supplied with the above distance is *isometrically embedded* into the space of real continuous functions defined on H as a complete cone, and the respective isometry is given by the formula:

$$\mathcal{D}(A,B) = \sup_{\|v\|=1}|\sigma_A(v)-\sigma_B(v)|\ . \tag{16}$$

Given now $F\in\operatorname{conv}H$ with nonempty interior and $v\in\operatorname{int}F$ let us denote by

$$r_F(v) := \sup\left\{r>0 : v+r\overline{\mathbf{B}}\subset F\right\}\ . \tag{17}$$

Being the set $(F-v)^0$ convex closed and bounded we have the following (local) Lipschitz inequality for the mapping $v \mapsto (F-v)^0$:

$$\mathcal{D}\left((F-v_1)^0,(F-v_2)^0\right) \le \frac{1}{r_F(v_1)\, r_F(v_2)}\,|v_1-v_2| \ , \tag{18}$$

$v_1, v_2 \in \operatorname{int} F$. It was obtained in [19, Lemma 2] for $H=\mathbb{R}^n$ but readily can be adapted to the Hilbert case.

In the rest of this section let us give some concepts and notations of Nonsmooth Analysis, which will be used in the sequel.

Given a proper lower semicontinuous function $f : H \to \mathbb{R} \cup \{+\infty\}$ we denote by $\partial^p f(x)$, $\partial^- f(x)$, $\partial^l f(x)$ and $\partial^c f(x)$ the *proximal*, *Fréchet*, *limiting* (*Mordukhovich*) *and Clarke–Rockafellar subdifferential* of $f(\cdot)$ at a point x, $f(x) < +\infty$, respectively. Let us recall their definitions (see, e.g., [12, 14, 31]):

- $\partial^p f(x) := \{\zeta \in H : \exists \eta > 0, \quad \sigma \ge 0 \quad \text{such that} \quad f(y) \ge f(x) + \langle \zeta, y-x\rangle - \sigma\|y-x\|^2 \quad \forall y \ , \ \|y-x\| \le \eta\}$;
- $\partial^- f(x) := \left\{\zeta \in H : \liminf\limits_{y\to x} \frac{f(y)-f(x)-\langle \zeta, y-x\rangle}{\|y-x\|} \ge 0\right\}$;
- $\partial^l f(x) := w\text{-}\limsup\limits_{\substack{f\\ y\to x}} \partial^- f(y) = \left\{w - \lim\limits_{i\to\infty} \zeta_i : \zeta_i \in \partial^- f(x_i),\ x_i \to x,\ f(x_i) \to f(x)\right\}$;
- $\partial^c f(x) := \{\zeta \in H : \langle \zeta, v\rangle \le f^\uparrow(x;v) \quad \forall v \in H\}$.

Here "w-$\lim\limits_{i\to\infty}$" stands for the weak limit, and

$$f^\uparrow(x;v) := \lim_{\varepsilon\to 0+} \limsup_{\substack{f\\ y\to x,\ h\to 0+}} \inf_{\|w-v\|\le\varepsilon} \frac{f(y+hw)-f(y)}{h}$$

is the *Rockafellar's generalized directional derivative* (see [35]); $y \xrightarrow{f} x$ means the convergence $y \to x$ together with $f(y) \to f(x)$.

Moreover, in order to treat viscosity solutions in the next section we define the *Fréchet superdifferential* $\partial^+ f(x)$ as the counterpart to $\partial^- f(x)$ assuming that the function $f(\cdot)$ is upper semicontinuous at x:

$$\partial^+ f(x) := \left\{\zeta \in H : \limsup_{y\to x} \frac{f(y)-f(x)-\langle \zeta, y-x\rangle}{\|y-x\|} \le 0\right\} \ . \tag{19}$$

It is well known (see, e.g., [31, p. 90]) that for a continuous function $f(\cdot)$ both $\partial^- f(x)$ and $\partial^+ f(x)$ are nonempty simultaneously if and only if $f(\cdot)$ is *Fréchet differentiable* at the point x. In this case $\partial^- f(x) = \partial^+ f(x) = \{\nabla f(x)\}$.

Notice that the inclusions

$$\partial^p f(x) \subset \partial^- f(x) \subset \partial^l f(x) \subset \partial^c f(x) \tag{20}$$

are always valid, and that $f^{\uparrow}(x;v)$ is reduced to the *Clarke's directional derivative*

$$f^{o}(x;v) := \limsup_{y\to x,\, h\to 0^{+}} \frac{f(y+hv)-f(y)}{h} \tag{21}$$

whenever $f(\cdot)$ is lipschitzean around x. In the latter case $\partial^{c} f(x)$ is bounded and can be represented as the convex closed hull of the limiting subdifferential. Taking into account that in turn $\partial^{l} f(x)$ can be expressed through proximal subgradients in the place of Fréchet ones (see [31, p. 240]), we have

$$\partial^{c} f(x) = \overline{\mathrm{co}} \left\{ w - \lim_{i\to\infty} \zeta_i : \zeta_i \in \partial^{p} f(x_i), \qquad x_i \to x \right\} \tag{22}$$

(see also [14, p. 88]). A function $f(\cdot)$ is said to be *proximally* (*lower*, *Clarke*) *regular* at x if $\partial^{p} f(x) = \partial^{l} f(x)$ (respectively, $\partial^{-} f(x) = \partial^{l} f(x)$ or $\partial^{-} f(x) = \partial^{c} f(x)$).

If $f(\cdot)$ is convex, then all the subdifferentials above coincide with the subdifferential in the sense of Convex Analysis. If instead $f(\cdot)$ is (Fréchet) continuously differentiable at x, then we can only affirm that $\partial^{-} f(x) = \partial^{l} f(x) = \partial^{c} f(x) = \{\nabla f(x)\}$ whereas the proximal subdifferential may be empty. However, this does not occur if the gradient $\nabla f(\cdot)$ is lipschitzean near x. So, in the latter case also $\partial^{p} f(x) = \{\nabla f(x)\}$. Let us observe that even Hölder continuity of $\nabla f(\cdot)$ with an exponent $0 < \alpha < 1$ does not guarantee the proximal regularity.

Given an open set $U \subset H$ in what follows we denote by $\mathcal{C}^{1,\alpha}_{\mathrm{loc}}(U)$, $0 < \alpha \leq 1$, the class of all continuously differentiable functions $f(\cdot)$ whose gradient $\nabla f(\cdot)$ is locally Hölderean on U with the exponent α. In this case we say that $f(\cdot)$ is of class $\mathcal{C}^{1,\alpha}_{\mathrm{loc}}$ on U.

The various notions of normal cones to a closed set C (all of them coincide with the cone $\mathbf{N}_C(x)$ if C is convex) can be given through the respective subdifferentials of the indicator function $\mathbf{I}_C(\cdot)$ equal to 0 on C and to $+\infty$ elsewhere. Thus, the *proximal*, *Fréchet* (ou *weak Bouligand*), *limiting* (or *Mordukhovich*) and *Clarke normal cones* are defined and denoted by $\mathbf{N}^{p}_C(x)$, $\mathbf{N}^{\sigma}_C(x)$, $\mathbf{N}^{l}_C(x)$, $\mathbf{N}^{c}_C(x)$, respectively. Various properties of the normal cones (as well as of the subdifferentials of lower semicontinuous functions) can be found, e.g., in [2, 6, 12, 14, 31, 36]). Similarly as for the subdifferentials, a closed set C is said to be *proximally* (*normally*, *Clarke*) *regular* at $x \in \partial C$ if $\mathbf{N}^{p}_C(x) = \mathbf{N}^{l}_C(x)$ (respectively, $\mathbf{N}^{\sigma}_C(x) = \mathbf{N}^{l}_C(x)$ or $\mathbf{N}^{\sigma}_C(x) = \mathbf{N}^{c}_C(x)$).

We say that a closed set $C \subset H$ *has smooth* (or $\mathcal{C}^1$) *boundary at* $x_0 \in \partial C$ if for each $x \in \partial C$ enough close to x_0 the limiting cone $\mathbf{N}^{l}_C(x)$ is reduced to $\mathfrak{n}_C(x)\,\mathbb{R}^{+}$ with some continuous function $\mathfrak{n}_C(\cdot)$, $\|\mathfrak{n}_C(x)\| = 1$. If, moreover, $\mathfrak{n}_C(\cdot)$ is Hölder continuous with an exponent $0 < \alpha \leq 1$, then we say that C *has* $\mathcal{C}^{1,\alpha}$*-boundary at* x_0.

In what follows by the *restriction* $\theta|_C$ we mean the function equal to $\theta(x)$ on C and to $+\infty$ elsewhere. If $\theta(\cdot)$ is defined also out of C, then clearly $\theta|_C = \theta + \mathbf{I}_C$. Due to this representation and to the proximal "sum rule" $\partial^{p} f + \partial^{p} g \subset \partial^{p}(f+g)$ we have that

- the subdifferential $\partial^{p}(\theta|_C)(x)$ is unbounded whenever $\mathbf{N}^{p}_C(x) \neq \{0\}$;
- $\partial^{p}(\theta|_C)(x) = \partial^{p}\theta(x)$ whenever $x \in \mathrm{int}\, C$.

3 Marginal Function as the Viscosity Solution

Let $\Omega \subset H$ be a nonempty open set and $\Gamma : \Omega \times \mathbb{R} \times H \to \mathbb{R}$ be a continuous mapping. Let us remind that a continuous function $u : \overline{\Omega} \to \mathbb{R}$ is said to be *viscosity solution* of the (stationary) *Hamilton–Jacobi equation*

$$\Gamma(x, u(x), \nabla u(x)) = 0 \tag{23}$$

if the following two conditions are fulfilled:

(I) $\Gamma(x, u(x), p) \leq 0$ for each $x \in \Omega$ and each $p \in \partial^+ u(x)$;
(II) $\Gamma(x, u(x), p) \geq 0$ for each $x \in \Omega$ and each $p \in \partial^- u(x)$.

For the main results of the viscosity theory for Hamilton–Jacobi equations we refer to [3] and to the bibliography therein (for a concise survey of viscosity solutions in finite dimensions see also the tutorial lessons by Bressan [7]). In particular, it is known that for each suitable boundary function $\theta(\cdot)$ there exists an unique viscosity solution $u(\cdot)$ of the Eq. (23) such that $u|_{\partial\Omega} = \theta$. Moreover, this solution is stable w.r.t. θ. Notice also that the class of viscosity solutions is consistent with other types of solutions. For instance, any continuously differentiable (by Fréchet) function $u(\cdot)$ satisfying (23) everywhere in Ω (so named *classical solution*) is also a viscosity solution.

Now we consider a hamiltonian Γ depending only on the gradient. Then, as shown in [9], under some geometric conditions the Hamilton–Jacobi equation (23) can be reduced to a particular case, where the hamiltonian takes the form $\rho_{F^0}(\xi) - 1$ with an appropriate gauge F [see (5) with $\Omega = H \backslash C$].

So, in what follows we deal with the boundary value problem for the Eq. (5), assuming that the boundary function $\theta : C \to \mathbb{R}$ satisfies the *slope condition* with respect to F, namely,

$$\theta(x) - \theta(y) \leq \rho_F(x - y) \qquad \forall x, y \in C \ . \tag{24}$$

Remark 1. By (9) the inequality (24) implies the Lipschitz continuity of the function $\theta(\cdot)$ on C with the Lipschitz constant $\left\|F^0\right\|$. Moreover, the function

$$\hat{u}(x) := \inf_{y \in C} \{\rho_F(x - y) + \theta(y)\} \tag{25}$$

is a sort of extension of $\theta(\cdot)$ to the whole H with keeping the property (24) (a generalization of McShane lemma, see also [9, Lemma 4.1]).

Indeed, on the one hand, it follows directly from (24) that $\hat{u}(x) = \theta(x)$ for all $x \in C$. On the other hand, given $y \in H$ and $\varepsilon > 0$ we find $z_y \in C$ such that

$$\hat{u}(y) \geq \rho_F(y - z_y) + \theta(z_y) - \varepsilon \ .$$

Then for each $x \in H$

$$\begin{aligned} \hat{u}(x) - \hat{u}(y) &\leq \rho_F(x - z_y) - \rho_F(y - z_y) + \varepsilon \\ &\leq \rho_F(x - y) + \varepsilon \ . \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we arrive at the same slope condition as (24):

$$\hat{u}(x) - \hat{u}(y) \le \rho_F(x-y) \qquad \forall x, y \in H \ . \tag{26}$$

It implies, in particular, that $\hat{u}(x)$ *admits finite value* for each $x \in H$ (in fact, $\hat{u}(x) \ge \theta(x_0) - \rho_F(x_0 - x)$, $x \in H$, where $x_0 \in C$ is an arbitrary fixed point).

Theorem 1. *If the inequality (24) is fulfilled, then the convolution (25) is the (unique) viscosity solution of the equation (5) such that* $\hat{u}(x) = \theta(x)$, $x \in C$.

Proof. In conformity with the definition above the proof splits into two parts as follows.

(I) Let us fix $x \notin C$ and $p \in \partial^+ \hat{u}(x)$. Then, given $\varepsilon > 0$ by using the formula (19) we find $\delta > 0$ such that

$$\hat{u}(y) - \hat{u}(x) - \langle p, y - x \rangle \le \varepsilon \|y - x\|$$

whenever $\|y - x\| \le \delta$. In the case $y \ne x$ dividing the latter inequality by $\rho_F(x-y)$ and taking into account that

$$\frac{\hat{u}(y) - \hat{u}(x)}{\rho_F(x-y)} \ge -1$$

[see (26)] we obtain

$$-1 + \left\langle p, \frac{x-y}{\rho_F(x-y)} \right\rangle \le \varepsilon \frac{\|x-y\|}{\rho_F(x-y)} \ .$$

Hence,

$$-1 + \sup_{0 < \|y-x\| < \delta} \left\langle p, \frac{x-y}{\rho_F(x-y)} \right\rangle \le \varepsilon \sup_{0 < \|y-x\| < \delta} \frac{\|y-x\|}{\rho_F(y-x)} \ . \tag{27}$$

Since, by the positive homogeneity of the gauge function,

$$\sup_{0 < \|y-x\| < \delta} \frac{\|y-x\|}{\rho_F(y-x)} = \sup_{z \ne 0} \frac{\|z\|}{\rho_F(z)} = \|F\|$$

and

$$\sup_{0 < \|y-x\| < \delta} \left\langle p, \frac{x-y}{\rho_F(x-y)} \right\rangle = \sup_{z \in \partial F} \langle p, z \rangle = \sigma_F(p) \ ,$$

it follows from (27) and (8) that

$$-1 + \rho_{F^0}(p) \le \varepsilon \|F\| \ .$$

Letting $\varepsilon \to 0^+$ we obtain $\rho_{F^0}(p) \leq 1$.

(II) Let us fix now $p \in \partial^- \hat{u}(x)$ ($x \notin C$ is given). We should prove that $\rho_{F^0}(p) \geq 1$. Assuming the contrary, we choose $\varepsilon > 0$ so small that $\rho_{F^0}(p) < 1 - \varepsilon$. Then, by the definition of the Fréchet subdifferential and by (9) there exists δ, $0 < \delta < \mathrm{d}_C(x)$, with

$$\hat{u}(x) - \hat{u}(y) + \langle p, y - x \rangle \leq \frac{\varepsilon}{4} \rho_F(x - y) \tag{28}$$

whenever $\|y - x\| \leq \delta$. On the other hand, let us take $z_x \in C$ such that

$$\hat{u}(x) \geq \rho_F(x - z_x) + \theta(z_x) - \frac{\varepsilon\delta}{8\|F\|} \ ,$$

and, consequently,

$$\hat{u}(x) - \hat{u}(y) \geq \rho_F(x - z_x) - \rho_F(y - z_x) - \frac{\varepsilon\delta}{8\|F\|} \qquad \forall y \notin C. \tag{29}$$

Since $\|z_x - x\| > \delta$ (by the choice of $\delta > 0$), there exists $y_x \in [x, z_x]$ with $\|y_x - x\| = \delta/2$. Representing this point as $\lambda x + (1 - \lambda) z_x$ for some $0 < \lambda < 1$, we clearly have

$$\begin{aligned} \rho_F(x - y_x) &= (1 - \lambda) \rho_F(x - z_x) \ ; \\ \rho_F(y_x - z_x) &= \lambda \rho_F(x - z_x) \ . \end{aligned}$$

Consequenlty (see also (9)),

$$\rho_F(x - z_x) - \rho_F(y_x - z_x) = \rho_F(x - y_x) \geq \frac{\delta}{2\|F\|} \ . \tag{30}$$

Therefore, applying successively (30), (29) and (28), we obtain

$$\begin{aligned} &\rho_F(x - y_x) - \frac{\varepsilon\delta}{8\|F\|} + \langle p, y_x - x \rangle \\ &= \rho_F(x - z_x) - \rho_F(y_x - z_x) - \frac{\varepsilon\delta}{8\|F\|} + \langle p, y_x - x \rangle \\ &\leq \hat{u}(x) - \hat{u}(y_x) + \langle p, y_x - x \rangle \leq \frac{\varepsilon}{4} \rho_F(x - y_x) . \end{aligned} \tag{31}$$

It follows from the inequality $\rho_{F^0}(p) \geq \left\langle p, \frac{x - y_x}{\rho_F(x - y_x)} \right\rangle$ [see (8)] and from the choice of $\varepsilon > 0$ that

$$\langle p, x - y_x \rangle < \rho_F(x - y_x)(1 - \varepsilon) \ .$$

Hence, recalling (31) and (30) we obtain

$$\begin{aligned} \rho_F(x - y_x) &\leq \frac{\varepsilon\delta}{8\|F\|} + \langle p, x - y_x \rangle + \frac{\varepsilon}{4} \rho_F(x - y_x) \\ &< \frac{\varepsilon}{4} \rho_F(x - y_x) + \frac{\varepsilon}{4} \rho_F(x - y_x) + \rho_F(x - y_x)(1 - \varepsilon) \\ &= \rho_F(x - y_x) \left(1 - \frac{\varepsilon}{2}\right) \ , \end{aligned}$$

which is a contradiction.

Combining the parts (I) and (II) proves the theorem. □

Remark 2. Observe that the slope condition (24) is always fulfilled if $\theta(\cdot)$ is defined and Lipschitz continuous on an open convex neighbourhood U of C and either $\nabla\theta(x) \in F^0$ for a.e. $x \in U$ (in finite dimensions), or $\partial^c\theta(x) \subset F^0 \quad \forall x \in U$ (in general case). This immediately follows from Lebourg's theorem (see [12, p. 41]). Vice versa (we use this property in the sequel), if $\theta(\cdot)$ is defined and satisfies (24) on a neighbourhood $U(\hat{x})$ of $\hat{x} \in C$, then $\partial^c\theta(\hat{x}) \subset F^0$. Indeed, it follows directly from (24) that given arbitrary $v \in H$ for each $x \in U(\hat{x})$ and sufficiently small $h > 0$ we have

$$\frac{\theta(x+hv) - \theta(x)}{h} \leq \rho_F(v) \ .$$

Then, passing to lim sup as $h \to 0+$ and $x \to \hat{x}$ we conclude from both (8) and (21) that $\theta^o(\hat{x}; v) \leq \sigma_{F^0}(v)$. So, the definition of the Clarke subdifferential gives:

$$\partial^c\theta(\hat{x}) \subset \{\zeta \in H : \langle v, \zeta\rangle \leq \sigma_{F^0}(v) \quad \forall v \in H\} = F^0 \ .$$

4 Marginal Function as the Minimal Time

In this section we relate the function (25) with an optimal time control problem having, in general, nonconstant (autonomous) dynamics. However, in order to do this we should impose much stronger hypothesis on the function $\theta(\cdot)$.

Theorem 2. *Let $U \subset H$ be an open convex set with $U \supset C$, and $\theta : U \to \mathbb{R}$ be a (Fréchet) continuously differentiable function such that*

$$\nabla\theta(x) \in \operatorname{int} F^0 \qquad \forall x \in U \ . \tag{32}$$

Then for each $x \in U$ the equality

$$\hat{u}(x) = \mathfrak{T}_C^{F,\theta}(x) + \theta(x) \tag{33}$$

holds, where $\mathfrak{T}_C^{F,\theta}(x)$ is the minimal time necessary to achieve (the boundary of) the set C from the point $x \in U$ by trajectories of the differential inclusion

$$\dot{x}(t) \in \left(-F^0 + \nabla\theta(x(t))\right)^0 \tag{34}$$

remaining inside U.

Proof. Observe first that under the condition (32) the right-hand side of the inclusion (34) is bounded for each $x \in U$ that is essential for proving (33). Moreover, it is easy to see that in this case the inequality (24) is strict.

Let us prove first that

$$\hat{u}(x) \leq \mathfrak{T}_C^{F,\theta}(x) + \theta(x) \quad . \tag{35}$$

To this end fix $x \in U$ and assume that $\mathfrak{T}_C^{F,\theta}(x) < +\infty$ [otherwise (35) is trivial]. So, the target C can be achieved from x by some trajectory $x(\cdot)$ of the inclusion (34), $x(t) \in U$, in some time moment $T > 0$.

On the other hand, since the function $\hat{u}(\cdot)$ satisfies on H the inequality (26), it admits nonempty, convex and closed Clarke subdifferential $\partial^c \hat{u}(z) \subset F^0$, $z \in H$ (see Remark 2). In other words, setting $g(z) := \hat{u}(z) - \theta(z)$ we have

$$-\partial^c g(z) \subset -F^0 + \nabla\theta(z) \tag{36}$$

(see [12, Propositions 2.3.1 and 2.3.2]). Now, the relations (34) and (36) imply that $\langle -p, \dot{x}(t)\rangle \leq 1$ for all $p \in \partial^c g(x(t))$ and a.e. $t \in [0,T]$. It follows then from [12, p. 27] that

$$g^o(x(t), -\dot{x}(t)) = \max_{p \in \partial^c g(x(t))} \langle p, -\dot{x}(t)\rangle \leq 1 \quad . \tag{37}$$

Let us consider the superposition $t \mapsto g(x(t))$, which is Lipschitz continuous because $x(t) \in U$, $t \in [0,T]$, and the right-hand side of (34) is bounded. Therefore, $t \mapsto g(x(t))$ admits derivative at a.e. $t \in [0,T]$. By (21) and (37) we successively obtain

$$\begin{aligned}
\frac{d}{dt} g(x(t)) &= -\limsup_{h \to 0+} \frac{g(x(t-h)) - g(x(t))}{h} \\
&= -\limsup_{h \to 0+} \frac{g(x(t) - h\dot{x}(t)) - g(x(t))}{h} \\
&\geq -g^o(x(t), -\dot{x}(t)) \geq -1
\end{aligned}$$

a.e. on $[0,T]$. Hence, by integrating the latter inequality on the interval $[0,T]$ and taking into account that $g(x(T)) = 0$ (due to the boundary condition on C) we have

$$-g(x) = g(x(T)) - g(x(0)) = \int_0^T \frac{d}{dt} g(x(t))\, dt \geq -T \quad .$$

Since the instant $T > 0$ was chosen arbitrarily, we arrive at (35).

In order to prove the opposite inequality we fix $x \in U$, $\varepsilon > 0$ and choose $z_x \in C$ such that

$$\hat{u}(x) \geq \rho_F(x - z_x) + \theta(z_x) - \varepsilon. \tag{38}$$

We should find a trajectory $x(\cdot)$ of the inclusion (34) remaining in U such that $x(0)=x$ and $x(T)=z_x\in C$ where

$$T:=\rho_F(x-z_x)+\theta(z_x)-\theta(x)>0 \tag{39}$$

[see (32)]. To this end we define first an approximative sequence $\{x_n(\cdot)\}$ as follows.

Given $n=1,2,\ldots$ let us divide the segment $[x,z_x]$ into small parts by the points $x_i^n:=x+\frac{i}{n}(z_x-x)\in U$, $i=1,2,\ldots,n$. Denote by

$$T_i^n:=\rho_F(x-x_i^n)+\theta(x_i^n)-\theta(x)$$

and observe that $T_0^n=0$, $T_n^n=T$, $T_i^n>0$ and

$$h_i^n:=T_i^n-T_{i-1}^n=\rho_F\left(x_{i-1}^n-x_i^n\right)+\theta(x_i^n)-\theta\left(x_{i-1}^n\right)>0\ , \tag{40}$$

$i=1,2,\ldots,n$, due to the strict slope condition. Defining on $[0,T]$ the continuous piecewise affine function

$$x_n(t):=x_{i-1}^n+\frac{t-T_{i-1}^n}{h_i^n}\left(x_i^n-x_{i-1}^n\right)\ ,\qquad t\in\left[T_{i-1}^n,T_i^n\right]\ , \tag{41}$$

we clearly have $x_n(0)=x$, $x_n(T)=z_x\in C$, $x_n(T_i^n)=x_i^n$, and

$$x_n(t)\in[x,z_x]\subset U\ ,\quad t\in[0,T]\ . \tag{42}$$

The composed function $t\mapsto\theta(x_n(t))$ is continuously differentiable on each interval $\left]T_{i-1}^n,T_i^n\right[$, $i=1,2,\ldots,n$, and by the mean value theorem there exists $\tau_i^n\in\left]T_{i-1}^n,T_i^n\right[$ such that

$$\begin{aligned}\theta(x_i^n)&=\theta\left(x_{i-1}^n\right)+\frac{d}{dt}\theta(x_n(t))\Big|_{t=\tau_i^n}h_i^n\\&=\theta\left(x_{i-1}^n\right)+\left\langle\nabla\theta(x_n(\tau_i^n)),x_i^n-x_{i-1}^n\right\rangle\end{aligned} \tag{43}$$

[see (41)]. Combining (40), (43) and (8) we have that

$$\begin{aligned}h_i^n&=\theta(x_i^n)-\theta\left(x_{i-1}^n\right)+\rho_F\left(x_{i-1}^n-x_i^n\right)\\&=\left\langle-\nabla\theta(x_n(\tau_i^n)),x_{i-1}^n-x_i^n\right\rangle+\sigma_{F^0}\left(x_{i-1}^n-x_i^n\right)\\&\geq\left\langle-\xi^*+\nabla\theta(x_n(\tau_i^n)),x_i^n-x_{i-1}^n\right\rangle\end{aligned}$$

whenever $\xi^*\in F^0$. Consequently,

$$\dot{x}_n(t)=\frac{x_i^n-x_{i-1}^n}{h_i^n}\in\left(-F^0+\nabla\theta(x_n(\tau_i^n))\right)^0 \tag{44}$$

for all $t\in\left]T_{i-1}^n,T_i^n\right[$, $i=1,2,\ldots,n$.

Thus, it remains to prove the convergence of $\{x_n(\cdot)\}$ (up to a subsequence) to a desired trajectory. To this end we observe, first, that the functions $x_n(\cdot)$ admit values in the same compact set $[x,z_x]$ [see (42)]. Furthermore, since there exists $r>0$ with

$$\nabla\theta(z) + r\overline{\mathbf{B}} \subset F^0 \tag{45}$$

whenever $z \in [x, z_x]$ [see (32)], passing to the polar sets and recalling (44) we have that

$$\|\dot{x}_n(t)\| \leq \left\| \left(-F^0 + \nabla\theta\left(x_n\left(\tau_i^n\right)\right)\right)^0 \right\| \leq \frac{1}{r} \ , \quad t \in \left]T_{i-1}^n, T_i^n\right[\ , \quad i = 1, 2, \ldots, n \ ,$$

$n = 1, 2, \ldots$. Therefore, applying successively Askoli and Banach-Alaoglu theorems without loss of generality we can assume that both the sequence $\{x_n(\cdot)\}$ converges uniformly on $[0, T]$ to some Lipschitz continuous function $x(\cdot)$ and $\{\dot{x}_n(\cdot)\}$ converges weakly in the space $L^\infty([0, T], H)$ to the derivative $\dot{x}(\cdot)$, which exists almost everywhere on $[0, T]$. Then, by Mazur's Lemma there exists a sequence $\{v_n(\cdot)\}$ of convex combinations of the functions $\dot{x}_n(\cdot)$, converging to $\dot{x}(\cdot)$ strongly in $L^\infty([0, T], H)$ and, consequently, almost everywhere on $[0, T]$. Let $\mathcal{T}$ be the set of all $t \in [0, T], t \neq T_i^n$, such that the derivative $\dot{x}(t)$ exists and $v_n(t) \to \dot{x}(t), n \to \infty$. Clearly, $\mathcal{T}$ is a set of full measure in $[0.T]$.

Fix now $t \in \mathcal{T}$ and choose i_n, $n = 1, 2, \ldots$, with $t \in \left]T_{i_n-1}^n, T_{i_n}^n\right[$. Since by (40), (9) and (24)

$$h_{i_n}^n = T_{i_n}^n - T_{i_n-1}^n \leq 2\left\|F^0\right\| \left\|x_{i_n-1}^n - x_{i_n}^n\right\| = \frac{2}{n}\left\|F^0\right\| \|z_x - x\| \to 0 \ ,$$

we have that $\tau_{i_n}^n \to t$ as $n \to \infty$. Then also $x_n\left(\tau_{i_n}^n\right) \to x(t)$ and $\nabla\theta\left(x_n\left(\tau_{i_n}^n\right)\right) \to \nabla\theta(x(t))$, $n \to \infty$ (we use here the continuity of the gradient $\nabla\theta(\cdot)$). Hence, recalling (18), (17) and (45) we obtain

$$\begin{aligned} &\mathcal{D}\left(\left(-F^0 + \nabla\theta\left(x_n\left(\tau_{i_n}^n\right)\right)\right)^0, \left(-F^0 + \nabla\theta(x(t))\right)^0\right) \\ &\leq \frac{1}{r^2}\left\|\nabla\theta\left(x_n\left(\tau_{i_n}^n\right)\right) - \nabla\theta(x(t))\right\| \to 0 \ . \end{aligned}$$

In particular, given $\delta > 0$ one can choose $N = N(\delta)$ such that

$$\left(-F^0 + \nabla\theta\left(x_n\left(\tau_{i_n}^n\right)\right)\right)^0 \subset \left(-F^0 + \nabla\theta(x(t))\right)^0 + \delta\overline{\mathbf{B}} \tag{46}$$

whenever $n \geq N$. Combining (46) with (44) by the convexity of the right-hand side of (46) it follows that

$$v_n(t) \in \left(-F^0 + \nabla\theta(x(t))\right)^0 + \delta\overline{\mathbf{B}}$$

for all $n \geq N$. Passing now to the limit and taking into account the arbitrarity of $\delta > 0$, we conclude that $x(\cdot)$ is indeed a solution of the differential inclusion (34) with $x(0) = x$ and $x(T) = z_x \in C$. Moreover, $x(\cdot)$ admits values in U because all the approximate solutions $x_n(\cdot)$ map $[0, T]$ into the compact segment $[x, z_x] \subset U$. Thus [see (38) and (39)],

$$\mathfrak{T}_C^{F,\theta}(x) \le T \le \hat{u}(x) - \theta(x) + \varepsilon \ ,$$

and getting $\varepsilon \to 0+$ we prove the second part of the theorem. □

Remark 3. Notice that although the right-hand side in (34) is nonconstant and the trajectories realizing the optimal time are, in general, nonaffine (as one can see from the second part of the proof above), this time optimal control problem satisfies an essential property that the target set is achieved for the shortest time along one fixed direction (or close to that).

5 Auxiliary Statement and Standing Assumptions

Our goal is to study regularity properties of $\hat{u}(\cdot)$ [see (25)], which (see the Sects. 3 and 4) can be seen either as viscosity solution of a stationary Hamilton–Jacobi equation or as translated value function in an associated optimal time control problem. Such regularity is strictly related to the existence, uniqueness and stability of minimizers of $y \mapsto \rho_F(x-y) + \theta(y)$ on C. In the particular case $\theta \equiv 0$ this relation was well studied in [18, 25, 26], while for a general marginal function $T(x)$ and a compact set $C = C(x)$ [see (1)] we find a justification of this property, for instance, in the result by F. Clarke on representation of the generalized gradient $\partial^c T(x)$ as a family of integrals of $f(x,\cdot)$ with respect to all Radon measures supported on the set of minimizers $\Pi(x)$ (see [12, p. 86]). In the sequel the set $\Pi(x)$ for the problem (7) will be denoted by $\pi_C^{F,\theta}(x)$, and we keep the same notation for its element if $\pi_C^{F,\theta}(x)$ is a singleton.

Our standing hypothesis in what follows is a slightly strengthened slope condition [compare with (24)]:

(**H**) there exists $0 < \gamma < \frac{1}{\|F\|\|F^0\|}$ such that

$$\theta(x) - \theta(y) \le \gamma \rho_F(x-y) \tag{47}$$

for all $x, y \in C$.

Extending if necessary $\theta(\cdot)$ in a suitable way (see Remark 1), without loss of generality we can assume that the function $\theta(\cdot)$ is defined and satisfies (47) on the whole space H. Due to Remark 2 the inequality (47) can be equivalently written as the inclusion

$$\partial^c \theta(x) \subset \gamma F^0 \ , \qquad x \in H \ . \tag{48}$$

By (9) and the convexity of F^0 it follows from (48) that

$$\partial^c \theta(x) + \frac{1-\gamma}{\|F\|}\overline{\mathbf{B}} \subset \partial^c \theta(x) + (1-\gamma)F^0 \subset F^0 \ , \tag{49}$$

or, recalling the definition (17),

$$r_{F^0}(\zeta) \geq \frac{1-\gamma}{\|F\|} \qquad \forall \zeta \in \partial^c \theta(x) \ , \quad x \in H \ . \tag{50}$$

On the other hand, passing in (49) to polar sets we have

$$\left\| \left(F^0 - \zeta\right)^0 \right\| \leq \frac{\|F\|}{1-\gamma} \qquad \forall \zeta \in \partial^c \theta(x) \ , \quad x \in H \ . \tag{51}$$

In particular, if $\theta(\cdot)$ is (Fréchet) differentiable at x, then it follows from (50) and (51) that

$$r_{F^0}(\nabla \theta(x)) \geq \frac{1-\gamma}{\|F\|} \tag{52}$$

and

$$\left\| \left(F^0 - \nabla \theta(x)\right)^0 \right\| \leq \frac{\|F\|}{1-\gamma} \ , \tag{53}$$

respectively (these estimates will be used further in Sect. 7).

Besides (**H**) in what follows we need a certain "slope-preserving" compatibility of the set C and the function $\theta(\cdot)$. Namely, set the following hypothesis:

($\hat{\mathbf{H}}$) for each $x \in \partial C$ there exist a (possibly empty) convex set $\Gamma(x) \subset \gamma F^0$ and a (possibly trivial) convex cone $\mathbf{N}_C^\theta(x)$ such that

$$\partial^p \left(\theta|_C\right)(x) = \Gamma(x) + \mathbf{N}_C^\theta(x) \ . \tag{54}$$

So, the subdifferential $\partial^p \left(\theta|_C\right)(x)$ is empty if and only if $\Gamma(x) = \varnothing$, while it is bounded iff $\mathbf{N}_C^\theta(x) = \{0\}$. Let us denote by $\partial^\theta C$ the part of ∂C consisting of the points x where the cone $\mathbf{N}_C^\theta(x)$ is nontrivial. Notice that in the case $\theta \equiv 0$ we have $\Gamma(x) = \{0\}$, $\mathbf{N}_C^\theta(x) = \mathbf{N}_C^p(x)$ for each $x \in \partial C$, and $\partial^\theta C = \partial^* C$ is the *reduced boundary* in the sense of [25, 26].

Observe that the equality (54) holds with $\Gamma(x) = \partial^p \theta(x)$ and $\mathbf{N}_C^\theta(x) = \mathbf{N}_C^p(x)$, in particular, whenever either both $\theta(\cdot)$ and C are proximally regular [because for the proximal subdifferentials the inclusion $\partial^p \left(\theta|_C\right)(x) \supset \partial^p \theta(x) + \mathbf{N}_C^p(x)$ always holds, while for the limiting ones we have $\partial^l \left(\theta|_C\right)(x) \subset \partial^l \theta(x) + \mathbf{N}_C^l(x)$ (see, e.g., [14, p. 62])] or $\theta(\cdot)$ is of class $\mathcal{C}^{1,1}$ near a given point (that can be proved easily by the same line as Proposition 2.11 [14, p. 38]). In the latter case, moreover, $\Gamma(x) = \{\nabla \theta(x)\}$.

Let us prove now an auxiliar assertion giving a property of minimizing sequences in (25), which generalizes the similar result [25, Lemma 5.1] obtained for the case $\theta \equiv 0$. We use here some tools of Variational and Proximal Analysis.

Lemma 1. *Let us suppose the standing assumptions* (**H**) *and* ($\hat{\mathbf{H}}$). *Then given a point* $z \in H \setminus C$ *and a minimizing sequence* $\{x_n\} \subset C$ *for the function* $x \mapsto$

$\rho_F(z-x)+\theta(x)$ on C one can find another minimizing sequence $\{x_n'\}\subset\partial^\theta C$ and sequences $\{x_n''\}$, $\{\mathbf{v}_n\}$, $\{\xi_n^*\}$ such that

$$\mathbf{v}_n \in \partial^p(\theta|_C)(x_n') \cap \partial F^0 \ , \tag{55}$$

$$\xi_n^* \in \partial\rho_F(z-x_n'') \tag{56}$$

and

$$\|x_n'-x_n\| + \|x_n''-x_n\| \to 0 \ , \tag{57}$$

$$\|\mathbf{v}_n-\xi_n^*\| \to 0 \tag{58}$$

as $n\to\infty$.

Proof. Let us take an arbitrary sequence $\varepsilon_n\to 0^+$ with

$$\rho_F(z-x_n)+\theta(x_n) \le \hat{u}(z)+\varepsilon_n$$

and applying the *Ekeland Variational Principle* [21, Corollary 11] choose a sequence $\{y_n\}\subset C$ such that

$$\rho_F(z-y_n)+\theta(y_n) \le \hat{u}(z)+\varepsilon_n \ ; \qquad \|x_n-y_n\| \le \sqrt{\varepsilon_n} \tag{59}$$

and

$$\rho_F(z-y_n)+\theta(y_n) \le \rho_F(z-y)+\theta(y)+\sqrt{\varepsilon_n}\,\|y-y_n\| \tag{60}$$

for all $y\in C$, $n=1,2,\ldots$.

The inequality (60) means that y_n minimizes the functional

$$F_n(y) := \rho_F(z-y)+\theta|_C(y)+\sqrt{\varepsilon_n}\,\|y-y_n\|$$

on H. Then the *necessary condition of optimality* in proximal form yields $0\in\partial^p F_n(y_n)$. Decomposing the proximal subdifferential in accordance with the *fuzzy sum rule* (see Theorem 8.3 [14, p. 56]) we find sequences $\{x_n'\}\subset C$ and $\{x_n''\}\subset H$, $\|x_n'-y_n\|\le\sqrt{\varepsilon_n}$, $\|x_n''-y_n\|\le\sqrt{\varepsilon_n}$, such that

$$\begin{aligned} 0 &\in -\partial\rho_F(z-x_n'') + \sqrt{\varepsilon_n}\frac{x_n''-y_n}{\|x_n''-y_n\|} + \partial^p(\theta|_C)(x_n') + \sqrt{\varepsilon_n}\overline{\mathbf{B}} \\ &\subset -\partial\rho_F(z-x_n'') + \partial^p(\theta|_C)(x_n') + 2\sqrt{\varepsilon_n}\overline{\mathbf{B}} \ . \end{aligned}$$

Hence, there exist vectors $\mathbf{v}_n'\in\partial^p(\theta|_C)(x_n')$ and $\xi_n^*\in\partial\rho_F(z-x_n'')$ with

$$\|\mathbf{v}_n'-\xi_n^*\| \le 2\sqrt{\varepsilon_n} \ . \tag{61}$$

It follows from (59), (47) and (9) that $\{x_n'\}$ is a minimizing sequence of $x\mapsto\rho_F(z-x)+\theta(x)$ on C. Indeed,

$$\begin{aligned}\rho_F\left(z-x_n'\right)+\theta\left(x_n'\right) &\le \rho_F\left(z-y_n\right)+\theta\left(y_n\right)\\ &\quad+\rho_F\left(y_n-x_n'\right)+\gamma\left\|F^0\right\|\left\|y_n-x_n'\right\|\\ &\le \hat{u}\left(z\right)+\left(\gamma+1\right)\left\|F^0\right\|\sqrt{\varepsilon_n}+\varepsilon_n\ .\end{aligned}$$

By using the hypothesis $(\hat{\mathbf{H}})$ we deduce that $x_n'\in\partial^\theta C$ since otherwise $\mathbf{v}_n'\in\Gamma\left(x_n'\right)\subset\gamma F^0$ contradicting the choice of ξ_n^* because $\left\|\mathbf{v}_n'-\xi_n^*\right\|\to 0$ [see (61)] and $\xi_n^*\in\partial F^0$ [see (12)]. So, by (54) $\mathbf{v}_n'$ can be decomposed in a sum $\mathbf{w}_n+\mathbf{u}_n$ where $\mathbf{w}_n\in\Gamma\left(x_n'\right)$ and $\mathbf{u}_n\in\mathbf{N}_C^\theta\left(x_n'\right)$ with $\mathbf{u}_n\neq 0$. Finally, let us define the vectors

$$\mathbf{v}_n:=\frac{\mathbf{u}_n}{\rho_{F^0-\mathbf{w}_n}\left(\mathbf{u}_n\right)}+\mathbf{w}_n\in\mathbf{N}_C^\theta\left(x_n'\right)+\Gamma\left(x_n'\right)=\partial^p\left(\theta\left|{}_C\right.\right)\left(x_n'\right)\ . \tag{62}$$

Obviously, $\mathbf{v}_n\in\partial F^0$ implying together with (62) the property (55). Furthermore, applying (9), the hypothesis $(\hat{\mathbf{H}})$ and the relations (48), (51) with $\Gamma\left(x_n'\right)$ in the place of $\partial^c\theta\left(x_n'\right)$ we successively obtain

$$\begin{aligned}\left\|\mathbf{v}_n'-\mathbf{v}_n\right\| &= \frac{\left\|\mathbf{u}_n\right\|}{\rho_{F^0-\mathbf{w}_n}\left(\mathbf{u}_n\right)}\left|\rho_{F^0-\mathbf{w}_n}\left(\mathbf{v}_n'-\mathbf{w}_n\right)-\rho_{F^0-\mathbf{w}_n}\left(\xi_n^*-\mathbf{w}_n\right)\right|\\ &\le \left\|F^0-\mathbf{w}_n\right\|\left\|\left(F^0-\mathbf{w}_n\right)^0\right\|\left\|\mathbf{v}_n'-\xi_n^*\right\|\\ &\le \frac{1+\gamma}{1-\gamma}\left\|F\right\|\left\|F^0\right\|\left\|\mathbf{v}_n'-\xi_n^*\right\|\ ,\end{aligned} \tag{63}$$

$n=1,2,\ldots$. Combining now (63) and (61) we arrive at (58), and the lemma is proved. □

6 Existence, Uniqueness and Stability of Minimizers

Given $x_0\in\partial C$ let us set now the local assumptions, under which the results on well-posedness and regularity near x_0 hold:

$(\mathbf{H}_1\left(x_0\right))$ the mapping $x\mapsto\mathfrak{J}_F\left(\partial^p\left(\theta\left|{}_C\right.\right)\left(x\right)\cap\partial F^0\right)$ is **single-valued** and **lipschitzean** (with Lipschitz constant $L=L\left(x_0\right)>0$) on the set

$$C_\delta\left(x_0\right):=\left\{x\in\partial^\theta C:\left\|x-x_0\right\|\le\delta\right\}\ ,\qquad\delta>0\ ;$$

$(\mathbf{H}_2\left(x_0\right))$ F is **uniformly rotund** w.r.t. the set

$$\mathfrak{U}_\delta\left(x_0\right):=\bigcup_{x\in C_\delta\left(x_0\right)}\partial^p\left(\theta\left|{}_C\right.\right)\left(x\right)\cap\partial F^0. \tag{64}$$

Observe that (like the case $\theta\equiv 0$) in finite dimensions the hypothesis $(\mathbf{H}_2\left(x_0\right))$ holds automatically if one requires just the strict convexity of F w.r.t. each vector $\xi^*\in\mathfrak{U}_\delta\left(x_0\right)$ that trivially follows from $(\mathbf{H}_1\left(x_0\right))$. So, the assumption $(\mathbf{H}_2\left(x_0\right))$ can be required only in an infinite dimensional space H, while in $\mathbb{R}^n$ it is superfluous.

Theorem 3. *Under the standing assumptions* $(\mathbf{H})$ *and* $(\hat{\mathbf{H}})$ *let us fix* $x_0 \in \partial C$ *and assume that the local hypotheses* $(\mathbf{H}_1(x_0))$ *and* $(\mathbf{H}_2(x_0))$ *are fulfilled. Then there exists a neighbourhood* $\mathcal{U}(x_0)$ *of* x_0 *where the mapping* $\pi_C^{F,\theta}(\cdot)$ *is single-valued and locally lipschitzean.*

Proof. Due to the choice of $\gamma > 0$ (see $(\mathbf{H})$) we can assume $\delta > 0$ from the hypotheses $(\mathbf{H}_1(x_0)) - (\mathbf{H}_2(x_0))$ so small that

$$\delta\gamma\left\|F^0\right\| < \frac{1-\gamma\|F\|\left\|F^0\right\|}{L} \ .$$

Now, using the upper semicontinuity of $\hat{u}(\cdot)$ and the equality $\hat{u}(x_0) = \theta(x_0)$ (see Theorem 1), we can define the (open) neighbourhood

$$\mathcal{U}(x_0) := \left\{ x \in H : \|x - x_0\| < \frac{\left(1-\gamma\|F\|\left\|F^0\right\|\right)\delta}{2\|F\|\left\|F^0\right\|} \ , \right.$$
$$\left. \hat{u}(x) < \theta(x_0) + \frac{1-\gamma\|F\|\left\|F^0\right\|}{L} - \delta\gamma\left\|F^0\right\| \right\} \ . \tag{65}$$

Fix $z \in \mathcal{U}(x_0) \setminus C$ and a minimizing sequence $\{x_n\} \subset C$ for the function $x \mapsto \rho_F(z-x) + \theta(x)$ on C. By Lemma 1 let us choose another minimizing sequence $\{x_n'\} \subset \partial^\theta C$ and sequences $\{x_n''\} \subset H$, $\mathbf{v}_n \in \partial^p(\theta|_C)(x_n') \cap \partial F^0$, $\xi_n^* \in \partial\rho_F(z - x_n'')$ satisfying (57) and (58).

Let us show first that $x_n' \in C_\delta(x_0)$ for $n \geq 1$ large enough. To this end we denote by

$$0 < \varepsilon_n := \rho_F\left(z - x_n'\right) + \theta\left(x_n'\right) - \hat{u}(z) \to 0+ \tag{66}$$

and using the inequalities (9) and (47) successively write

$$\begin{aligned} \rho_F\left(x_0 - x_n'\right) &\leq \rho_F(x_0 - z) + \rho_F\left(z - x_n'\right) \\ &= \rho_F(x_0 - z) + \hat{u}(z) - \theta\left(x_n'\right) + \varepsilon_n \\ &\leq \rho_F(x_0 - z) + \rho_F(z - x_0) + \theta(x_0) - \theta\left(x_n'\right) + \varepsilon_n \\ &\leq 2\left\|F^0\right\|\|z - x_0\| + \gamma\left\|F^0\right\|\left\|x_n' - x_0\right\| + \varepsilon_n \ . \end{aligned} \tag{67}$$

Hence, again by (9) we have

$$\left(\frac{1}{\|F\|} - \gamma\left\|F^0\right\|\right)\left\|x_n' - x_0\right\| \leq \varepsilon_n + 2\left\|F^0\right\|\|z - x_0\| \ ,$$

and by the choice of z [see (65)] conclude that $\|x_n' - x_0\| < \delta$.

Then, due to one of the characterizations of the convex subdifferential [see (12)] $\xi_n^* \in \mathbf{N}_F(\xi_n) \cap \partial F^0$, where

$$\xi_n := \frac{z - x_n''}{\rho_F(z - x_n'')} \ ,$$

and, consequently [see (10)], $\xi_n \in \mathfrak{J}_F(\xi_n^*)$, $n = 1, 2, \ldots$.

Set now

$$\beta_n := \max\left\{\varepsilon_n \ , \quad \left\|x_n' - x_n\right\| + \left\|x_n'' - x_n\right\| \ , \quad \left\|\mathfrak{J}_F(\mathbf{v}_n) - \xi_n\right\|\right\} \tag{68}$$

and deduce from (57), (58), the hypothesis $(\mathbf{H}_2(x_0))$ and [26, Proposition 2.1] that $\beta_n \to 0+$ as $n \to \infty$.

Taking into account the representation $x_n'' = z - \xi_n \rho_F(z - x_n'')$, for given $m, n \geq 1$ we write

$$\left\|x_n'' - x_m''\right\| \leq \rho_F\left(z - x_n''\right)\left\|\xi_n - \xi_m\right\| + \left\|\xi_m\right\|\left|\rho_F\left(z - x_n''\right) - \rho_F\left(z - x_m''\right)\right| \ . \tag{69}$$

Let us estimate each term of the latter inequality. First, by the definition of ε_n [see (66)], (9), (47) and (68) we obtain that

$$\begin{aligned}\rho_F\left(z - x_n''\right) &\leq \left\|F^0\right\|\left\|x_n'' - x_n'\right\| + \hat{u}(z) - \theta\left(x_n'\right) + \varepsilon_n \\ &\leq \left(\left\|F^0\right\| + 1\right)\beta_n + \hat{u}(z) - \theta(x_0) + \gamma\delta\left\|F^0\right\|\end{aligned} \tag{70}$$

and that

$$\begin{aligned}&\left|\rho_F\left(z - x_n''\right) - \rho_F\left(z - x_m''\right)\right| \\ &\leq \left|\rho_F\left(z - x_n'\right) - \rho_F\left(z - x_m'\right)\right| + \left\|F^0\right\|(\beta_n + \beta_m) \\ &\leq \left|\rho_F\left(z - x_n'\right) + \theta\left(x_n'\right) - \hat{u}(z)\right| + \left\|F^0\right\|(\beta_n + \beta_m) \\ &\quad + \left|\rho_F\left(z - x_m'\right) + \theta\left(x_m'\right) - \hat{u}(z)\right| + \left|\theta\left(x_n'\right) - \theta\left(x_m'\right)\right| \\ &\leq \left(\left\|F^0\right\| + 1\right)(\beta_n + \beta_m) + \gamma\left\|F^0\right\|\left\|x_n' - x_m'\right\| \ .\end{aligned} \tag{71}$$

Furthermore, since $\mathbf{v}_n \in \partial^p(\theta|_C)(x_n') \cap \partial F^0$, applying the main hypothesis $(\mathbf{H}_1(x_0))$ we have

$$\begin{aligned}\left\|\xi_n - \xi_m\right\| &\leq \left\|\mathfrak{J}_F(\mathbf{v}_n) - \mathfrak{J}_F(\mathbf{v}_m)\right\| + \beta_n + \beta_m \\ &\leq L\left\|x_n' - x_m'\right\| + \beta_n + \beta_m \ .\end{aligned} \tag{72}$$

After substituting (70)–(72) into (69) and joining all the infinitesimal constants we finally arrive at:

$$\begin{aligned}\left\|x_n' - x_m'\right\| \leq \left\|x_n'' - x_m''\right\| + \beta_n + \beta_m \leq \big[&L\left(\hat{u}(z) - \theta(x_0) + \gamma\delta\left\|F^0\right\|\right) \\ &+ \gamma\|F\|\left\|F^0\right\|\big]\left\|x_n' - x_m'\right\| + \mu_{n,m}\end{aligned} \tag{73}$$

where $\mu_{n,m} \to 0+$ as $n, m \to \infty$. Since

$$\gamma\|F\|\left\|F^0\right\| + L\left(\hat{u}(z) - \theta(x_0) + \gamma\delta\left\|F^0\right\|\right) < 1$$

by the choice of z [see (65)], we conclude from (73) that $\{x_n'\}$ (and, consequently, $\{x_n\}$) is a Cauchy sequence in H.

In fact, we have proved that each minimizing sequence of the function $x \mapsto \rho_F(z-x)+\theta(x)$ on C is a Cauchy sequence. Therefore, its limit $\bar{x}$ is the (unique) element of the set of minimizers $\pi_C^{F,\theta}(z)$. Moreover, by using the same argument we can prove the continuity of the mapping $z \mapsto \pi_C^{F,\theta}(z)$ on $\mathcal{U}(x_0)$. Indeed, taking a sequence $\{z_n\} \subset \mathcal{U}(x_0)$, $z_n \to z \in \mathcal{U}(x_0)$, and denoting by $\bar{z}_n$ the unique element of $\pi_C^{F,\theta}(z_n)$, we observe that $\{\bar{z}_n\}$ is a minimizing sequence of $x \mapsto \rho_F(z-x)+\theta(x)$ on C. Indeed, we have

$$\begin{aligned}\hat{u}(z) &\leq \rho_F(z-\bar{z}_n)+\theta(\bar{z}_n) \leq \rho_F(z-z_n)+\rho_F(z_n-\bar{z}_n)+\theta(\bar{z}_n)\\ &\leq \hat{u}(z)+2\left\|F^0\right\|\|z-z_n\| \ ,\end{aligned}$$

where the latter inequality follows from the lipschitzeanity of the function $\hat{u}(\cdot)$. So, $\{\bar{z}_n\}$ converges to the (unique) element of $\pi_C^{F,\theta}(z)$.

In the second part of the proof we show that the single-valued function $\pi_C^{F,\theta}(\cdot)$ is actually Lipschitz continuous on $\mathcal{U}(x_0)$. To do this fix an arbitrary point $x \in \mathcal{U}(x_0)$ and choose $\tau > 0$ and $0 < \varepsilon \leq \frac{\tau}{2\|F^0\|}$ so small that

$$\hat{u}(x)-\theta(x_0)+\gamma\delta\left\|F^0\right\|+\tau < \frac{1-\gamma\|F\|\left\|F^0\right\|}{L} \tag{74}$$

and

$$\frac{2\|F\|\left\|F^0\right\|}{1-\gamma\|F\|\|F^0\|}(\|x-x_0\|+\varepsilon) < \delta \ . \tag{75}$$

Let us take $z_1, z_2 \in \mathcal{U}(x_0)$, $\|z_i - x\| < \varepsilon$, $i = 1,2$, and assume first that both (different) points z_1 and z_2 are out of C. Setting $\beta := \|z_1 - z_2\|/2 > 0$, in virtue of the hypothesis $(\mathbf{H}_2(x_0))$ and [26, Proposition 2.1 (ii)] we find $0 < \nu \leq \varepsilon \wedge \beta$ such that

$$\|\mathfrak{J}_F(\eta^*)-\xi\| \leq \beta$$

whenever $\xi \in \mathfrak{J}_F(\xi^*)$, $\xi^* \in \partial F^0$ and $\eta^* \in \mathfrak{U}_\delta(x_0)$ with $\|\xi^* - \eta^*\| \leq \nu$. Without loss of generality one may suppose that

$$\nu + \frac{2\|F\|\left\|F^0\right\|}{1-\gamma\|F\|\|F^0\|}(\|x-x_0\|+\varepsilon) < \delta \tag{76}$$

and that $\left(z_i+\nu\overline{B}\right)\cap C = \varnothing$, $i = 1,2$. Set also $\bar{z}_i := \pi_C^{F,\theta}(z_i)$. Now we apply the tools used for proving Lemma 1 but without recurrence to the Ekeland Principle (because the exact minimizer is already known). Namely, $\bar{z}_i$ minimizes the function

$$F_i(z) := \rho_F(z_i - z) + \theta|_C(z)$$

on H. Therefore $0 \in \partial^p F_i(\bar{z}_i)$, $i = 1,2$. By the fuzzy sum rule similarly as in the proof of Lemma 1 we find points $z_i' \in \partial^\theta C$ and $z_i'' \in H$ both close to $\bar{z}_i$ (say $\|z_i' - \bar{z}_i\| + \|z_i'' - \bar{z}_i\| \leq \nu$) and vectors $\mathbf{v}_i \in \partial^p(\theta|_C)(z_i') \cap \partial F^0$, $\xi_i^* \in \partial\rho_F(z_i - z_i'')$ such that

$$\|\mathbf{v}_i - \xi_i^*\| \leq \nu \ . \tag{77}$$

Let us show now that $z_i' \in C_\delta(x_0)$. Similarly as in (67) we have:

$$\begin{aligned}\rho_F(x_0 - \bar{z}_i) &\leq \rho_F(z_i - \bar{z}_i) + \rho_F(x_0 - z_i) \\ &= \hat{u}(z_i) - \theta(\bar{z}_i) + \rho_F(x_0 - z_i) \\ &\leq \rho_F(z_i - x_0) + \theta(x_0) - \theta(\bar{z}_i) + \|F^0\| \|z_i - x_0\| \\ &\leq 2\|F^0\| \|z_i - x_0\| + \gamma\|F^0\| \|\bar{z}_i - x_0\| \ ,\end{aligned}$$

and, hence,

$$\frac{1 - \gamma\|F^0\| \|F\|}{\|F\|} \|\bar{z}_i - x_0\| \leq 2\|F^0\| \|z_i - x_0\| \ .$$

Recalling (76), from the latter inequality we obtain

$$\begin{aligned}\|z_i' - x_0\| &\leq \|z_i' - \bar{z}_i\| + \|\bar{z}_i - x_0\| \\ &\leq \nu + \frac{2\|F^0\| \|F\|}{1 - \gamma\|F^0\| \|F\|} (\varepsilon + \|x - x_0\|) < \delta \ .\end{aligned}$$

Thus $z_i' \in C_\delta(x_0)$ and $\mathbf{v}_i \in \mathfrak{U}_\delta(x_0)$ [see (64)].

Setting now $\xi_i := \frac{z_i - z_i''}{\rho_F(z_i - z_i'')}$ we see that $\xi_i \in \mathfrak{J}_F(\xi_i^*)$, and it follows from (77) and from the choice of $\nu > 0$ that

$$\|\mathfrak{J}_F(\mathbf{v}_i) - \xi_i\| \leq \beta \ . \tag{78}$$

Joining together the inequalities (78) for $i = 1, 2$ and using the hypothesis $(\mathbf{H}_1(x_0))$, we have

$$\begin{aligned}\|\xi_1 - \xi_2\| &\leq 2\beta + L\|z_1' - z_2'\| \\ &\leq 2\beta + L(2\nu + \|\bar{z}_1 - \bar{z}_2\|) \ .\end{aligned} \tag{79}$$

In order to estimate the distance $\|\bar{z}_1 - \bar{z}_2\|$ we use first the proximity of each minimizer $\bar{z}_i$ to $z_i'' = z_i - \xi_i \rho_F(z_i - z_i'')$. Namely,

$$\begin{aligned}\|\bar{z}_1 - \bar{z}_2\| &\leq 2\nu + \|z_1'' - z_2''\| \\ &\leq 2\nu + \|z_1 - z_2\| + \|\xi_1 \rho_F(z_1 - z_1'') - \xi_2 \rho_F(z_2 - z_2'')\| \\ &\leq 2\nu + \|z_1 - z_2\| + \rho_F(z_1 - z_1'') \|\xi_1 - \xi_2\| \\ &\quad + \|F\| |\rho_F(z_1 - z_1'') - \rho_F(z_2 - z_2'')| \ .\end{aligned} \tag{80}$$

On the other hand, similarly to (70) and (71) we successively have

$$\begin{aligned}\rho_F(z_1 - z_1'') &\leq \rho_F(z_1 - \bar{z}_1) + \|F^0\| \|\bar{z}_1 - z_1''\| \\ &\leq \|F^0\| \nu + \hat{u}(z_1) - \theta(\bar{z}_1)\end{aligned}$$

$$\begin{aligned} &\leq \left\|F^0\right\|(\nu+\|z_1-x\|)+\hat{u}(x)-\theta(x_0)+\gamma\left\|F^0\right\|\|\bar{z}_1-x_0\| \\ &\leq \hat{u}(x)-\theta(x_0)+\gamma\left\|F^0\right\|\delta+\tau \end{aligned} \tag{81}$$

(recall that $\nu \leq \varepsilon \leq \frac{\tau}{2\left\|F^0\right\|}$), and

$$\begin{aligned} &\left|\rho_F\left(z_1-z_1''\right)-\rho_F\left(z_2-z_2''\right)\right| \\ &\leq \left|\rho_F(z_1-\bar{z}_1)-\rho_F(z_2-\bar{z}_2)\right|+2\left\|F^0\right\|\nu \\ &\leq |\hat{u}(z_1)-\hat{u}(z_2)|+|\theta(\bar{z}_1)-\theta(\bar{z}_2)|+2\left\|F^0\right\|\nu \\ &\leq \left\|F^0\right\|(2\nu+\|z_1-z_2\|)+\gamma\left\|F^0\right\|\|\bar{z}_1-\bar{z}_2\| \ . \end{aligned} \tag{82}$$

Taking into account the inequalities (81), (79), (82) and recalling that $\nu \leq \beta = \|z_1-z_2\|/2$ we deduce from (80):

$$\left[1-L\left(\hat{u}(x)-\theta(x_0)+\gamma\left\|F^0\right\|\delta+\tau\right)-\gamma\|F\|\left\|F^0\right\|\right]\|\bar{z}_1-\bar{z}_2\| \leq K\|z_1-z_2\|$$

where

$$K=K(x):=2\left(1+\|F\|\left\|F^0\right\|\right)+(L+1)\left(\hat{u}(x)-\theta(x_0)+\gamma\left\|F^0\right\|\delta+\tau\right)>0 \ .$$

Finally,

$$\mu=\mu(x):=1-L\left(\hat{u}(x)-\theta(x_0)+\gamma\left\|F^0\right\|\delta+\tau\right)-\gamma\|F\|\left\|F^0\right\|>0$$

by (74), and we arrive at the (local) Lipschitz inequality

$$\left\|\pi_C^{F,\theta}(z_1)-\pi_C^{F,\theta}(z_2)\right\| \leq \frac{K(x)}{\mu(x)}\|z_1-z_2\| \ . \tag{83}$$

In the case when one of the points z_i (say z_2) belongs to C, we obviously have $\pi_C^{F,\theta}(z_2)=z_2$ and

$$\begin{aligned} \|\bar{z}_1-\bar{z}_2\| &= \|\bar{z}_1-z_2\| \leq \|F\|\rho_F(z_1-\bar{z}_1)+\|z_1-z_2\| \\ &= \|F\|(\hat{u}(z_1)-\theta(\bar{z}_1))+\|z_1-z_2\| \\ &= \|F\|(\hat{u}(z_1)-\hat{u}(z_2))+\|F\|(\theta(z_2)-\theta(\bar{z}_1))+\|z_1-z_2\| \\ &\leq \left(\|F\|\left\|F^0\right\|+1\right)\|z_1-z_2\|+\gamma\|F\|\left\|F^0\right\|\|\bar{z}_1-\bar{z}_2\| \ . \end{aligned}$$

Hence, (83) holds as well with

$$K(x):=\|F\|\left\|F^0\right\|+1$$

and

$$\mu(x):=1-\gamma\|F\|\left\|F^0\right\|>0 \ .$$

Theorem is completely proved. □

Remark 4. Notice that the Lipschitz constant of the mapping $\pi_C^{F,\theta}(\cdot)$ depends essentially on the distance from the boundary of the neighbourhood $\mathcal{U}(x_0)$ controlled by the parameter τ. In fact, $\mathcal{U}(x_0)$ is defined by means of two inequalities: the first one gives direct proximity to the boundary point x_0, while the second derives from the upper semicontinuity of the marginal function $\hat{u}(\cdot)$ at x_0. Thus, the Lipschitz constant of $\pi_C^{F,\theta}(\cdot)$ depending on $x \in \mathcal{U}(x_0)$ tends to $+\infty$ ($\mu(x) \to 0$) whenever the strict upper semicontinuity inequality

$$\hat{u}(x) < \hat{u}(x_0) + \frac{1-\gamma\|F\|\left\|F^0\right\|}{L} - \delta\gamma\left\|F^0\right\|$$

tends to become an equality, i.e., the value of the function $\hat{u}(\cdot)$ at x is most distant from its value at x_0. This generalizes the well-known property of the metric projections onto prox-regular sets (see, e.g., [8]).

Remark 5. If the conditions $(\mathbf{H}_1(x_0))$ and $(\mathbf{H}_2(x_0))$ are fulfilled at each point $x_0 \in \partial C$, then the marginal mapping $x \mapsto \pi_C^{F,\theta}(x)$ is single-valued and locally lipschitzean on the open neighbourhood

$$\mathfrak{A}(C) := \operatorname{int} C \cup \bigcup_{x_0 \in \partial C} \mathcal{U}(x_0)$$

of the target set.

7 Regularity of the Value Function

At the beginning of this section we study the Clarke (and lower) regularity of the function $\hat{u}(\cdot)$ at a given point $\hat{x}$ out of the target set under an a priori assumption that for each x near $\hat{x}$ the infimum in (7) is attended at an unique point, and a kind of stability of the minimizer takes place. Furthermore, we give a representation formula for the Clarke (Fréchet or Mordukhovich) subdifferential of $\hat{u}(\cdot)$ at $\hat{x}$ in terms of the respective constructions for F, $\theta(\cdot)$ and C. A similar result was obtained in [26] in the case $\theta \equiv 0$ (see also [16]).

For one step of the proof we need the following simple observation.

Lemma 2. *Fix $x \notin C$ such that $\pi_C^{F,\theta}(x)$ is a singleton (say $\bar{x}$) and denote by $\overline{\xi} := \frac{x-\bar{x}}{\rho_F(x-\bar{x})}$. Then for all $0 \le t \le \rho_F(x-\bar{x})$ the inequality*

$$\hat{u}\left(x - t\overline{\xi}\right) \le \hat{u}(x) - t \tag{84}$$

holds.

Proof. Setting $y_t := x - t\overline{\xi}$ for $0 \le t \le \rho_F(x-\bar{x})$, we have

$$\begin{aligned}\hat{u}(y_t) &\le \rho_F(y_t-\bar{x}) + \theta(\bar{x}) = \rho_F\left((x-\bar{x})\left(1-\frac{t}{\rho_F(x-\bar{x})}\right)\right) + \theta(\bar{x}) \\ &= \rho_F(x-\bar{x}) - t + \theta(\bar{x}) = \hat{u}(x) - t \ ,\end{aligned}$$

and (84) is proved. □

Theorem 4. *Let us fix $\hat{x} \notin C$ and assume that the mapping $x \mapsto \pi_C^{F,\theta}(x)$ is single-valued in a neighbourhood $U(\hat{x})$ of $\hat{x}$ and such that*

$$\lim_{r\to 0+} \frac{\omega(\hat{x};r)}{\sqrt{r}} = 0 \ , \tag{85}$$

where $\omega(\hat{x};r)$ is the modulus of continuity of $\pi_C^{F,\theta}(\cdot)$ at the point $\hat{x}$, namely,

$$\omega(\hat{x};r) := \sup\left\{\left\|\pi_C^{F,\theta}(x) - \pi_C^{F,\theta}(\hat{x})\right\| : \|x-\hat{x}\| \le r\right\} \ .$$

Suppose also that the restriction $\theta|_C$ is proximally regular at $\overline{x} := \pi_C^{F,\theta}(\hat{x})$. Then the function $\hat{u}(\cdot)$ is Clarke (and, hence, lower) regular at $\hat{x}$. Furthermore, the following formula takes place:

$$\partial^c \hat{u}(\hat{x}) = \partial^l \hat{u}(\hat{x}) = \partial^- \hat{u}(\hat{x}) = \partial\rho_F\left(\hat{x}-\overline{x}\right) \cap \partial^-\left(\theta|_C\right)\left(\overline{x}\right) \neq \varnothing \ . \tag{86}$$

Proof. Our proof is divided into several steps.

Step 1. Let us show first that $\partial^- \hat{u}(x) \subset \partial\rho_F(x-\bar{x})$ for each $x \in U(\hat{x})$ where $\bar{x} := \pi_C^{F,\theta}(x)$. To this end we use the representation of the subdifferential $\partial\rho_F(x-\bar{x})$ via the normal cone to F [see (12)]. Since the function $\hat{u}(\cdot)$ satisfies the slope condition (26), by (20) and Remark 2 we have $\partial^- \hat{u}(x) \subset \partial^c \hat{u}(x) \subset F^0$. On the other hand, by Theorem 1 $\hat{u}(\cdot)$ is the viscosity solution of (5). So, in particular,

$$\rho_{F^0}(p) \ge 1 \tag{87}$$

for each $p \in \partial^- \hat{u}(x)$. Thus $\partial^- \hat{u}(x) \subset \partial F^0$.

Besides that, given $p \in \partial^- \hat{u}(x)$ let us choose $\varepsilon > 0$ and $\delta > 0$ such that

$$\hat{u}(y) - \hat{u}(x) - \langle p, y-x\rangle \ge -\varepsilon\|x-y\|$$

for all y, $\|y-x\| \le \delta$. In particular, setting $y = x - t\bar{\xi}$, where

$$\bar{\xi} := \frac{x-\bar{x}}{\rho_F(x-\bar{x})} \ ,$$

and applying Lemma 2 we have that

$$\begin{aligned}\hat{u}(x) - t &\ge \hat{u}\left(x - t\bar{\xi}\right) \\ &\ge \hat{u}(x) - t\left\langle p, \bar{\xi}\right\rangle - \varepsilon t\left\|\bar{\xi}\right\|\end{aligned}$$

for sufficiently small $t>0$. Hence, letting $\varepsilon\to 0+$ we arrive at $1\le\left\langle p,\bar{\xi}\right\rangle$ and, consequently,

$$\left\langle p,\xi-\bar{\xi}\right\rangle\le 0$$

whenever $\xi\in F$, i.e., $p\in\mathbf{N}_F\left(\bar{\xi}\right)$. So, due to (12) we conclude that $\partial^-\hat{u}(x)\subset\partial\rho_F(x-\bar{x})$.

Step 2. We prove the inclusion $\partial^p\hat{u}(x)\subset\partial^p(\theta|_C)(\bar{x})$, $x\in U(\hat{x})$. Given $p\in\partial^p\hat{u}(x)$ let us choose $\eta>0$ and $\sigma>0$ such that

$$\hat{u}(y)-\hat{u}(x)-\langle p,y-x\rangle\ge-\sigma\|y-x\|^2$$

for all y, $\|y-x\|\le\eta$. In particular, for $y=z-\bar{x}+x$, where $z\in C$, $\|z-\bar{x}\|\le\eta$, we have:

$$\begin{aligned}-\sigma\|z-\bar{x}\|^2&\le\hat{u}(z-\bar{x}+x)-\hat{u}(x)-\langle p,z-\bar{x}\rangle\\&\le\rho_F(x-\bar{x})+\theta(z)-\rho_F(x-\bar{x})-\theta(\bar{x})-\langle p,z-\bar{x}\rangle\ ,\end{aligned}$$

or, in other words,

$$-\sigma\|z-\bar{x}\|^2\le\theta|_C(z)-\theta|_C(\bar{x})-\langle p,z-\bar{x}\rangle$$

for all $z\in H$ with $\|z-\bar{x}\|\le\eta$ that means $p\in\partial^p(\theta|_C)(\bar{x})$.

Thus, joining Steps 1 and 2 we conclude that

$$\partial^p\hat{u}(x)\subset\partial\rho_F(x-\bar{x})\cap\partial^p(\theta|_C)(\bar{x})\tag{88}$$

for each x close to $\hat{x}$.

Step 3. Let us prove now a kind of opposite inclusion

$$\partial\rho_F(x-\bar{x})\cap\partial^p(\theta|_C)(\bar{x})\subset\partial^-\hat{u}(x)\tag{89}$$

but only at the point $x=\hat{x}$. To this end fix p from the left-hand side of (89). Then, in particular,

$$\rho_F(\xi)\ge\rho_F\left(\hat{x}-\bar{\hat{x}}\right)+\left\langle p,\xi-\hat{x}+\bar{\hat{x}}\right\rangle\tag{90}$$

for all $\xi\in H$. Given $x\notin C$ sufficiently close to $\hat{x}$ let us set $\xi=x-\bar{x}$ in (90) and rewrite the latter inequality as

$$\rho_F(x-\bar{x})-\rho_F\left(\hat{x}-\bar{\hat{x}}\right)-\langle p,x-\hat{x}\rangle\ge-\left\langle p,\bar{x}-\bar{\hat{x}}\right\rangle\ .\tag{91}$$

On the other hand, one can choose $\eta>0$ and $\sigma>0$ such that

$$\theta(z)\ge\theta\left(\bar{\hat{x}}\right)+\left\langle p,z-\bar{\hat{x}}\right\rangle-\sigma\left\|z-\bar{\hat{x}}\right\|^2\tag{92}$$

whenever $z\in C$ with $\left\|z-\bar{\hat{x}}\right\|\le\eta$. Due to the continuity of the mapping $\pi_C^{F,\theta}(\cdot)$ at $\hat{x}$ the inequality (92) holds for $z=\bar{x}$ with x enough close to $\hat{x}$. Combining this with both (91) and the condition (85), we successively obtain:

$$\begin{aligned}
&\liminf_{x\to\hat{x}} \frac{\hat{u}(x)-\hat{u}(\hat{x})-\langle p,x-\hat{x}\rangle}{\|x-\hat{x}\|} \\
&= \liminf_{x\to\hat{x}} \frac{\rho_F(x-\bar{x})+\theta(\bar{x})-\rho_F(\hat{x}-\bar{\hat{x}})-\theta(\bar{\hat{x}})-\langle p,x-\hat{x}\rangle}{\|x-\hat{x}\|} \\
&\geq \liminf_{x\to\hat{x}} \frac{-\langle p,\bar{x}-\bar{\hat{x}}\rangle+\theta(\bar{x})-\theta(\bar{\hat{x}})}{\|x-\hat{x}\|} \\
&\geq \liminf_{x\to\hat{x}} \frac{-\sigma\|\bar{x}-\bar{\hat{x}}\|^2}{\|x-\hat{x}\|} \geq -\sigma\left(\lim_{r\to 0+}\frac{\omega(\hat{x},r)}{\sqrt{r}}\right)^2 = 0 \ ,
\end{aligned}$$

i.e., $p\in\partial^-\hat{u}(\hat{x})$.

Taking into account the proximal regularity of $\theta|_C$ at the point $\bar{\hat{x}}$ and the inclusions (20) we deduce from (89) that

$$\partial\rho_F(\hat{x}-\bar{\hat{x}})\cap\partial^-(\theta|_C)(\bar{\hat{x}})\subset\partial^-\hat{u}(\hat{x})\subset\partial^c\hat{u}(\hat{x}) \ . \tag{93}$$

Step 4. To complete the proof recall the representation formula (22) for the Clarke subdifferential through the proximal subgradients in neighbour points and apply the inclusion (88). So, $\partial^c\hat{u}(\hat{x})$ is contained in the closed convex hull of the set of all weak limits of sequences $\zeta_i\in\partial\rho_F(x_i-\bar{x}_i)\cap\partial^p(\theta|_C)(\bar{x}_i)$ such that $x_i\to\hat{x}$ as $i\to\infty$. Furthermore, since $\bar{x}_i\to\bar{\hat{x}}$, the subdifferential of a convex function *has $s\times w$ sequentially closed graph* and

$$w-\limsup_{x\to\bar{\hat{x}},\,x\in C}\partial^p(\theta|_C)(x)=\partial^l(\theta|_C)(\bar{\hat{x}})$$

(see [31, p. 240]), we have

$$\begin{aligned}
\partial^c\hat{u}(\hat{x}) &\subset \overline{\mathrm{co}}\left(\partial\rho_F(\hat{x}-\bar{\hat{x}})\cap\partial^l(\theta|_C)(\bar{\hat{x}})\right) \\
&= \partial\rho_F(\hat{x}-\bar{\hat{x}})\cap\partial^-(\theta|_C)(\bar{\hat{x}}) \ ,
\end{aligned} \tag{94}$$

where the latter equality follows from the lower regularity of the function $\theta|_C$ (it is a consequence of the proximal regularity) and from the convexity of the Fréchet subdifferential $\partial^-(\theta|_C)(\bar{\hat{x}})$. Hence, in particular, the right-hand side of (94) is nonempty because $\hat{u}(\cdot)$ is a lipschitzean function.

Combining now (94) with (93) proves the theorem. □

Corollary 1. *If the condition (85) is fulfilled not only at the point $\hat{x}$ itself but at each $x\in U(\hat{x})$ (in particular, if $\pi_C^{F,\theta}(\cdot)$ is Hölder continuous with an exponent $\beta>1/2$ on this neighbourhood), then the equality*

$$\partial^c\hat{u}(\hat{x})=\partial\rho_F(\hat{x}-\bar{\hat{x}})\cap\partial^-(\theta|_C)(\bar{\hat{x}}) \tag{95}$$

takes place whenever $\theta|_C$ is just lower regular at $\bar{\hat{x}}$. If, moreover, $\theta|_C$ is lower regular at each point of ∂C close to $\bar{\hat{x}}$ then the same equality as (95) holds also at each $x\in U(\hat{x})$.

Proof. Indeed, under the assumptions of the Corollary the inclusion (89) is valid at all points x close to $\hat{x}$, and passing to the *weak Kuratowski–Painlevé upper limits* we have

$$\begin{aligned}\partial\rho_F\left(\hat{x}-\bar{\hat{x}}\right)\cap\partial^-\left(\theta|_C\right)\left(\bar{\hat{x}}\right) &= \partial\rho_F\left(\hat{x}-\bar{\hat{x}}\right)\cap\partial^l\left(\theta|_C\right)\left(\bar{\hat{x}}\right)\\ &\subset \partial^l\hat{u}\left(\hat{x}\right)\subset\partial^c\hat{u}\left(\hat{x}\right)\ ,\end{aligned}$$

while the opposite inclusion is already proved [see (94)]. The second assertion is obvious. □

Observe that in the framework of Corollary 1 by strengthening the condition (85) we can achieve the Clarke regularity of $\hat{u}(\cdot)$ as well.

Proposition 1. *Given $\hat{x}\notin C$ let us assume that the mapping $x\mapsto\pi_C^{F,\theta}(x)$ is single-valued in a neighbourhood $U(\hat{x})$ of $\hat{x}$ and satisfies the Lipschitz type inequality*

$$\left\|\pi_C^{F,\theta}(x)-\pi_C^{F,\theta}(\hat{x})\right\|\le L\|x-\hat{x}\| \tag{96}$$

for all $x\in U(\hat{x})$ with some constant $L>0$. If, moreover, the restriction $\theta|_C$ is lower regular at $\bar{\hat{x}}:=\pi_C^{F,\theta}(\hat{x})$, then the statement of Theorem 4 holds.

Proof. The equalities in (86) split essentially into two inclusions. The first one is (94), which is obtained by using only the lower regularity of $\theta|_C$, and the second is

$$\partial\rho_F\left(\hat{x}-\bar{\hat{x}}\right)\cap\partial^-\left(\theta|_C\right)\left(\bar{\hat{x}}\right)\subset\partial^-\hat{u}\left(\hat{x}\right) \tag{97}$$

[compare with (89)].

Taking $p\in\partial\rho_F\left(\hat{x}-\bar{\hat{x}}\right)\cap\partial^-\left(\theta|_C\right)\left(\bar{\hat{x}}\right)$ similarly as in the proof of Theorem 4 (see Step 3) we write the inequality (91) for all $x\notin C$ sufficiently close to $\hat{x}$. Furthermore, given $\varepsilon>0$ we choose $\eta>0$ such that

$$\theta(z)\ge\theta\left(\bar{\hat{x}}\right)+\left\langle p,z-\bar{\hat{x}}\right\rangle-\frac{\varepsilon}{L}\left\|z-\bar{\hat{x}}\right\| \tag{98}$$

whenever $z\in C$ with $\left\|z-\bar{\hat{x}}\right\|\le\eta$ [compare with (92)]. By the continuity of $\pi_C^{F,\theta}(\cdot)$ let us choose $\delta>0$ such that $\left\|\bar{x}-\bar{\hat{x}}\right\|\le\eta$ whenever $\|x-\hat{x}\|\le\delta$. Setting then $z=\bar{x}$ in (98) and taking into account the inequality (96) we have:

$$\theta(\bar{x})\ge\theta\left(\bar{\hat{x}}\right)+\left\langle p,\bar{x}-\bar{\hat{x}}\right\rangle-\varepsilon\|x-\hat{x}\|\ . \tag{99}$$

Joining together (91) and (99) we obtain that

$$\begin{aligned}\hat{u}(x)-\hat{u}(\hat{x})-\langle p,x-\hat{x}\rangle &= \rho_F(x-\bar{x})+\theta(\bar{x})\\ &\quad-\rho_F\left(\hat{x}-\bar{\hat{x}}\right)-\theta\left(\bar{\hat{x}}\right)-\langle p,x-\hat{x}\rangle\\ &\ge-\left\langle p,\bar{x}-\bar{\hat{x}}\right\rangle+\theta(\bar{x})-\theta\left(\bar{\hat{x}}\right)\ge-\varepsilon\|x-\hat{x}\|\end{aligned}$$

for all x, $\|x-\hat{x}\| \leq \delta$. So, $p \in \partial^- \hat{u}(\hat{x})$, and the inclusion (97) is proved. □

We see that under the assumption of Corollary 1 the question of the Fréchet continuous differentiability of the value function $\hat{u}(\cdot)$ is reduced to the single-valuedness and the continuity of the mapping

$$\Phi(x) := \partial^- (\theta|_C)(\bar{x}) \cap \partial \rho_F (x-\bar{x}) \tag{100}$$

near a given point $\hat{x}$. Indeed, if $\Phi(x) = \partial^c \hat{u}(x)$ is a singleton, then by [12, Proposition 2.2.4] $\hat{u}(\cdot)$ is *strictly differentiable* and $\partial^c \hat{u}(x) = \{\nabla_H \hat{u}(x)\}$, where $\nabla_H \hat{u}(x)$ stands for the (strict) *Hadamard derivative* coinciding with the Fréchet one by the continuity. Observe that in finite dimensions the mapping (100) is continuous as soon as it is single-valued. This follows from the lower regularity of $\theta|_C$ and from the properties of the subdifferentials $\partial^l(\theta|_C)$ and $\partial \rho_F$.

Thus, recalling (12) we have a representation formula for the gradient $\nabla \hat{u}(\cdot)$ in a neighbourhood of $\hat{x}$ through the (unique) minimizer $\bar{x}$ of the function $y \mapsto \rho_F(x-y) + \theta(y)$ on C:

$$\nabla \hat{u}(x) = \partial^- (\theta|_C)(\bar{x}) \cap \mathbf{N}_F \left(\frac{x-\bar{x}}{\rho_F(x-\bar{x})} \right) \cap \partial F^0 \ . \tag{101}$$

Although the condition guaranteeing the continuous Fréchet differentiability in such a form [single-valuedness and continuity of the mapping (100)] and the formula (101) have a certain theoretical interest, their practical applicability is very restrictive because they are given in terms of an a priori unknown minimizer. To overcome this difficulty we propose first an alternate hypothesis regarding the regularity properties of either the function $\theta|_C$ or the gauge F. Namely, observing that the right-hand side in (101) is reduced to a singleton whenever either the Fréchet subdifferential $\partial^- (\theta|_C)(\bar{x})$ or the normal cone $\mathbf{N}_F \left(\frac{x-\bar{x}}{\rho_F(x-\bar{x})} \right)$ (both unbounded) becomes a semiline, we arrive at the following result.

Theorem 5. *Given $\hat{x} \in H \setminus C$ assume that in some neighbourhood $U(\hat{x})$ of $\hat{x}$ the mapping $x \mapsto \pi_C^{F,\theta}(x)$ is single-valued and Hölder continuous with an exponent $\beta > 1/2$, and that the restriction $\theta|_C$ is lower regular at $\bar{x} = \pi_C^{F,\theta}(x)$ for each x close to $\hat{x}$. Then the function $\hat{u}(\cdot)$ is (Fréchet) continuously differentiable on $U(\hat{x})$ if at least one of the conditions below holds:*

(**i**) *F is smooth at $\xi := \frac{x-\bar{x}}{\rho_F(x-\bar{x})}$ for each $x \in U(\hat{x})$;*

(**ii**) *C has smooth boundary at $\bar{x}$, and the function $\theta(\cdot)$ is of class $\mathcal{C}^1$ near this point.*

Furthermore, in the first case

$$\nabla \hat{u}(x) = \nabla \rho_F (x-\bar{x}) \tag{102}$$

(it coincides with the unique normal vector to F at the point ξ, belonging to the boundary ∂F^0), while in the second

$$\nabla \hat{u}(x) = \nabla \theta(\bar{x}) + \lambda(\bar{x})\, \mathfrak{n}_C(\bar{x}) \ , \tag{103}$$

where $\lambda = \lambda(\bar{x}) > 0$ *is the unique positive root of the equation*

$$\rho_{F^0}(\nabla \theta(\bar{x}) + \lambda \mathfrak{n}_C(\bar{x})) = 1 \ , \tag{104}$$

and $\mathfrak{n}_C(\bar{x})$ *is the (unique) unit normal vector to* C *at* $\bar{x}$.

Proof. The first assertion follows directly from (101) because the smoothness of F at ξ means exactly that the convex subdifferential $\partial \rho_F(\xi)$ is reduced to the unique point $\xi^* \in \mathfrak{J}_{F^0}(\xi) = \mathbf{N}_F(\xi) \cap \partial F^0$ [see (10)], which is nothing else than the Fréchet gradient $\nabla \rho_F(x - \bar{x})$. The continuity of $\nabla \hat{u}(\cdot)$ instead follows from both [25, Proposition 3.3 (ii)] and the continuity of $\pi_C^{F,\theta}(\cdot)$.

In the second case let us decompose the Fréchet subdifferential of $\theta|_C$ into the sum of the gradient $\nabla \theta(\bar{x})$ and the (Fréchet) normal cone to C (see [31, p. 112]):

$$\partial^-(\theta|_C)(\bar{x}) = \nabla \theta(\bar{x}) + \mathbf{N}_C^{\sigma}(\bar{x}) \ .$$

Since the boundary ∂C is assumed to be smooth at $\bar{\hat{x}}$, and $\bar{x}$ is close to $\bar{\hat{x}}$ whenever $x \in U(\hat{x})$, we have that

$$\mathbf{N}_C^{\sigma}(\bar{x}) = \mathbf{N}_C^{l}(\bar{x}) = \{\lambda \mathfrak{n}_C(\bar{x}) : \lambda \geq 0\} \ ,$$

where $\mathfrak{n}_C(\cdot)$ is a continuous function defined on ∂C near $\bar{\hat{x}}$ with $\|\mathfrak{n}_C(\bar{x})\| = 1$. So, the intersection $\partial^-(\theta|_C)(\bar{x}) \cap \partial F^0$ is the singleton $\nabla \theta(\bar{x}) + \lambda(\bar{x})\, \mathfrak{n}_C(\bar{x})$ where $\lambda(\bar{x}) > 0$ can be uniquely determined from the Eq. (104), and the gradient $\nabla \hat{u}(x)$ takes the form (103) [see (101)]. In order to show continuity let us fix a sequence $\{x_n\}$ converging to $x \in U(\hat{x})$. Then $\{\bar{x}_n\}$ converges to $\bar{x} \in \partial C$. Let us denote by $\lambda_n = \lambda(\bar{x}_n)$ the respective positive root of (104) and observe that the sequence $\{\lambda_n\}$ is bounded. Consequently, some of its subsequences (assume that $\{\lambda_n\}$ itself) converge to $\bar{\lambda} \geq 0$. Passing to limit in the equality

$$\rho_{F^0}(\nabla \theta(\bar{x}_n) + \lambda_n \mathfrak{n}_C(\bar{x}_n)) = 1$$

and using continuity of the involved functions we arrive at

$$\rho_{F^0}\left(\nabla \theta(\bar{x}) + \bar{\lambda}\, \mathfrak{n}_C(\bar{x})\right) = 1 \ .$$

Hence $\bar{\lambda} > 0$ [see (48)], and by the uniqueness $\bar{\lambda} = \lambda(\bar{x})$. Therefore $\nabla \hat{u}(x_n) \to \nabla \hat{u}(x)$, and the continuity is proved. □

Recall now that under the standing assumptions $(\mathbf{H})$, $(\hat{\mathbf{H}})$ and the local hypotheses $(\mathbf{H}_1(x_0))$, $(\mathbf{H}_2(x_0))$ the (single-valued) mapping $x \mapsto \pi_C^{F,\theta}(x)$ is locally lipschitzean, so satisfies the hypotheses of both Corollary 1 and Proposition 1 near a fixed point $x_0 \in C$ (see Theorem 3). Taking into account that $\pi_C^{F,\theta}(x)$ is close to x_0 whenever x approaches x_0, we give a version of the regularity theorem, which does not use explicitly the minimizers.

Theorem 6. *Let us fix $x_0 \in \partial C$ and suppose all the hypotheses of Theorem 3 to be valid. Assume, in addition, that $\theta|_C$ is lower regular near x_0, and that for some $\delta > 0$ at least one of the conditions below is fulfilled:*

(**i**) *F is smooth at each $\xi \in \mathfrak{J}_F\left(\partial^-(\theta|_C)(x) \cap \partial F^0\right)$, $x \in \partial C$, $\|x - x_0\| \leq \delta$;*
(**ii**) *C has smooth boundary, and $\theta(\cdot)$ is of class $\mathcal{C}^1$ on $\partial C \cap \left(x_0 + \delta\overline{\mathbf{B}}\right)$.*

Then the marginal function $\hat{u}(\cdot)$ is (Fréchet) continuously differentiable on a neighbourhood of x_0 (outside of C), and the gradient $\nabla\hat{u}(x)$ can be computed by the formula (102) or (103), respectively.

Proof. By Theorem 3 there exists a neighbourhood $\mathcal{U}(x_0)$ of x_0 such that for each $x \in \mathcal{U}(x_0)$ the set $\pi_C^{F,\theta}(x)$ is a singleton, say $\{\bar{x}\}$, and $\|\bar{x} - x_0\| \leq \delta$. Now, in the case (**i**) we apply Corollary 1 at the point $x \in \mathcal{U}(x_0) \setminus C$ and consider the unique vector

$$\xi^* \in \partial^-(\theta|_C)(\bar{x}) \cap \mathbf{N}_F(\xi) \cap \partial F^0$$

where $\xi := \frac{x-\bar{x}}{\rho_F(x-\bar{x})}$ [see (101)]. In particular, $\xi^* \in \mathfrak{J}_{F^0}(\xi)$, or dually $\xi \in \mathfrak{J}_F(\xi^*)$. So, we are led to the hypothesis (**i**) of Theorem 5. The case (**ii**) instead is directly reduced to Theorem 5 (**ii**). □

By duality, in the place of the *smoothness* of F we may require here the *rotundity* of F^0 with respect to each $\xi \in \mathfrak{J}_F\left(\partial^-(\theta|_C)(x) \cap \partial F^0\right)$ (compare with the condition $(\mathbf{H}_2(x_0))$).

If the hypotheses $(\mathbf{H}_1(x_0))$ and $(\mathbf{H}_2(x_0))$ hold at each point $x_0 \in \partial C$, and the restriction $\theta|_C$ is lower regular everywhere on the boundary ∂C, then in order to have the continuous differentiability of $\hat{u}(\cdot)$ in some neighbourhood $\mathfrak{A} \supset C$ (outside of C) one can alternate the conditions (**i**) and (**ii**) from one point $x_0 \in \partial C$ to other.

In conclusion let us strengthen the hypotheses on F, C and $\theta(\cdot)$ in order to have more regularity for the value function $\hat{u}(\cdot)$. Remind that in the case $\theta \equiv 0$ and $F = \overline{\mathbf{B}}$ under well-posedness assumptions, which are reduced to the φ-convexity of C, the function $\hat{u}(\cdot) = \mathrm{d}_C(\cdot)$ is of class $\mathcal{C}^{1,1}$ near C, while in the case of an arbitrary gauge the lipschitzeanity (hölderianity, in general) of $\nabla\hat{u}(\cdot)$ depends on the order of smoothness of the input data (see [26, Theorems 5.6 and 5.7]). The same happens in the case $\theta \neq 0$.

Theorem 7. *Given $x_0 \in \partial C$ let us assume that all the hypotheses of Theorem 6 hold. Moreover, suppose that in the case* (**i**) *the gradient $\nabla\rho_F(\cdot)$ is Hölder continuous with an exponent $0 < \alpha \leq 1$ on the set*

$$\mathfrak{M}_\delta(x_0) := \bigcup_{x \in \partial C, \|x - x_0\| \leq \delta} \mathfrak{J}_F\left(\partial^-(\theta|_C)(x) \cap \partial F^0\right)$$

(equivalently, the unit normal vector to F moves in a hölderean way along the part $\mathfrak{M}_\delta(x_0)$ of the boundary ∂F), while in the case (**ii**) *both the gradient $\nabla\theta(\cdot)$ and the normal $\mathfrak{n}_C(\cdot)$ are Hölder continuous (with the exponent $0 < \alpha \leq 1$) near x_0. Then the value function $\hat{u}(\cdot)$ is of class $\mathcal{C}^{1,\alpha}_{loc}$ in a neighbourhood of x_0 (outside of C).*

Proof. After application of Theorem 6 the proving consists in the verification (locally) the Hölder inequality for the gradient $\nabla\hat{u}(x)$ in both cases. Let $\mathcal{U}(x_0)$ be the neighbourhood of x_0 constructed in Theorem 3. Without loss of generality assume that $\delta > 0$ from the formula (65) is the same as in the conditions (**i**) and (**ii**) of Theorem 6. Fix $x \in \mathcal{U}(x_0) \setminus C$ and choose $\bar{\delta} > 0$ such that $x + \bar{\delta}\overline{\mathbf{B}} \subset \mathcal{U}(x_0) \setminus C$. We have $\mathcal{U}(x_0) \subset x_0 + \delta\overline{\mathbf{B}}$. Moreover, for each $z \in x + \left(\bar{\delta}/2\right)\overline{\mathbf{B}}$ the (unique) minimizer $\bar{z} := \pi_C^{F,\theta}(z)$ also belongs to $x_0 + \delta\overline{\mathbf{B}}$ as shown in the first part of the proof of Theorem 3. Now, let us consider the respective estimates in each case separately.

(**i**) Given $z_1, z_2 \in x + \left(\bar{\delta}/2\right)\overline{\mathbf{B}}$ we denote by

$$\xi_i := \frac{z_i - \bar{z}_i}{\rho_F(z_i - \bar{z}_i)} \ , \quad \bar{z}_i := \pi_C^{F,\theta}(z_i) \ , \quad i = 1,2 \ ,$$

and by the positive homogeneity of the Minkowski functional deduce from (102) that

$$\|\nabla\hat{u}(z_1) - \nabla\hat{u}(z_2)\| \leq \mathfrak{h}\,\|\xi_1 - \xi_2\|^{\alpha} \ , \tag{105}$$

where $\mathfrak{h} > 0$ is the Hölder constant of $\nabla\rho_F(\cdot)$ on $\mathfrak{M}_\delta(x_0)$. Setting for the sake of brevity $\rho_i := \rho_F(z_i - \bar{z}_i)$, we further have

$$\begin{aligned}
\|\xi_1 - \xi_2\| &\leq \frac{1}{\rho_1\rho_2}\left(\|\bar{z}_1 - z_1\|\,|\rho_2 - \rho_1| + \rho_1\,\|(\bar{z}_1 - \bar{z}_2) + (z_2 - z_1)\|\right) \\
&\leq \frac{1}{\rho_2}\left(\|F\|\,\left\|F^0\right\| + 1\right)\|(\bar{z}_1 - \bar{z}_2) + (z_2 - z_1)\| \\
&\leq \frac{1}{\rho_2}\left(\|F\|\,\left\|F^0\right\| + 1\right)\left(\|\bar{z}_1 - \bar{z}_2\| + \|z_1 - z_2\|\right) \ .
\end{aligned} \tag{106}$$

Notice that $\|z_2 - \bar{z}_2\| > \bar{\delta}/2$ because otherwise $\|\bar{z}_2 - x\| \leq \bar{\delta}$, contradicting the choice of $\bar{\delta} > 0$. Consequently, $\rho_2 \geq \frac{\bar{\delta}}{2\|F\|}$ [see (9)]. Hence, using the Lipschitz continuity of $\pi_C^{F,\theta}(\cdot)$ (with the Lipschitz constant $K > 0$) we obtain from (106) that

$$\|\xi_1 - \xi_2\| \leq \frac{2\|F\|\left(\|F\|\,\left\|F^0\right\| + 1\right)(K+1)}{\bar{\delta}}\,\|z_1 - z_2\| \ .$$

Joining the latter inequality with (105) we arrive at

$$\|\nabla\hat{u}(z_1) - \nabla\hat{u}(z_2)\| \leq \mathfrak{H}\,\|z_1 - z_2\|^{\alpha} \ , \tag{107}$$

where the constant $\mathfrak{H} > 0$ essentially depends on x (through $\bar{\delta}$ and K) and tends to $+\infty$ as the point x approaches the target C.

(**ii**) In this case we prove a Hölder inequality like (107) in the neighbourhood $x + \bar{\delta}\overline{\mathbf{B}}$. To this end we apply the Lipschitz continuity of $\pi_C^{F,\theta}(\cdot)$ on $\mathcal{U}(x_0) \supset x + \bar{\delta}\overline{\mathbf{B}}$

and the Hölder continuity of both $\nabla\theta(\cdot)$ and $\mathfrak{n}_C(\cdot)$ on $\partial C \cap \left(x_0 + \delta\overline{\mathbf{B}}\right)$. Let us take $z_1, z_2 \in x + \bar{\delta}\overline{\mathbf{B}}$ and set as usual $\bar{z}_i := \pi_C^{F,\theta}(z_i)$, $i = 1,2$. Then it follows from (103) that

$$\begin{aligned}\|\nabla\hat{u}(z_1) - \nabla\hat{u}(z_2)\| \leq \|\nabla\theta(\bar{z}_1) - \nabla\theta(\bar{z}_2)\| + |\lambda(\bar{z}_1) - \lambda(\bar{z}_2)| \\ + \lambda(\bar{z}_1)\|\mathfrak{n}_C(\bar{z}_1) - \mathfrak{n}_C(\bar{z}_2)\|\end{aligned} \tag{108}$$

where $\lambda(\bar{z}_i) > 0$, $i = 1,2$, satisfy the equality

$$\rho_{F^0}(\nabla\theta(\bar{z}_i) + \lambda(\bar{z}_i)\mathfrak{n}_C(\bar{z}_i)) = 1 \ . \tag{109}$$

Notice that (109) is equivalent to

$$\frac{1}{\lambda(\bar{z}_i)} = \rho_{F^0 - \nabla\theta(\bar{z}_i)}(\mathfrak{n}_C(\bar{z}_i)) \ .$$

Then, due to (9) and to the hypothesis (**H**) [see (48)]

$$\lambda(\bar{z}_i) \leq \left\|F^0 - \nabla\theta(\bar{z}_i)\right\| \leq (1+\gamma)\left\|F^0\right\| \ . \tag{110}$$

On the other hand, by the lipschitzeanity of the gauge function, (8), (53), (16), (18) and (52) we successively have:

$$\begin{aligned}\left|\frac{1}{\lambda(\bar{z}_1)} - \frac{1}{\lambda(\bar{z}_2)}\right| &\leq \left|\rho_{F^0-\nabla\theta(\bar{z}_1)}(\mathfrak{n}_C(\bar{z}_1)) - \rho_{F^0-\nabla\theta(\bar{z}_1)}(\mathfrak{n}_C(\bar{z}_2))\right| \\ &\quad + \left|\rho_{F^0-\nabla\theta(\bar{z}_1)}(\mathfrak{n}_C(\bar{z}_2)) - \rho_{F^0-\nabla\theta(\bar{z}_2)}(\mathfrak{n}_C(\bar{z}_2))\right| \\ &\leq \left\|\left(F^0 - \nabla\theta(\bar{z}_1)\right)^0\right\| \|\mathfrak{n}_C(\bar{z}_1) - \mathfrak{n}_C(\bar{z}_2)\| \\ &\quad + \left|\sigma_{\left(F^0-\nabla\theta(\bar{z}_1)\right)^0}(\mathfrak{n}_C(\bar{z}_2)) - \sigma_{\left(F^0-\nabla\theta(\bar{z}_2)\right)^0}(\mathfrak{n}_C(\bar{z}_2))\right| \\ &\leq \frac{\|F\|}{1-\gamma}\|\mathfrak{n}_C(\bar{z}_1) - \mathfrak{n}_C(\bar{z}_2)\| \\ &\quad + \mathcal{D}\left(\left(F^0 - \nabla\theta(\bar{z}_1)\right)^0, \left(F^0 - \nabla\theta(\bar{z}_2)\right)^0\right) \\ &\leq \frac{\|F\|}{1-\gamma}\|\mathfrak{n}_C(\bar{z}_1) - \mathfrak{n}_C(\bar{z}_2)\| \\ &\quad + \left(\frac{\|F\|}{1-\gamma}\right)^2 \|\nabla\theta(\bar{z}_1) - \nabla\theta(\bar{z}_2)\| \ .\end{aligned} \tag{111}$$

Applying the Hölder inequality for both $\mathfrak{n}_C(\cdot)$, $\nabla\theta(\cdot)$ with the exponent $0 < \alpha \leq 1$ and a Hölder constant $\mathfrak{h} > 0$ we obtain from (110) and (111) that

$$\begin{aligned}|\lambda(\bar{z}_1) - \lambda(\bar{z}_2)| &= \lambda(\bar{z}_1)\lambda(\bar{z}_2)\left|\frac{1}{\lambda(\bar{z}_1)} - \frac{1}{\lambda(\bar{z}_2)}\right| \\ &\leq \frac{(1+\gamma)^2}{1-\gamma}\left(1 + \frac{\|F\|}{1-\gamma}\right)\|F\|\left\|F^0\right\|^2 \mathfrak{h}\|\bar{z}_1 - \bar{z}_2\|^\alpha \ .\end{aligned} \tag{112}$$

Again by the Hölder continuity of the functions $\nabla\theta(\cdot)$ and $\mathfrak{n}_C(\cdot)$ and by the inequalities (108), (110) and (112) it follows that

$$\|\nabla\hat{u}(z_1) - \nabla\hat{u}(z_2)\| \le \bar{\mathfrak{h}}\,\|\bar{z}_1 - \bar{z}_2\|^{\alpha}$$

with some constant $\bar{\mathfrak{h}} > 0$, which is proportional to $\mathfrak{h}$. Recalling also the Lipschitz inequality for minimizers (with a Lipschitz constant $K > 0$) we arrive, finally, at (107) where $\mathfrak{H} := \bar{\mathfrak{h}}K^{\alpha}$.

So the theorem is completely proved. □

Observe that unlike the item (**i**) the Hölder constant in the case (**ii**) depends possibly on how close to the boundary ∂C the point x is just through the lipschitzeanity of the mapping $\pi_C^{F,\theta}(\cdot)$ and the hölderianity of both $\nabla\theta(\cdot)$ and $\mathfrak{n}_C(\cdot)$ (remind that in the case of metric projections onto convex and prox-regular sets the gradient $\nabla\hat{u}(x)$ exactly coincides with the unit normal $\mathfrak{n}_C(\bar{x})$). However, in both cases $\mathfrak{H}$ tends to $+\infty$ as the point x approaches the boundary of the neighbourhood $\mathcal{U}(x_0)$, where the regularity hypotheses $(\mathbf{H}_1(x_0))$ and $(\mathbf{H}_2(x_0))$ fail (see Remark 4).

Remark 6. Notice that Theorems 5–7 have been proved under the lower regularity assumption for the function $\theta|_C$, while in [26] we supposed the target set C to be proximal regular that is a much stronger property.

8 Examples

In this section we illustrate the obtained results with two simple examples restricting ourselves just to the case $H = \mathbb{R}^2$. We recommend to compare them with the examples given earlier for the case $\theta \equiv 0$ (see [25, 26]).

Example 1.

$$\begin{aligned} F &:= \left\{(\xi_1,\xi_2) \in \mathbb{R}^2 : |\xi_2| \le 1 - \xi_1^4,\ -1 \le \xi_1 \le 1\right\}\ ; \\ C &:= \left\{(x_1,x_2) \in \mathbb{R}^2 : x_1 \le x_2^2\right\}\ ; \\ \theta(x) &:= \frac{1}{2}\mathrm{arctg}\left(x_1^2 + x_2^2\right)\ ,\quad x = (x_1,x_2) \in C\ . \end{aligned}$$

In [25, Example 8.3] the well-posedness problem for the same target set C and the same dynamics F with $\theta \equiv 0$ was considered. Now we somewhat complicate the problem by introducing a (smooth) nonlinear boundary function.

Let us verify first the standing slope condition (24), guaranteeing that the value function $\hat{u}(\cdot)$ is the viscosity solution of the associated Hamilton–Jacobi equation (5) with the boundary datum $u|_C = \theta$ (see Theorem 1). To this end we calculate the gradient

$$\nabla\theta(x) = \left(\frac{x_1}{1+\left(x_1^2+x_2^2\right)^2}, \frac{x_2}{1+\left(x_1^2+x_2^2\right)^2} \right)$$

and the dual gauge function

$$\rho_{F^0}(\xi^*) = \sigma_F(\xi^*) = \begin{cases} 3\frac{\left(|\xi_1^*|/4\right)^{4/3}}{|\xi_2^*|^{1/3}} + |\xi_2^*| & \text{if} \quad |\xi_2^*| \geq |\xi_1^*|/4 \ ; \\ |\xi_1^*| & \text{if} \quad |\xi_2^*| < |\xi_1^*|/4 \ , \end{cases} \tag{113}$$

$\xi^* = (\xi_1^*, \xi_2^*) \in \mathbb{R}^2$. Comparing the radii of two circles centered at the origin, one inscribed into F and other circumscribed around it, we find that $\|F\| \leq 7/6$ and $\left\|F^0\right\| \leq 9/8$. Substituting $\nabla\theta(x)$ in the place of ξ^* in (113) we obviously have

$$\rho_{F^0}(\nabla\theta(x)) \leq \begin{cases} \frac{3}{4}\frac{|x_1|}{1+\left(x_1^2+x_2^2\right)^2} + \frac{|x_2|}{1+\left(x_1^2+x_2^2\right)^2} & \text{if} \quad |x_2| \geq \frac{|x_1|}{4} \ ; \\ \frac{|x_1|}{1+\left(x_1^2+x_2^2\right)^2} & \text{if} \quad |x_2| < \frac{|x_1|}{4} \ . \end{cases} \tag{114}$$

The function in the right-hand side of (114) attends its maximum at the point $\left(\frac{3^{3/4}}{5}, \frac{4}{3^{1/4}\cdot 5}\right)$, and the maximum is $\frac{5\cdot 3^{3/4}}{16} < \frac{35}{48}$. Thus, $\nabla\theta(x) \in \gamma F^0$ with

$$\gamma := \frac{35}{48} < \frac{1}{\|F\|\,\|F^0\|} \ , \tag{115}$$

and we have not merely the slope condition (24) but also the (stronger) standing hypothesis (**H**) required for the well-posedness and regularity results. Moreover, by [14, p. 38] the second standing assumption ($\hat{\mathbf{H}}$) holds as well with $\Gamma(x) = \{\nabla\theta(x)\}$ and

$$\begin{aligned} \mathbf{N}_C^\theta(x) &= \mathbf{N}_C^p(x) = \mathbf{N}_C^l(x) \\ &= \{\lambda \mathfrak{n}_C(x) : \lambda \geq 0\} = \{(\lambda, -2\lambda x_2) : \lambda \geq 0\} \ , \end{aligned}$$

where $\mathfrak{n}_C(x)$ is the unit normal vector to C,

$$\mathfrak{n}_C(x) = \frac{1}{\sqrt{1+4x_2^2}}(1, -2x_2) \ ,$$

$x = (x_1, x_2) \in \partial C$. Thus,

$$\partial^p(\theta|_C)(x) = \left\{ \left(\frac{x_1}{1+\left(x_1^2+x_2^2\right)^2} + \lambda, \frac{x_2}{1+\left(x_1^2+x_2^2\right)^2} - 2\lambda x_2 \right) : \lambda \geq 0 \right\} \ .$$

Let us fix now $x_0 = \left(x_1^0, x_2^0\right) \in \partial C$ and verify the local hypotheses of Theorem 3, or, rather, just the hypothesis $(\mathbf{H}_1(x_0))$ according to the observation before Theorem 3. To this end we compute first the value $\mathfrak{J}_F(\xi^*)$, $\xi^* = (\xi_1^*, \xi_2^*) \in \partial F^0$, restricting ourselves just to the case $\xi_1^* > 0$, since the first coordinate of the (unique) element of $\partial^p(\theta|_C)(x) \cap \partial F^0$, $x \in C$, is positive. By the formula (10) due to the continuous differentiability of $\rho_{F^0}(\cdot)$ we have that $\mathfrak{J}_F(\xi^*) = \{\nabla \rho_{F^0}(\xi^*)\}$, while the direct derivation of (113) gives

$$\nabla \rho_{F^0}(\xi^*) = \left(f\left(\frac{\xi_1^*}{4\left|\xi_2^*\right|}\right), \operatorname{sgn}(\xi_2^*)\, g\left(\frac{\xi_1^*}{4\left|\xi_2^*\right|}\right)\right) \ , \tag{116}$$

where $f(\cdot)$ and $g(\cdot)$ are real functions defined on $]0, +\infty[$ by

$$f(t) := \begin{cases} t^{1/3} & \text{if} \quad 0 < t \le 1 \ , \\ 1 & \text{if} \quad t > 1 \ , \end{cases} \tag{117}$$

$$g(t) := \begin{cases} 1 - t^{4/3} & \text{if} \quad 0 < t \le 1 \ , \\ 0 & \text{if} \quad t > 1 \ . \end{cases} \tag{118}$$

Then we substitute in the place of ξ^* in (116) the unique subgradient of $\theta|_C$ belonging to ∂F^0, i.e.,

$$\xi_1^* = \frac{x_1}{1 + \left(x_1^2 + x_2^2\right)^2} + \lambda \ ; \tag{119}$$

$$\xi_2^* = \frac{x_2}{1 + \left(x_1^2 + x_2^2\right)^2} - 2\lambda x_2 \ , \tag{120}$$

where $\lambda = \lambda(x)$ is the (unique) positive root of the equation $\rho_{F^0}(\xi_1^*, \xi_2^*) = 1$, $x \in \partial C$ (i.e., $x_1 = x_2^2$) with $\|x - x_0\| \le \delta$. Similarly as in the proof of Theorem 7 [see (110) and (53)] we obtain the following estimates for the parameter λ [see also (115)]:

$$\frac{1}{5} < \frac{13}{56} \le \frac{1-\gamma}{\|F\|} \le \lambda(x) \le (1+\gamma)\left\|F^0\right\| \le \frac{249}{128} < 2 \tag{121}$$

and establish the lipschitzeanity of the function $\lambda(\cdot)$ near x_0 [see (112)]. It follows from (119)–(121) that

- $\frac{\xi_1^*}{4\left|\xi_2^*\right|} \ge 1$ whenever $\left|\xi_2^*\right| \le \frac{1}{20}$;
- $\frac{\xi_1^*}{4\left|\xi_2^*\right|} \ge \frac{\lambda(x)}{4\left\|F^0\right\|} \ge s := \frac{2}{45}$ whenever $x \in \partial C$.

Taking this into account and observing that the functions (117) and (118) are lipschitzean on $[s, +\infty[$ with the Lipschitz constant $1/3 \max\left\{4, s^{-2/3}\right\}$, that they are constant for $t \ge 1$, and that the mapping $\xi^* \mapsto \frac{\xi_1^*}{4\left|\xi_2^*\right|}$ is lipschitzean on the set

$$\left\{\xi^* \in \partial F^0 : |\xi_2^*| \geq \frac{1}{20}\right\} ,$$

we conclude that the gradient $\nabla \rho_{F^0}(\cdot)$ [see (116)] is lipschitzean on

$$\left\{\xi^* \in \partial F^0 : \frac{\xi_1^*}{4|\xi_2^*|} \geq s\right\} .$$

Consequently, estimating further the second derivative of the function $\theta(\cdot)$ we obtain the lipschitzeanity of the composed mapping

$$x \mapsto \nabla \rho_{F^0}\left(\frac{x_1}{1+\left(x_1^2+x_2^2\right)^2}+\lambda(x), \frac{x_2}{1+\left(x_1^2+x_2^2\right)^2}-2\lambda(x)x_2\right)$$

on the set $C_\delta(x_0)$.

Thus, all the conditions of Theorem 3 are fulfilled, and we can affirm that the function

$$\frac{1}{2}\operatorname{arctg}\left(z_1^2+z_2^2\right)+\rho_F(x_1-z_1, x_2-z_2)$$

admits an unique minimizer $\pi_C^{F,\theta}(x)$ on C, which is Lipschitz continuous w.r.t. x in a neighbourhood of each point $x_0 \in \partial C$ (out of C). This neighbourhood is given by the formula (65), where the Lipschitz constant $L > 0$ of the mapping

$$x \mapsto \mathfrak{J}_F\left(\partial^p(\theta|_C)(x) \cap \partial F^0\right) , \quad x \in C_\delta(x_0) ,$$

can be computed by using the above arguments.

Furthermore, the restriction $\theta|_C$ is obviously lower (even proximally) regular on ∂C, and the condition (**ii**) of Theorems 6 and 7 holds (the condition (**i**) is violated in the "angle" point $(1,0)$). Therefore, applying Theorem 7 we see that the value function $\hat{u}(x)$ in the above mathematical programming problem, which can be interpreted also as the viscosity solution to the Hamilton–Jacobi equation

$$\min\left\{\left|\frac{\partial u}{\partial x_1}\right|, \frac{3}{4}\left|\frac{\partial u}{\partial x_1}\right| \sqrt[3]{\frac{\left|\frac{\partial u}{\partial x_1}\right|}{4\left|\frac{\partial u}{\partial x_2}\right|}}+\left|\frac{\partial u}{\partial x_2}\right|\right\}=1 ,$$

$$u\left(x_2^2, x_2\right)=\frac{1}{2}\operatorname{arctg}\left(x_1^2+x_2^2\right)$$

[see (113)], is of class $\mathcal{C}_{\text{loc}}^{1,1}$ on an open set $\left\{(x_1,x_2) : x_2^2 < x_1 < x_2^2+\eta(x_2)\right\}$, where $\eta(\cdot)$ is a positive real function.

By the next example we test the case of nonsmooth both a target C and a boundary function $\theta(\cdot)$. It shows, in particular, that the hypotheses of Theorem 3 can be fulfilled even if the target C has an *"inward" angle point*.

Example 2.

$$\begin{aligned} F &:= \left\{x \in \mathbb{R}^2 : \|x\| \le 1 \quad \text{and} \quad \langle \mathbf{v}, x\rangle + \mu \|x - \mathbf{v}\| \le 1\right\} \ ; \\ C &:= \left\{x \in \mathbb{R}^2 : \min\{x_1, x_2\} \le 0\right\} \ ; \\ \theta(x) &:= \max\{\langle \mathbf{a}, x\rangle, \langle \mathbf{b}, x\rangle\} \ , \quad x = (x_1, x_2) \in C \ . \end{aligned} \tag{122}$$

Here $0 < \mu < 1$ and $\mathbf{v}$, $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^2$ are such that $\|\mathbf{v}\| = 1$, $\mathbf{v}_i > 0$, $\|\mathbf{a}\| < 1$, $\|\mathbf{b}\| < 1$, $\mathbf{a}_i \ge 0$, $\mathbf{b}_i \ge 0$, $i = 1, 2$, and $\mathbf{a}_1 \ne \mathbf{b}_1$, $\mathbf{a}_2 \ne \mathbf{b}_2$. Our goal is to find conditions on the choice of the parameters $\mathbf{v}$, $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^2$ and μ, under which the well-posedness and regularity results of the previous sections hold.

Since F can be represented as $\overline{\mathbf{B}} \cap \left(\mathbf{v} + K_{\mathbf{v},\mu}\right)$ where

$$K_{\mathbf{v},\mu} := \{x : \langle -\mathbf{v}, x\rangle \ge \mu \|x\|\}$$

is a closed convex cone in $\mathbb{R}^2$, from elementary geometric considerations we obtain that F^0 is the convexification of the unit circle $\overline{\mathbf{B}}$ and a symmetric segment of the tangent line at the point $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)$. Namely,

$$\begin{aligned} F^0 &= \overline{\text{co}}\left(\overline{\mathbf{B}} \cup \left(\mathbf{v} + K_{\mathbf{v},\mu}\right)^0\right) \\ &= \overline{\text{co}}\left(\overline{\mathbf{B}} \cup \left\{(\mathbf{v}_1 + \lambda \mathbf{v}_2, \mathbf{v}_2 - \lambda \mathbf{v}_1) : |\lambda| \le \frac{\mu}{\sqrt{1-\mu^2}}\right\}\right) \ . \end{aligned} \tag{123}$$

We obviously have $\|F\| = 1$, $\left\|F^0\right\| = \frac{1}{\sqrt{1-\mu^2}}$, and $\mathfrak{J}_F(\xi^*) = \{\mathbf{v}\}$ for each $\xi^* = (\mathbf{v}_1 + \lambda \mathbf{v}_2, \mathbf{v}_2 - \lambda \mathbf{v}_1) \in \partial F^0$ with $|\lambda| < \frac{\mu}{\sqrt{1-\mu^2}}$.

The target set C admits the unit normal vector

$$\mathfrak{n}_C(x) = \begin{cases} (1,0) & \text{if} \quad x_1 = 0 \ , \quad x_2 > 0 \ , \\ (0,1) & \text{if} \quad x_1 > 0 \ , \quad x_2 = 0 \end{cases} \tag{124}$$

at each point of the boundary ∂C except the origin, where the proximal and the Fréchet normal cones are trivial.

The function $\theta(\cdot)$ is convex and admits the piecewise constant gradient

$$\nabla\theta(x) = \begin{cases} \mathbf{a} & \text{if} \quad \langle \mathbf{a} - \mathbf{b}, x\rangle > 0 \ , \\ \\ \mathbf{b} & \text{if} \quad \langle \mathbf{a} - \mathbf{b}, x\rangle < 0 \ , \end{cases} \tag{125}$$

while

$$\partial\theta(x) = \partial^c\theta(x) = \{\lambda \mathbf{a} + (1-\lambda)\mathbf{b} \ , \quad 0 \le \lambda \le 1\}$$

whenever $\langle \mathbf{a} - \mathbf{b}, x\rangle = 0$ (see [12, Theorem 2.5.1]). Hence, we deduce the first condition, under which the standing hypothesis (**H**) is fulfilled [see (48)]:

$$\max\{\|\mathbf{a}\|,\|\mathbf{b}\|\} < \frac{1}{\|F\|\,\|F^0\|} = \sqrt{1-\mu^2}\ . \tag{126}$$

Since at each point $x \in \partial C$ with $x \neq 0$ both the function $\theta(\cdot)$ and the set C are *proximally regular* (moreover, $\theta(\cdot)$ is of class $\mathcal{C}^2$), we have that

$$\begin{aligned}\partial^p(\theta|_C)(x) &= \partial\theta(x) + \mathbf{N}_C^p(x)\\ &= \{\nabla\theta(x) + \lambda\,\mathrm{n}_C(x) : \lambda \geq 0\}\ .\end{aligned} \tag{127}$$

Taking into account (125) and (124) we may further represent (127) in an alternate form depending on the mutual location of the vectors $\mathbf{a}$ and $\mathbf{b}$. Let us restrict ourselves just to the case when $\mathbf{a}_1 < \mathbf{b}_1$ and $\mathbf{a}_2 > \mathbf{b}_2$. Then

$$\partial^p(\theta|_C)(x) = \begin{cases}\{(\mathbf{a}_1+\lambda,\mathbf{a}_2) : \lambda \geq 0\} & \text{if} \quad x_1 = 0\ ,\ \ x_2 > 0\ ,\\ \{(\mathbf{b}_1,\mathbf{b}_2+\lambda : \lambda \geq 0)\} & \text{if} \quad x_1 > 0\ ,\ \ x_2 = 0\ .\end{cases} \tag{128}$$

At the origin instead we compute this subdifferential directly by the definition. In fact, $\partial^p(\theta|_C)(0)$ is the triangle $\Delta := co\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ where $\mathbf{c} := (\mathbf{b}_1,\mathbf{a}_2)$. So the hypothesis $(\hat{\mathbf{H}})$ also holds with

$$\Gamma(x) = \begin{cases}\partial\theta(x) & \text{if} \quad x \neq 0\ ;\\ \Delta & \text{if} \quad x = 0\ ,\end{cases}$$

and $\mathbf{N}_C^\theta(x) = \mathbf{N}_C^p(x)$ at each $x \in \partial C$ if we set

$$\gamma := \|\mathbf{c}\| = \sqrt{\mathbf{b}_1^2+\mathbf{a}_2^2} < \sqrt{1-\mu^2}.$$

Notice that here $\partial^\theta C = \partial C \setminus \{0\}$.

Given $\delta > 0$ simple geometric considerations give that the condition

$$\max\left\{\frac{|\mathbf{b}_1-\mathbf{v}_1|}{\mathbf{v}_2}, \frac{|\mathbf{a}_2-\mathbf{v}_2|}{\mathbf{v}_1}\right\} \leq \frac{\mu}{\sqrt{1-\mu^2}} \tag{129}$$

ensures that the intersection $\partial^p(\theta|_C)(x) \cap \partial F^0$ is contained in the line segment $\left\{(\mathbf{v}_1+\lambda\mathbf{v}_2,\mathbf{v}_2-\lambda\mathbf{v}_1) : |\lambda| \leq \frac{\mu}{\sqrt{1-\mu^2}}\right\}$ [see(123)] for each $x \in C_\delta(0)$. Furthermore, if the inequality (129) is strict, then (see above)

$$\mathfrak{J}_F\left(\partial^p(\theta|_C)(x) \cap \partial F^0\right) = \{\mathbf{v}\}\ \ ,\ \ x \in C_\delta(0)\ .$$

So, the hypothesis $(\mathbf{H}_1(0))$ is trivially fulfilled. Notice that the verification of the respective hypothesis at each point $x_0 \in \partial C$, $x_0 \neq 0$, is reduced to the (more general) case $x_0 = 0$. So, everything said above is sufficient to be able to apply Theorem 3 and to conclude that under the assumptions

$$\sqrt{b_1^2+a_2^2} < \sqrt{1-\mu^2}; \tag{130}$$

$$\max\left\{\frac{|\mathbf{b}_1-\mathbf{v}_1|}{\mathbf{v}_2},\frac{|\mathbf{a}_2-\mathbf{v}_2|}{\mathbf{v}_1}\right\}<\frac{\mu}{\sqrt{1-\mu^2}} \tag{131}$$

the minimization problem for the function

$$\max\left\{\langle\mathbf{a},z\rangle,\langle\mathbf{b},z\rangle\right\}+\rho_F(x-z) \tag{132}$$

subject to $\min\{z_1,z_2\}\le 0$, where F is defined by (122), admits an unique minimizer, which is Lipschitz continuous w.r.t. x from an open neighbourhood $\mathfrak{A}$ of the constraint set.

For instance, setting $\mathbf{v}=\left(\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}\right)$, $\mathbf{a}=(0,q)$ and $\mathbf{b}=(q,0)$, $0<q<1$, we see that the conditions (130) and (131) hold whenever the parameters μ and q satisfy the inequalities

$$1-\frac{\mu}{\sqrt{1-\mu^2}}<\sqrt{2}q<\sqrt{1-\mu^2}$$

(e.g., $\mu=\sqrt{3}/2$ and $q=1/3$). In this case we are led to minimize the function

$$q\max\{z_1,z_2\}+\rho_F(x-z)\quad,\quad z\in C\ . \tag{133}$$

Unfortunately, we are not able to deduce anything about the Fréchet continuous differentiability of $\hat{u}(\cdot)$ near the origin due to the lack of smoothness of the input data F, C and $\theta(\cdot)$.

Acknowledgements Work is fulfilled in framework of the project "Variational Analysis: Theory and Applications" (PTDC/MAT/111809/2009) financially supported by Fundação para Ciência e Tecnologia (FCT), the Portuguese institutions COMPETE, QREN and the European Regional Development Fund (FEDER).

References

1. Aubin, J.-P., Cellina, A.: Differential Inclusions. Springer, Berlin (1984)
2. Aubin, J.-P., Frankowska, H.: Set-Valued Analysis. Birkhauser, Boston (1990)
3. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations. Birkhäuser, Boston (1997)
4. Bernard, F., Thibault, L., Zlateva, N.: Characterizations of prox-regular sets in uniformly convex Banach spaces. J. Convex Anal. **13**, 525–559 (2006)
5. Bernard, F., Thibault, L., Zlateva, N.: Prox-regular sets and epigraphs in uniformly convex Banach spaces: various regularities and other properties. Trans. Am. Math. Soc. **363**, 2211–2247 (2011)
6. Bounkhel, M., Thibault, L.: On various notions of regularity of sets in nonsmooth analysis. Nonlin. Anal. Theory Methods Appl. **48**, 223–246 (2002)
7. Bressan, A.: Hamilton-Jacobi Equations and Optimal Control. An Illustrated Tutorial. NTNU, Trondheim (2001)
8. Canino, A.: On p-convex sets and geodesics. J. Differ. Eq. **75**, 118–157 (1988)
9. Cardaliaguet, P., Dacorogna, B., Gangbo, W., Georgy, N.: Geometric restrictions for the existence of viscosity solutions. Ann. Inst. Henri Poincaré **16**, 189–220 (1999)

10. Chong, Li: On well posed generalized best approximation problems. J. Approx. Theory **107**, 96–108 (2000)
11. Chong, Li, Renxing, Ni: Derivatives of generalized distance functions and existence of generalized nearest points. J. Approx. Theory **115**, 44–55 (2002)
12. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley-Interscience, New York (1983)
13. Clarke, F.H., Stern, R.J., Wolenski, P.R.: Proximal smoothness and the lower-$\mathcal{C}^2$ property. J. Convex Anal. **2**, 117–144 (1995)
14. Clarke, F.H., Ledyaev, Yu.S., Stern, R.J., Wolenski, P.R.: Nonsmooth Analysis and Control Theory. Springer, New York (1998)
15. Colombo, G., Goncharov, V.V.: Variational inequalities and regularity properties of closed sets in Hilbert spaces. J. Convex Anal. **8**, 197–221 (2001)
16. Colombo, G., Goncharov, V.V., Mordukhovich, B.S.: Well-posedness of minimal time problems with constant dynamics in Banach spaces. Set-Valued Var. Anal. **18**, 349–372 (2010)
17. Colombo, G., Thibault, L.: Prox-regular sets and applications. In: Gao, D.Y., Motreano, D. (eds.) Handbook on Nonconvex Analysis. International Press, Boston (2010)
18. Colombo, G., Wolenski, P.R.: Variational analysis for a class of minimal time functions in Hilbert spaces. J. Convex Anal. **11**, 335–361 (2004)
19. Dal Maso, G., Goncharov, V.V., Ornelas, A.: A Lipschitz selection from the set of minimizers of a nonconvex functional of the gradient. Nonlin. Anal. Theory Meth. Appl. **37**, 707–717 (1999)
20. De Blasi, F.S., Myjak, J.: On a generalized best approximation problem. J. Approx. Theory **94**, 54–72 (1998)
21. Ekeland, I.: Nonconvex minimization problems. Bull. Am. Math. Soc. **1**, 443–474 (1979)
22. Ekeland, I., Lebourg, G.: Generic Fréchet differentiability and perturbed optimization problems in Banach spaces. Trans. Am. Math. Soc. **224**, 193–216 (1976)
23. Federer, H.: Curvature measures. Trans. Am. Math. Soc. **93**, 418–491 (1959)
24. Georgiev, P.G.: Submonotone mappings in Banach spaces and applications. Set-Valued Anal. **5**, 1–35 (1997)
25. Goncharov, V.V., Pereira, F.F.: Neighbourhood retractions of nonconvex sets in a Hilbert space via sublinear functionals. J. Convex Anal. **18**, 1–36 (2011)
26. Goncharov, V.V., Pereira, F.F.: Geometric conditions for regularity in a time-minimum problem with constant dynamics. J. Convex Anal. **19**, 631–669 (2012)
27. Goncharov, V.V., Pereira, F.F.: Geometric conditions for regularity of viscosity solution to the simplest Hamilton-Jacobi equation. In: Hömberg, D., Tröltzsch, F. (eds.) Proceedings of the 25th IFIP TC7 Conference on System Modeling and Optimization (CSMO 2011), pp. 245–254. Springer, Berlin (2012)
28. Hiriart-Urruti, J.-B.: Gradients generalises de fonctiones marginales. SIAM J. Contr. Optim. **16**, 301–316 (1978)
29. Ioffe, A.D., Penot, J.-P.: Subdifferentials of performance functions and calculus of coderivatives of set-valued mappings. Serdica Math. J. **22**, 359–384 (1996)
30. Kruzhkov, S.N.: Generalized solutions of the Hamilton-Jacobi equation of the Eikonal type. I. Math. USSR Sbornik **27**, 406–446 (1975)
31. Mordukhovich, B.S.: Variational Analysis and Generalized Differentiation I. Basic Theory. Springer, Berlin (2006)
32. Mordukhovich, B.S., Nam, N.M., Yen, N.D.: Subgradients of marginal functions in parametric mathematical programming. Math. Program. Ser. B. **116**, 369–396 (2009)
33. Ngai, H.V., Penot J.-P.: Approximately convex functions and approximately monotonic operators. Nonlin. Anal. Theory Meth. Appl. **66**, 547–564 (2007)
34. Poliquin, R.A., Rockafellar, R.T., Thibault, L.: Local differentiability of distance functions. Trans. Am. Math. Soc. **352**, 5231–5249 (2000)
35. Rockafellar, R.T.: Generalized directional derivatives and subgradients of nonconvex functions. Can. J. Math. **32**, 257–280 (1980)
36. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)

37. Shapiro, A.: Existence and differentiability of metric projections in Hilbert spaces. SIAM J. Optim. **4**, 130–141 (1994)
38. Thibault, L.: On subdifferentials of optimal value functions. SIAM J. Contr. Optim. **29**, 1019–1036 (1991)
39. Wolenski, P.R., Zhuang, Yu.: Proximal analysis and the minimal time function. SIAM J. Contr. Optim. **36**, 1048–1072 (1998)
40. Yiran, He, Kung, Fu Ng: Subdifferentials of a minimum time function in Banach spaces. J. Math. Anal. Appl. **321**, 896–910 (2006)
41. Zajicek, L.: A generalization of an Ekeland-Lebourg theorem and the differentiability of distance functions. Supl. Rend. del Circolo Mat. Palermo, II **3**, 403–410 (1984)

On Solving Optimization Problems with Hidden Nonconvex Structures

Alexander S. Strekalovsky

1 Introduction

The most of real-life problems, according to Leibnitz and Euler [3, 10, 13], can be stated as optimization problems, because the lows of the nature just follow the principles of Fermat, Lagrange, Euler, and other equations provided by extremum principles.

On the other hand, the contemporary situation can be characterized by the crucial impact and the increasing value of the numerical methods in view of computational solving the problems of practical interest.

It is worthy to note that the optimization problems must be separated into two parts: convex and nonconvex. From the viewpoint of the numerical processing of the problem of a rather general kind

$$\begin{cases} f_0(x) \downarrow \min\limits_{x}, \quad x \in S \subset \mathbb{R}^n, \\ f_i(x) \leq 0, \quad i = 1, 2, \ldots, m, \end{cases} \tag{$\mathcal{P}_0$}$$

there exists a "solvable case"—this one of the convex optimization problems, those where the domain S and the functions f_0 and f_i are all convex [3, 13, 22, 40].

Under minimal additional computability assumptions a convex optimization problem is computationally tractable [3, 22]. It means that the computational effort required to solve the problem to a given accuracy grows moderately with the dimension of the problem and the required number of accuracy digits.

A.S. Strekalovsky (✉)
Institute for System Dynamics and Control Theory of SB RAS,
Lermontov St. 134, Irkutsk, 664033, Russia
e-mail: strekal@icc.ru

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_23,

In contrast to this, general-type nonconvex problems are too difficult for numerical solution, since in a real-life nonconvex optimization problem there can exist a lot (often a huge!) of local extrema and stationary points which are rather far from a global solution [9, 17, 25, 28].

As a consequence, the classical optimization methods (conjugate gradients, Newton's and quasi-Newton's methods, TRM, SQP, IPM, etc.) turn out to be inoperative, in general, and ineffective as to finding a global solution in nonconvex problems because they are not able to escape a local pit.

Moreover, specialists in applied problems do not think about the correctness of direct application of classical optimization methods in nonconvex problems, while the numerical results are interpreted only in the content aspect, without thinking of the fact that all classical optimization methods converge to the global solution only in convex problems [3, 22].

At the same time, in nonconvex problems, the direct application of standard methods may have unpredictable consequences [3, 17, 24, 28, 40], and sometimes may even distract one from the desired solution. So, these arise various approaches which completely neglect the classical optimization methods and use a direct selection way employing, for example, the B&B idea or cut's method. As well known the latter algorithms suffer the curse of dimension, when the volume of computations grows exponentially side by side with the growth of the problem's dimension [17]. We are sure, there exists also another way of solving nonconvex problems of high dimension [18, 19, 26–37].

In the recent two decades, we have managed to construct a theory of global search, which is harmonic from the viewpoint of the theory of optimization and which unexpectedly has turned out to be rather efficient in the aspect of computations, especially for the problems of high dimensions. Simultaneously necessary and sufficient Global Optimality Conditions (GOCs) for the principal classes of nonconvex problems can be viewed as the kernel of the theory (see below) [28].

Furthermore, we have proposed a family of local search methods (LSMs), which, on the one hand, in some cases develop methods earlier known for the special problems and, on the other hand, this family of LSMs represents a joint ensemble of methods, which is harmonic from the viewpoint of GOCs [28, 31, 33].

Moreover, the procedures of escape from stationary or local solutions, which are based on GOCs, are unique and quite efficient even in case of any simplest implementation [18, 19, 26–28, 31–37].

Besides, the approach elaborated has been tested on a wide field of popular nonconvex problems (some part of which is represented below). It has demonstrated an unexpected efficiency during the numerical solving problems of high dimension. Note, convex optimization methods are successfully used "inside" the procedures of local and global search proposed [18, 19, 26–28, 31–37].

Finally, we have to add that, according to the opinion of numerous confirmed specialists in optimization, the most attractive and promising fields of investigation and, may be, even modelling paradigms in optimization in twenty-first century can be represented (see [24]), in particular, by the following examples which both possess the hidden nonconvex structures:

- the search for equilibriums in competitions (conflict situations or games);
- hierarchical optimization problems.

Unexpectedly for us, we turned out to be on this main stream, but rather prepared, i.e. possessing a suitable mathematical apparatus.

2 Examples of Applied Problems

2.1 Linear Complementarity Problems

As well known [7], the linear complementary problem (LCP) aims at finding the pair of vector $(x,w) \in \mathbb{R}^{n+n}$, which satisfy the following conditions:

$$\left.\begin{array}{c} Mx+q=w, \quad \langle x,w\rangle = 0, \\ x \geq 0, \quad w \geq 0, \end{array}\right\} \tag{1}$$

for a given vector $q \in \mathbb{R}^n$ and a given real $(n \times n)$-matrix M which is, in general, indefinite. Many physical, engineering problems (the braking problem; the problem of contact; the problem of viscoelastic twisting, etc.), some economic problems (the problems of market equilibrium, the problem of optimal constant basic capital, etc.) and problems of computational geometry can often be stated as LSP. From the first glance any nonconvexity is not visible in (1). Even if we consider a similar formulation of LSP, for example,

$$\left.\begin{array}{c} \langle x, Mx+q\rangle = 0, \quad M = M^\top, \\ x \geq 0, \quad Mx+q \geq 0, \end{array}\right\} \tag{1$'$}$$

a nonconvexity does not appear since all data stays to be linear. However, if one looks at the problem as optimization problem:

$$\left.\begin{array}{c} \Phi(x) := \langle x, Mx\rangle + \langle x,q\rangle \downarrow \min, \quad x \in S, \\ S \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid x \geq 0, \ Mx+q \geq 0\}, \end{array}\right\} \tag{2}$$

it becomes clear that the properties and the structure of the LSP (1) depend on the features of the matrix M, as follows.

(a) If M is nonnegative definite, the problem (2) is convex, i.e. solvable with the classical methods, for instance, the conjugate gradient method (CGM).
(b) If M is negative definite, then the problem (2) turns out to be nonconvex (anticonvex) optimization problem of concave minimization (that is equivalent to convex maximization).
(c) If M is indefinite, i.e. it has positive and negative eigenvalues, then the problem (2) must be classified as a d.c. minimization problem:

$$\Phi(x) = g(x) - h(x) \downarrow \min_x, \quad x \in S, \tag{3}$$

where $g(x) = \langle x, M_1 x\rangle + \langle q, x\rangle$, $M_1 = M_1^\top > 0$, $h(x) = \langle x, M_2 x\rangle$, $M_2 = M_2^\top > 0$, $M = M_1 - M_2$, $g(\cdot)$ and $h(\cdot)$ are strongly convex functions on $\mathbb{R}^n$. In each of cases (a), (b), and (c) one has to apply the different methods of local and global search in order to find a global solution to Problem (2) or, what is equivalent, to find a solution to Problem (1).

The conclusion is unexpected: if anyone needs to have a solution to the LCP (1), then he has to choose the only one way (global search method (GSM)) among three different paths (GS methods) dependent on the properties of the matrix M. Below we will show how to do it.

So, very simple, from the first sight, Problem (1) may turn out to be very difficult to solve, since it possesses, in general, a hidden nonconvexity.

2.2 Search for an Equilibrium

As example of equilibrium problems, let us consider the bimatrix games [33] which reflect the conflict of two parties (players), each one having a finite number of strategies. After having introduced the mixed strategies, we obtain

$$\left.\begin{array}{c} \langle x, Ay\rangle \uparrow \max\limits_{x}, \quad x \in S_m, \\ \langle x, By\rangle \uparrow \max\limits_{y}, \quad y \in S_n, \\ S_p = \left\{ x \in \mathbb{R}^p_+ \mid \sum\limits_{i=1}^{p} x_i = 1 \right\}, \quad p = m, n. \end{array}\right\} \tag{4}$$

Some economics, engineering and ecological problems can be represented in the form of bimatrix games, in which the Nash equilibrium is the common concept and can be represented as follows: find an equilibrium situation $(x^*, y^*) \in S_m \times S_n$:

$$\left.\begin{array}{ll} \langle x^*, Ay* \rangle \geq \langle x, Ay^* \rangle & x \in S_m, \\ \langle x^*, By^* \rangle \geq \langle x^*, By \rangle & y \in S_n. \end{array}\right\} \tag{5}$$

In formulae (4) and (5), from the first sight, any nonconvexity also is not yet visible, since all the data is linear, and the problem (4)–(5) seems to be convex, i.e. solvable by the classical methods and approaches.

However, it turns out that the search for the Nash equilibrium can be reduced [20, 21] to solving the following nonconvex (in general) problem of mathematical programming:

$$\left.\begin{array}{c} F(x, y, \alpha, \beta) := \langle x, (A + B)y\rangle - \alpha - \beta \uparrow \max, \\ x^\top B - \beta e_n \leq 0_n, \quad x \in S_m, \\ Ay - \alpha e_m \leq 0_m, \quad y \in S_n, \end{array}\right\} \tag{6}$$

where $\alpha, \beta \in \mathbb{R}$, $e_p = (1, 1, \ldots, 1)^\top \in \mathbb{R}^p$, $p = m, n$. Note that the numbers α_* and β_* in a global solution $(x^*, y^*, \alpha_*, \beta_*)$ to Problem (6) are the optimal profits of the first

and second players, respectively, in the game (4)–(5), while the pair (x^*, y^*) turns out to be just a Nash equilibrium point in the game (4)–(5).

On account of the formulation (6) it becomes clear that a way (method) of finding a Nash point strongly depends on the properties of the matrix $(A+B)$.

So, the conclusion is obvious here and consists in the fact that the initial statement (4)–(5) of a bimatrix game is deceptive in the sense that it has, in general, a hidden (implicit) nonconvexity.

2.3 Hierarchical Optimization Problems

Hierarchical problems are encountered in practice because of impossibility of accumulation of the total available information at the upper level in the process of investigation of structurally complex control systems (social, economic, ecological-economic ones, etc.) and, as a consequence, possess some hidden nonconvexity generated by just hierarchical structures.

For example, the financial systems in the economic power countries are usually constructed as bilevel systems. Besides, the electric energy system in USSR was organized as a four-level system.

As to mathematical aspects of the statement, problems of bilevel programming represent extremum problems, that side by side with standard constraints which are expressed in terms of equalities and inequalities, include the constraints described with the aid of optimization subproblem representing the lower level of the bilevel problem (or the player called the follower in difference with the player of the upper level called the leader).

To begin with, let us consider the linear bilevel problem

$$(\mathcal{LBP}): \quad \begin{cases} F(x,y) := \langle c,x\rangle + \langle d,y\rangle \downarrow \min\limits_{x,y}, \\ x \in X = \{x \in \mathbb{R}^m \mid Ax \le b\}, \\ y \in Y_*(x) := \operatorname{Arg}\min\limits_{y}\{\langle d_1,y\rangle \mid y \in Y(x)\}, \\ Y(x) = \{y \in \mathbb{R}^n \mid A_1x + B_1y \le b_1\}, \end{cases}$$

where $c \in \mathbb{R}^m$, $d, d_1 \in \mathbb{R}^n$, $b \in \mathbb{R}^p$, $b_1 \in \mathbb{R}^q$, and A, A_1, B_1 are matrices of corresponding dimensions. Suppose, that
$(\mathcal{H}_1)$: the function $F(x,y)$ is bounded below on the nonempty set Z,

$$Z := \{x \in \mathbb{R}^m,\ y \in \mathbb{R}^n \mid Ax \le b,\ A_1x + B_1y \le b_1\};$$

$(\mathcal{H}_2)$: the function $\langle d_1,y\rangle$ is bounded below on the set $Y(x)$ for all $x \in X$.
Even in this very simple case it is easy to construct an example showing the nonconvexity of the problem $(\mathcal{LBP})$.

Example 1. ([8]) Consider the problem

$$\left.\begin{array}{l} F(x,y) = x + 3y \downarrow \min\limits_{x,y}, \quad x,y \in \mathbb{R}, \\ 1 \le x \le 6, \quad y \in Y_*(x) = Sol(\mathcal{P}_L), \end{array}\right\} \qquad (\mathcal{LBP}_1)$$

$$(\mathcal{P}_L): \quad \left\{\begin{array}{c} f(y) = -y \downarrow \min\limits_{y}, \\ x + y \le 8, \quad x + 4y \ge 8, \\ x + 2y \le 13. \end{array}\right.$$

Regardless of the convexity of the set

$$Z = \{\, (x,y) \in \mathbb{R}^2 \mid 1 \le x \le 6,\ x + y \le 8,\ x + 4y \ge 8,\ x + 2y \le 13 \,\},$$

it is easy to see even geometrically that the set

$$Z_* = \{\, (x,y) \in Z \mid y \in Y_*(x) \,\}$$

is nonconvex which provides for the nonconvexity in the problem $(\mathcal{LBP}_1)$. □

So, the Example 1 shows the importance of the preliminary theoretical study of hierarchical optimization problems and, as, may be, the simplest case of $(\mathcal{BP})$, the $(\mathcal{LBP})$.

2.4 Problems of Financial and Medical Diagnostics

Such problems are well known as applied ones, and on the other hand, these problems are often interpreted as the problems of generalized separability. For example, if the two sets of points $\mathcal{A}$ and $\mathcal{B}$ are characterized by the matrices $A = [a^1, \ldots, a^M]$, $B = [b^1, \ldots, b^N]$, $a^i, b^j \in \mathbb{R}^n$, then the problem of polyhedral separability may be reduced to the problem of minimization of the nonconvex nondifferentiable error function ($V = (v^p), \Gamma = (\gamma_p),\ \gamma_p \in \mathbb{R},\ v^p \in \mathbb{R}^n,\ p = 1, \ldots, P$)

$$F(V,\Gamma) = F_1(V,\Gamma) + F_2(V,\Gamma), \qquad (7)$$

$$\left.\begin{array}{l} F_1(V,\Gamma) = \dfrac{1}{M} \displaystyle\sum_{i=1}^{M} \max\{0; \max\limits_{1 \le p \le P} (\langle a^i, v^p \rangle - \gamma_p + 1)\}, \\ F_2(V,\Gamma) = \dfrac{1}{N} \displaystyle\sum_{j=1}^{N} \max\{0; \min\limits_{1 \le p \le P} (-\langle b^i, v^p \rangle + \gamma_p + 1)\}. \end{array}\right\} \qquad (8)$$

In this problem it is also not clear with which kind of nonconvexity we are dealing and how to overcome not only the nonsmoothness of the problem but also a nonconvexity generated apparently by $F_2(\cdot)$.

But, anyway, the question arises how to attack optimization problems with a hidden or an explicit nonconvexities.

3 Optimization Problems with the Functions of A.D.Alexandrov

The targets of our presentation can be bounded by consideration of the class $DC(\mathbb{R}^n)$ of the functions $f(\cdot)$ which can be represented as the difference of two convex functions (d.c. functions). The class was, for the first time, introduced in 1934 [1, 2] by the Russian mathematician A.D.Alexandrov, the member of AS of USSR.

Nowadays, this class is viewed by the specialists [9, 17, 25, 39] to be rather wise for consideration. Furthermore, the $DC(\mathbb{R}^n)$ possess several remarkable properties.

(a) The set $DC(\mathbb{R}^n)$ is generated by the well-studied class—the convex cone of convex functions and forms a linear space [12, 13, 17, 28, 39].
(b) $DC(\mathbb{R}^n)$ includes the well-known classes such as twice differentiable functions, power and trigonometric polynomials [12, 13, 17, 39].
(c) Any continuous function on a compact set $\mathcal{K} \subset \mathbb{R}^n$ can be approximated at any desired accuracy (in the topology of homogeneous convergence) by a function from $DC(\mathcal{K})$ [12]. Consequently, any optimization problem with continuous functions can be approximated at any desired accuracy by an extremum problem with functions of A.D.Alexandrov.

Only note, that, if $f(\cdot)$ is a d.c. function, then there exists an infinite number of d.c. representations of the form (3), for example, in the form of difference of strongly convex functions.

Closedness of the set $DC(\mathbb{R}^n)$ of functions of A.D.Alexandrov with respect to the majority of operations, which are used in optimization, is also essential from the optimization viewpoint. For example, a sum, a difference, the module, the maximum, the minimum, etc. of the family of d.c. functions occur also in the class $DC(\mathbb{R}^n)$.

Besides, the number of problems with d.c. functions is so large that the majority of the specialists, who have a long-time experience of solving problems of d.c. programming are sure [9, 12, 13, 17, 39] that all (or almost all) nonconvex optimization problems turn out to be really d.c. problems.

In this connection, the following statement of optimization problem can be viewed as rather general:

$$\left.\begin{array}{l} f_0(x) = g_0(x) - h_0(x) \downarrow \min\limits_x, \ \ x \in S, \\ f_i(x) = g_i(x) - h_i(x) \le 0, \ \ i = 1, \ldots, m; \\ f_j(x) = g_j(x) - h_j(x) = 0, \ \ j = 1, \ldots, N. \end{array}\right\} \tag{9}$$

Here g_i, g_j, h_i, h_j are convex functions and S is convex set from $\mathbb{R}^n$.

Apparently, almost all the specialists in optimization areas could estimate Problem (9) as very difficult and unsolvable by the existing approaches and methods even for the case of middle dimension (say, $n = 100, \ldots, 1{,}000$.)

Actually, even very simple (from the viewpoint of Problem (9)) the convex maximization quadratic problem on a box (which is a very particular case of (9)):

$$\left.\begin{array}{l} h(x) = \frac{1}{2}\langle x, Qx\rangle \uparrow \max\limits_{x}, \;\; Q = Q^T > 0, \\ x \in S := \Pi = \{x \in \mathbb{R} \mid \alpha_i \leq x_i \leq \beta_i \;\; i = 1, \ldots, n\} \end{array}\right\} \tag{10}$$

is proved to be NP-hard [9]. Therefore, to begin with, let us simplify the situation and start with rather simple (from the first glance) nonconvex optimization problems.

1. *D.C. minimization*

$$(\mathcal{P}): \qquad f(x) = g(x) - h(x) \downarrow \min, \;\; x \in D, \tag{11}$$

   where $g(\cdot)$, $h(\cdot)$ are convex functions, and D is a convex set, $D \subset \mathbb{R}^n$.
2. *D.C. constraint problem*

$$(\mathcal{DCC}): \qquad \left.\begin{array}{l} f_0(x) \downarrow \min\limits_{x}, \; x \in S, \\ F(x) = g(x) - h(x) \leq 0, \end{array}\right\} \tag{12}$$

   where $g(\cdot)$ and $h(\cdot)$ are as above, $S \subset I\!R^n$, $f_0(\cdot)$ is a continuous function.
3. *Convex maximization*

$$h(x) \uparrow \max, \;\; x \in D, \tag{13}$$

   (when $g \equiv 0$ in (11)).
4. *Reverse-convex constraint problem*

$$\left.\begin{array}{l} f_0(x) \downarrow \min, \; x \in S, \\ h(x) \geq 0, \end{array}\right\} \tag{14}$$

   ($g \equiv 0$ in (12)).

Note, that any quadratic optimization problem with arbitrary matrices occurs in the classification (11)–(14) or takes the form (9).

4 Global Search Methodology

Since in our approach the general global search procedure includes two principal parts:

(a) local search;
(b) procedures of escaping a critical point provided by a LSM; we are going, first, to consider special (for each class of d.c. programming problems) LSMs.

4.1 Local Search

The ideas of the most of LSM are rather simple and may consist in the consecutive solution of the (partially) linearized problems which for Problems (11)–(14) turn out to be convex. As a consequence, it becomes possible to apply classical convex optimization methods (Newtonians, CGM, TRM, etc.) in order to find a solution to linearized problems, i.e. within the framework of Local Search Schemes.

So, unlike that in well-known methods of the so-called Global Optimization (such as B&B, cuts methods), which, say, "deny and ignore" the modern and classical optimization methods, we insist on the obligatory, but "indirect" application of these methods.

For example, as regards the problem of d.c. minimization $(\mathcal{P})$–(11), the basic element, the "cornerstone" of the Global and LSMs is solving the following (linearized at a current iteration point $x^s \in D$) convex problem

$$(\mathcal{PL}_s): \qquad \Phi_s(x) := g(x) - \langle h'(x^s), x\rangle \downarrow \min_x, \quad x \in D, \tag{15}$$

where $h'(x^s) = h'_s \in \partial h(x^s),\ s = 1,2,\ldots$ is a subgradient of the convex function $h(\cdot)$ at the point x^s [13]. It is clear that in the differentiable case h'_s coincides with the usual gradient ∇h_s [13].

Furthermore, the LSM itself for $(\mathcal{P})$–(11) may consist in the consecutive solving (likewise in the method of "direct iterations") Problems $(\mathcal{PL}_s)$–(15). More precisely, given $x^s \in D$, we can find $x^{s+1} \in D$ as an approximate solution to $(\mathcal{PL}_s)$ by means of some suitable convex optimization method (for example, BFGS), or one of the packages of applied software (Xpress-MP, IBM CPLEX etc).

So, we produce the sequence $\{x^s\}$ according to the inequality:

$$\Phi_s(x^{s+1}) := g(x^{s+1}) - \langle h'(x^s), x^{s+1}\rangle \ \le\ \inf_x\{g(x) - \langle h'(x^s), x\rangle \mid x \in D\} + \delta_s \tag{16}$$

where the sequence $\{\delta_s\}$ fulfils the condition

$$\sum_{s=0}^{\infty} \delta_s < +\infty,\ \ \delta_s > 0,\ \ s = 1,2,\ldots.$$

It was rather surprising that the process in this case converges in the following sense.

Theorem 1. *Suppose the cost function of Problem $(\mathcal{P})$–(11) is bounded below, so that*

$$\mathcal{V}(\mathcal{P}) := \inf(f, D) \stackrel{\triangle}{=} \inf_x\{f(x) \mid x \in D\} > -\infty.$$

Then the sequence $\{x^s\} \in D$ generated by the rule (16) satisfies the following conditions.

(a) The number sequence $\{f_s\},\ f_s = f(x^s)$ converges in the sense, as follows:

$$\lim_{s\to\infty} f_s = f_* \ge \mathcal{V}(\mathcal{P}). \tag{17}$$

(b) $$\lim_{s\to\infty}[\inf_{x}\{g(x)-g(x^{s+1})+\langle h'(x^s),x^{s+1}-x^s\rangle \mid x\in D\}]=0. \tag{18}$$

or, what is the same (see $(\mathcal{PL}_s)$*–(15)),*

$$\lim_{s\to\infty}[\mathcal{V}(\mathcal{PL}_s)-\Phi_s(x^{s+1})]=0, \tag{18'}$$

where

$$\mathcal{V}_s:=\mathcal{V}(\mathcal{PL}_s):=\inf_{x}\{g(x)-\langle h'(x^s),x\rangle \mid x\in D\} \tag{19}$$

is the optimal value of the linearized problem $(\mathcal{PL}_s)$*–(15)*

(c) If the function $h(\cdot)$ *in (15) is strongly convex, then we have*

$$\lim_{s\to\infty}\|x^s-x^{s+1}\|=0. \tag{20}$$

(d) Any limit point x_* *of the sequence* $\{x^s\}$ *generated by LSM (16) is a solution of the following linearized problem*

$$(\mathcal{PL}_*):\qquad \Phi_*(x):=g(x)-\langle y_*,x\rangle\downarrow\min_{x},\quad x\in D, \tag{21}$$

where $y_*=h'(x_*)\in\partial h(x_*)$.

Note that very frequently for small dimension $(n\le 7,8,10)$ cases LSM (16) provides for a global solution to $(\mathcal{P})$–(11).

It is interesting that historically the particular case ($g\equiv 0$) of LSM (16) for differentiable convex maximization problem (13) has been proposed by Bulatov in 1969 [4] and can be represented in the modern form as follows:

$$\langle h'(x^s),x^{s+1}\rangle+\delta_s\ge\sup_{x}\{\langle h'(x^s),x\rangle \mid x\in D\}. \tag{22}$$

Besides, the well-known "power" method for finding the maximal eigenvalue of a symmetric positive definite matrix A, or what is the same, for solving the problem [38]

$$\langle x,Ax\rangle\uparrow\max_{x},\quad \|x\|\le 1, \tag{23}$$

turns out to be very particular case of LSM (22), when the linearized problem $(\mathcal{PL}_s)$ (with $g\equiv 0$) can be solved analytically. Note that the method (22) in Problem (23) converges to the global solution [38].

Thus, one can conclude that the idea of linearization with respect to the basic nonconvexity of a nonconvex problem has certainly some age. Anyway, it is worth noting to mention the works of the group of Pham Dinh Tao in which the idea of linearization with respect to the basic nonconvexity also demonstrated its effectiveness [14–16].

Furthermore, special methods of local search have been developed for the problems with d.c. constraints (12), (14) (see [31]). These methods have also been grounded, for example, on considering linearized problems of the form

$$\left.\begin{array}{c} g(x) - \langle h'(x^s), x\rangle \downarrow \min\limits_{x}, \\ x \in S, \quad f_0(x) \le \zeta_s = f_0(x^s), \end{array}\right\} \tag{24}$$

and the duality of Tuy [39].

4.2 Global Optimality Conditions

The second step in the global search methodology can be viewed as the most important one and even crucial, because the question is how to escape a critical point (provided by an LSM and that is not a global solution).

Such a procedure is substantiated by the theoretical basis produced with the help of the so-called GOC which for the case of d.c. minimization problem $(\mathcal{P})$–(11) takes the following form.

Theorem 2. *If z is a global solution to $(\mathcal{P})$, $z \in Sol(\mathcal{P})$, $\zeta := f(z)$, then*

$$(\mathcal{E}): \qquad \left\{\begin{array}{l} \forall (y,\beta) \in \mathbb{R}^n \times \mathbb{R}: \quad h(y) = \beta - \zeta, \\ g(x) - \beta \ge \langle h'(y), x - y\rangle \quad \forall x \in D. \end{array}\right. \tag{25}$$

Proof. Suppose, for some pair (y,β) satisfying (25) and a feasible point $\hat{x} \in D$ the inequality in (25) is violated

$$g(\hat{x}) < \beta + \langle h'(y), \hat{x} - y\rangle.$$

Then due to convexity of $h(\cdot)$ we have

$$f(\hat{x}) \stackrel{\triangle}{=} g(\hat{x}) - h(\hat{x}) < h(y) + \zeta - h(y) = f(z),$$

or $f(\hat{x}) < f(z)$. Thus, $\hat{x}$ is "better" than z, which contradicts to $z \in Sol(\mathcal{P})$. □

So, when selecting the "perturbation parameters" (y,β) satisfying (25) and solving the linearized problem (sf. (15))

$$\Phi_y(x) := g(x) - \langle h'(y), x\rangle \downarrow \min_{x}, \quad x \in D, \tag{26}$$

(where $y \in \mathbb{R}^n$ is not obligatory feasible!) we obtain a family of starting points $x(y,\beta)$ for a further (assume) local search.

Moreover, on each level $\zeta_k = f(z^k)$ it is not necessary to investigate all the pairs (y,β) satisfying (25), $\zeta_k = \beta - h(y)$, but it is sufficient to discover the violation of the variational inequality (25) only for one pair $(\hat{y},\hat{\beta})$.

After that, one proceeds to the next iteration of the global search: $z^{k+1} := \hat{x}$, $\zeta_{k+1} := f(z^{k+1})$, and starts the procedure from the very beginning. So, the idea of the GSM becomes considerably more clear.

For the case of d.c. constraint problem (12) the character of GOC is a little bit different. It particular, the necessary conditions are rather far from the sufficient ones. More precisely, we have the result as follows:

Theorem 3. *Assume that in Problem (DCC)–(12) the following condition holds:*

$$(\mathcal{G}): \qquad \left.\begin{array}{l} \text{there does not exist a solution } x_* \in S \\ \text{to Problem (12) such that } F(x_*) < 0. \end{array}\right\} \tag{27}$$

If a point $z \in S$ is a global solution to Problem (12) such that $F(z) = 0$, then

$$(\mathcal{E}_1): \qquad \left\{\begin{array}{l} \forall (y,\beta) \in \mathbb{R}^n \times \mathbb{R}: \quad \beta = h(y), \ \ \forall h'(y) \in \partial h(y) \\ g(x) - \beta \geq \langle h'(y), x - y \rangle \ \ \forall x \in S, \ \ f_0(x) \leq f_0(z). \end{array}\right. \tag{28}$$

Proof. Suppose we find some parameters (y_0, β_0), $h'(y_0) \in \partial h(y_0)$ and a point $x_0 \in S$ such that

$$\beta_0 = h(y_0), \quad f_0(x_0) \leq f_0(z), \text{ and } g(x_0) - \beta_0 < \langle h'(y_0), x_0 - y_0 \rangle.$$

Then due to convexity of $h(\cdot)$ we obtain

$$0 < \beta_0 - g(x_0) + h(x_0) - h(y_0) = -F(x_0).$$

Hence, we have the feasible point $x_0 \in S$, $F(x_0) < 0 = F(z)$ with the property $f_0(x_0) \leq f_0(z)$. It means that x_0 is a solution to Problem (12) as well as the point z. The latter contradicts to the condition $(\mathcal{G})$–(27). □

A procedure of escaping a local pit can be conducted in a similar manner as it was explained after Theorem 2.

In the next subsection such a procedure will be precised for the case of d.c. minimization problem (11).

4.3 Global Search Methods

In order to deal with nonconvex optimization problems and, in addition, on the basis of the rather large computational experience [11, 18, 19, 26–28, 31–37] we propose three principles on which can be produced a search for a global solution to d.c. optimization problems of the forms considered above.

1. Linearization with respect to the basic nonconvexities of the problem under scrutiny and, consequently, the reduction of the original problem to a family of (partially) linearized problems.
2. Application of contemporary convex optimization methods for solving linearized problems and, as a consequence, "within" special LSMs.
3. Construction of "good" approximations (resolving sets) of the level surfaces/epigraph boundaries of convex functions.

Moreover, by working rather long time (about 30 years) on the field of nonconvex optimization, several practical rules have been elaborated, which can be represented as follows:

1. Never apply convex optimization methods directly.
2. Exact classification of the problem under scrutiny.
3. Application of special (for the class of problems to which belongs your problem) LSM, or (your problem's) specific methods.
4. Application of GSM specialized for the class which includes your problem.
5. Construction of suitable approximations of level surfaces (and the boundaries of the epigraphs) of convex functions with the aid of the experience obtained during solving similar problems.
6. Application of convex optimization methods for solving linearized problems and within the framework of special LSM.

These rules may be explained otherwise and by examples, following the instances.

1. Never apply CGM or BFGS if you are not convinced that your problem is convex.
2. Try to separate the data of your problem into two parts—convex and anticonvex.

For example, dealing with a quadratic function of the kind

$$q(x) = \frac{1}{2}\langle x, Qx\rangle,$$

where the matrix Q is indefinite, you have to separate the matrix $Q(n \times n)$ into a difference $Q = Q_1 - Q_2$ of two symmetric positive definite matrices $Q_i = Q_i^T > 0$, $i = 1, 2$. Note there exists infinity of such representations and several methods and ways to obtain its [28, 38].

Further, it is very important where your quadratic function $q(x)$ is situated—in the objective function or among the constraint's data, because depending on the situation you have different types of the problem to solve—d.c. minimization (11) or d.c. constraint problem (12), respectively. And as a consequence, you have to follow the different strategy (GSM, see below).

To demonstrate the effectiveness of these practical rules, let us consider the following example.

Example 2 (Incorrect Classification). Consider the problem

$$\left.\begin{aligned} \varphi(x) = \sum_{i=1}^{n} \ln(1 + x_i) \downarrow \min_x, \\ x \in \Pi = \left\{x \in \mathbb{R}^n \mid -\tfrac{1}{2} \le x_i \le 3\right\} \subset \mathbb{R}^n. \end{aligned}\right\} \tag{29}$$

Obviously, the point $z = \left(-\frac{1}{2}, \ldots, -\frac{1}{2}\right)^T$ is the solution to the problem. Suppose, the current iterate is $x^k = (0, \ldots, 0)^T$

$$\nabla f(x) = \left(\frac{1}{1+x_1}, \ldots, \frac{1}{1+x_n}\right)^T, \quad \nabla f(x^k) = (1, \ldots, 1)^T,$$

$$\nabla^2 f(x) = \begin{bmatrix} -\frac{1}{(1+x_1)^2} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & -\frac{1}{(1+x_n)^2} \end{bmatrix}, \quad \nabla^2 f(x^k) = \begin{bmatrix} -1 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & -1 \end{bmatrix}.$$

The auxiliary problem of Newton's method

$$\Phi(d) = \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} d_i^2 \downarrow \min, \quad d \in \Pi = (\Pi - x^k)$$

has obviously the solution $d = (3, \ldots, 3)^T$ (take the case $n = 2$), which is a direction to the worst feasible point $x = (3, \ldots, 3)^T$. As a consequence, the iteration of the line-search method $x^{k+1} = x^k + td^k$ cannot escape from $x^k = (0, \ldots, 0)^T$ and the process is stopped at x^k. Besides, note that the auxiliary problem conserves the nonconvex character of the original problem (29).

In contrast to the incorrectness of the above classification, let us look at the goal function of Problem (29) as a concave function, and correspondingly at the problem (29) as the concave minimization problem. Then we immediately conclude that we are dealing just with Problem (13) ($h(\cdot) = -\varphi(\cdot)$). Hence, we have to apply, first, the special LSM (16) where $g(x) \equiv 0$. So, beginning at arbitrary feasible point $x^0 \in \Pi$, we have to solve the linearized problem

$$(\mathcal{PL}_0): \quad \langle \nabla \varphi(x^0), x \rangle = -\langle \nabla h(x^0), x \rangle \downarrow \min_x, \quad x \in \Pi \subset \mathbb{R}^n,$$

i.e.

$$\sum_{i=1}^{n} \frac{1}{1+x_i^0} \cdot x_i \downarrow \min_x \quad -\frac{1}{2} \leq x_i \leq 3, \quad i = 1, \ldots, n$$

that provides for the global solution to the original problem (29)

$$x^1 = z \stackrel{\triangle}{=} \left(-\frac{1}{2}, \ldots, -\frac{1}{2}\right)^T \in Sol(\mathcal{P}).$$

So, the special LSM (in one step!) has found the global solution to Problem (29).

□

Let us return now to the construction of a GSM (strategy) based on GOC presented in Theorem 2 and specialized only for Problem $(\mathcal{P})$–(11).

The basic stages of such a GSM (strategy) can be described as follows:

I. Find a critical point z by means of the special LSM ((16), for example).
II. Choose a number $\beta \in [\beta_-, \beta_+]$, where $\beta_- = \inf(g, D)$, $\beta_+ = \sup(g, D)$ can be approximated by rather rough estimates.
III. Construct an approximation

$$\mathcal{A}(\beta) = \{y^1, \ldots, y^N \mid h(y^i) = \beta - \zeta,\ i = 1, \ldots, N = N(\beta)\}$$

of the level surface of the function $h(\cdot)$.

IV. Beginning at every point y^i of the approximation $\mathcal{A}(\beta)$ find a feasible point u^i by means of the special local search algorithm (16).

V. Verify the VI (25) from GOC

$$g(u^i) - \beta \geq \langle h'(w^i), u^i - w^i \rangle \quad \forall i = 1, \ldots, N, \tag{30}$$

where w^i may be found as the projection of the point u^i onto the convex set

$$\mathcal{L}(h, \beta - \zeta) = \{x \in \mathbb{R}^n \mid h(x) \leq \beta - \zeta\}.$$

VI. If $\exists j \in \{1, \ldots, N\}$ such that (30) is violated, then set $x^{k+1} := u^j$ and return to Stage I. Otherwise change β and return to Stage III.

Example 3. Consider the problem

$$f(x) \downarrow \min, \quad x \in \mathbb{R}, \tag{31}$$

$$f(x) = \begin{cases} \frac{1}{4}x^4 - \frac{1}{2}x^2, & x \geq 0, \\ \frac{1}{2}x^4 - x^2, & x < 0. \end{cases} \tag{32}$$

The d.c. representation is here obvious, for example

$$f(x) = g(x) - h(x),$$

where

$$g(x) = \begin{cases} \frac{1}{4}x^4, & x \geq 0, \\ \frac{1}{2}x^4, & x < 0, \end{cases} \qquad h(x) = \begin{cases} \frac{1}{2}x^2, & x \geq 0, \\ x^2, & x < 0. \end{cases} \tag{33}$$

Let us choose the starting point $x_0 = 100$, while the global solution is $z = -1$, which can be readily seen.

(A) *Local search*

We have $s = 0$, $x_0 = 100$, $\nabla h(x_0) = x_0 = 100$, since

$$\nabla h(x) = \begin{cases} \nabla h_1(x) = x, & x \geq 0, \\ \nabla h_2(x) = 2x, & x < 0. \end{cases} \tag{34}$$

Then the linearized (convex) problem $(\mathcal{PL}_0)$–(15) takes the form

$$(\mathcal{PL}_0): \quad \Phi_0(x) = g(x) - \langle \nabla h(x_0), x \rangle = \frac{1}{4}x^4 - 100x \downarrow \min_x, \quad x \in \mathbb{R}.$$

To simplify the situation, in order to solve (PL_s) let us apply the Fermat rule instead of any numerical method. This yields

$$\nabla \Phi(x) = x^3 - 100 = 0.$$

Taking into account that $4^3 = 64$, $5^3 = 125$, let us risk to set $x_1 \approx 4.5$. Further, we have to solve the next linearized problem $(s = 1)$

$$(\mathcal{PL}_1): \qquad \Phi_1(x) = \frac{1}{4}x^4 - 4.5x \downarrow \min_x, \quad x \in \mathbb{R},$$

and, as a consequence, the equation

$$\nabla \Phi_1(x) = x^3 - 4.5 = 0,$$

whence it follows that $x_2 \approx 1.7$.

For $s = 2$ consider the next linearized problem

$$(\mathcal{PL}_2): \qquad \Phi_2(x) = \frac{1}{4}x^4 - 1.7x \downarrow \min_x, \quad x \in \mathbb{R},$$

and the corresponding equation

$$x^3 - 1,7 = 0,$$

that provides for the solution $x_3 \approx 1.2$. Hence, it is clear that $\{x^s\}$ tends to $z^0 = 1 \in \text{Arglocmin}$ (31).

(B) *Global search.*

Step 1. Thus, beginning at $x_0 = 100$ LSM provided for the point $z^0 = 1$, $\zeta_0 := f(z^0) = -\frac{1}{4}$.

Step 2. To begin with let us choose $\beta_0 = g(z^0) = g(1) = \frac{1}{4}$.

Step 3. Now we need to construct an approximation

$$\mathcal{A}_0 = \left\{ y^1, y^2, \ldots, y^N \mid h(y^i) = \beta_0 - \zeta_0 = \frac{1}{4} - \left(-\frac{1}{4}\right) = \frac{1}{2} \right\}.$$

$$i = 1, \quad h_1(y) \stackrel{\triangle}{=} \frac{1}{2}y^2 = \frac{1}{2}, \quad y > 0, \quad y_1 = 1,$$

$$i = 2, \quad h_2(y) \stackrel{\triangle}{=} y^2 = \frac{1}{2}, \quad y < 0, \quad y_2 = -\frac{\sqrt{2}}{2}.$$

So, we obtain $\mathcal{A}_0 = \{ y_1 = 1, \ y_2 = -\frac{\sqrt{2}}{2} \}$.

Step 4. Further, we have to solve the linearized problems $(i = 1, 2)$

$$(\mathcal{PL}_i): \qquad \Phi_i(x) = g(x) - \nabla h(y_i), x\rangle \downarrow \min_x, \quad x \in \mathbb{R}.$$

a) $i = 1$,

$$\Phi_1(x) = \frac{1}{4}x^4 - \langle \nabla h_1(y_1), x\rangle = \frac{1}{4}x^4 - x \downarrow \min_x.$$

The Fermat rule provides for $u_1 = 1$.

$i = 2$,

$$\Phi_2(x) = \frac{1}{4}x^4 - \langle \nabla h_2(y_2), x\rangle = \frac{1}{4}x^4 - \langle 2y_2, x\rangle = \frac{1}{4}x^4 + \sqrt{2}x \downarrow \min_x, \quad x \in \mathbb{R},$$

whence it follows $\nabla\Phi_2(x) = x^3 + \sqrt{2} = 0$ that yields $u_2 = -(2)^{1/6}$.

b) $i = 1$,

$$\Phi_3(x) = \frac{1}{2}x^4 - \langle \nabla h_1(y_1), x\rangle = \frac{1}{2}x^4 - x \downarrow \min_x.$$

As above, one has $u_3 = \sqrt[3]{0.5}$.

Here, it is necessary to note that the points u_2 and u_3 are unacceptable because the initial data and the final results are incompatible, i.e.

a) $i = 2$, $g(x) = g_1(x) = \frac{1}{4}x^4$ when $x \geq 0$ meanwhile the solution $u_2 = -(2)^{1/6}$ is negative;

b) $i = 1$. $g(x) = g_2(x) = \frac{1}{2}x^2$ when $x < 0$, while $u_3 = \sqrt[3]{0.5} > 0$.
Further we consider the last case.

c) $i = 2$.

$$\Phi_4(x) = g_2(x) - \langle \nabla h_2(y_2), x\rangle = \frac{1}{2}x^4 - \langle 2y_2, x\rangle = \frac{1}{2}x^4 + \sqrt{2}x \downarrow \min_x.$$

It is easy to see that the Fermat rule $\nabla\Phi_4(x) = 2x^3 + \sqrt{2} = 0$ yields $u_4 = -\left(\frac{\sqrt{2}}{2}\right)^{1/3}$.

Now, we have to verify VI (30).
Step 5. a) $i = 1$.

$$g(u_1) - \beta_0 - \langle \nabla h(y_1), u_1 - y_1\rangle = \frac{1}{4}u_1^4 = \frac{1}{4} - \langle y_1, u_1 - y_1\rangle = \frac{1}{4} - \frac{1}{4} - 0 = 0.$$

b) $i = 2$.

$$g_2(u_2) - \beta_0 - \langle \nabla h_2(y_2), u_2 - y_2\rangle = \tfrac{1}{2}u_4^4 - \tfrac{1}{4} - \langle 2y_2, u_4 - y_2\rangle$$

$$= \tfrac{1}{2}\left(\tfrac{\sqrt{2}}{3}\right)^{4/3} - \tfrac{1}{4} + \langle \sqrt{2}, -\left(\tfrac{\sqrt{2}}{2}\right)^{1/3} + \tfrac{\sqrt{2}}{2}\rangle$$

$$= \tfrac{1}{2}\left[\left(\tfrac{1}{2}\right)^{2/3} - \tfrac{1}{2}\right] + \langle \sqrt{2}, \left(\tfrac{\sqrt{2}}{2}\right) - \left(\tfrac{\sqrt{2}}{2}\right)^{1/3}\rangle.$$

Without a computer it is rather difficult to decide about the sign of the latter expression. Suppose, our program was incorrect at this point, and we turned out to be unsuccessful to violate the VI (30). What do we have to do further? It is necessary to change β for another value, i.e. to loop to Step 2.

Step 2. Change β_0 for $\beta_1 = \frac{3}{4}$.
Step 3. We need a new set $\mathcal{A}_1$ of points y_i satisfying

$$h(y) = \left\{ \begin{array}{ll} \frac{1}{2}y^2, & y \geq 0, \\ y^2, & y < 0 \end{array} \right\} = \beta_1 - \zeta_0 = \frac{3}{4} - \left(-\frac{1}{4}\right) = 1,$$

whence it follows

$$i = 1, \quad y_1 = \sqrt{2},$$
$$i = 2, \quad y_2 = -1.$$

On account of (34) we have to solve the linearized (convex) problems ($i = 1, 2$)

$$(\mathcal{PL}_i): \qquad \Phi_i(x) = g(x) - \langle \nabla h(y_i), x \rangle \downarrow \min_x, \quad x \in \mathbb{R}.$$

Note that here it is sufficient to investigate only the case ($i = 1$, a) and ($i = 2$, b), because other two ($i = 1$, b) and ($i = 2$, a) turn out to be unacceptable, as above.
$i = 1$, a)

$$g_1(x) - \langle \nabla h_1(y_1), x \rangle = \frac{1}{4}x^4 - x\sqrt{2} \downarrow \min_x, \quad x \in \mathbb{R}.$$

With the help of Fermat rule one has

$$x^3 - \sqrt{2} = 0, \quad x \geq 0, \quad u_1 = (2)^{1/6}.$$

$i = 2$, b)

$$g_2(x) - \langle \nabla h_2(y_2), x \rangle = \Phi_4(x) = \frac{1}{2}x^4 + 2x \downarrow \min, \quad x \in \mathbb{R},$$

which provides for

$$2x^3 + 2 = 0, \quad u_4 = -1.$$

Now, we need to verify VI (30). However, it is sufficient to consider only the case ($i = 2$, b) with $u_4 = -1$, $y_2 = -1$. Actually, in this case due to (34) we have

$$g_2(u_4) - \beta_1 - \langle \nabla h_2(y_2), u_4 - y_2 \rangle = \frac{1}{2}u_4^4 - 1 - \langle 2(-1), -1 - 1 \rangle = \frac{1}{2} - 1 - 0 = -\frac{1}{2} < 0.$$

The latter inequality means that GOCs have been violated, and, moreover, we were successful to "jump" out the local pit $z^0 = 1$ directly to the global solution $z^1 = -1$ by means of *Global Search Strategy (Method)*. □

5 Numerical Solution of the Applied Problems

5.1 Linear Complementarity Problem

As it was said in Sect. 2, we have to look at the LCP (1) as the optimization problem (2). Besides, we will consider the most difficult case when (2) is nonconvex. More precisely, the matrix M in the statement (2) is indefinite, i.e. possesses positive and negative eigenvalues.

Note that the LCP (1) represents necessary optimality conditions for the problem

$$f(x) = \frac{1}{2}\langle x, Mx\rangle + \langle q, x\rangle \downarrow \min, \quad x \geq 0. \tag{35}$$

However, this problem cannot replace (2) because M is indefinite. As a consequence, $f(\cdot)$ in (35) can be unbounded below, while the objective function in (2) is nonnegative and takes the zero value only at a solution to Problem (2). It is clear that this provides an additional information for the computational process.

Now, let us describe the principal stage of the *Global Search Algorithm* (GSA).

0. *Classification*. Thus, we decide to classify LCP with an indefinite matrix M as a d.c. minimization problem (3) with the strongly convex functions $g(\cdot)$ and $h(\cdot)$:

$$g(x) = \langle x, M_1 x\rangle + \langle q, x\rangle, \quad h(x) = \langle x, M_2 x\rangle,$$
$$M_i = M_i^{\top} > 0, \quad i = 1, 2, \quad M = M_1 - M_2.$$

I. The next stage is the local search. In order to do it, let us apply the LSM (16) which (for the LCP (2)) takes the form of the consecutive solutions of the following linearized problem

$$\left.\begin{array}{c} \Phi_s(x) = \langle M_1 x, x\rangle + \langle q - 2M_2 x^s, x\rangle \downarrow \min, \\ x \geq 0, \quad Mx + q \geq 0. \end{array}\right\} \tag{36}$$

To the end of solving Problems (36) we used the well-known XPress solver, which was especially designed for solving convex quadratic and linear programming problems.

On the other hand, to organize rather effective testing of the LSM (16), the data of the $(n \times n)$ matrix M were randomly generated in the interval $[-n, n]$ (see [34]). So, the set of randomly generated LCPs of type (2) and of the dimension varying from $n = 2$ to $n = 1000$ has been formed. Moreover, for every LCP we used three different starting points. Further, the local solution process has been performed and analyzed on this field of test LCPs. In particular, the results enable us to observe the behavior of the method (15)–(16) and to choose appropriate starting points for global search ("good-bad" points, starting at which the LSM (15)–(16) does not provide for a global solution to Problem (2)).

Note separately that due to Theorem 1 the linearized problems (36) may be solved at a low accuracy at the first steps; further, the accuracy δ_s can be gradually improved ($\delta_s \downarrow 0$), for example, $\delta_0 = 0.1$, $\delta_{s+1} = 0.5\delta_s$ until the condition $\delta_s \leq \delta$ is fulfilled with a given accuracy $\delta > 0$. The results of computational testing of LSM (16) have been presented for the first time in [34] and after have been considerably improved (till the dimension $n = 1{,}000$ with a 3.4 GHz Pentium computer with 1 Gb of memory). The auxiliary linearized problems (36) have been solved by XPress solver.

On the basis of the analysis of the results of computational testing [34] one can conclude that the LSM (16) showed itself rather effective for LCP (2). Moreover, it was considerably more effective in comparison with X-Press solver, because the latter was unable to deal with nonconvex LCP (2) of dimension $n \geq 10$, meanwhile

the LSM (15)–(16) yielded a critical (feasible) point for (2) in all considered test problems till the dimension $n = 1{,}000$ with obligatory and considerable decreasing of the goal function $\Phi(\cdot)$ in (2) (see [34]).

So, LSM (15)–(16) can be applied in a GSA, although it is not able, in general, to reach a global solution.

Now we can pass to a global search described in Sect. 4. To begin with, first, one has to propose a numeric solution of the equation

$$h(y) \stackrel{\triangle}{=} \langle y, M_2 y \rangle = \beta - \zeta_k, \;\; M_2 = M_2^T > 0,$$

where $\zeta_k := \Phi(z^k), \;\; \beta \in [\beta_-, \beta_+]$, more precisely to construct an approximation

$$\mathcal{A}_k(\beta) = \{y^1, \ldots, y^N \mid h(y^i) = \beta - \zeta_k, \;\; i = 1, \ldots, N_k\}$$

of the level surface $U_k(\beta) = \{x \mid h(x) = \beta - \zeta_k\}$. The construction of such an approximation is a key point in the implementation of the global search.

Regardless of the importance of this procedure, the construction may be performed in rather simple fashion, for example,

$$y^i = \mu_i d^i, \quad i = 1, \ldots, N, \tag{37}$$

where d^i are elements of some set in $\mathbb{R}^n$, for instance, $d^i = e^i, \;\; \{e^1, \ldots, e^n\}$ being the Euclidian basis of $\mathbb{R}^n$, and the numbers μ_i are chosen as the roots of the quadratic equation $h(\mu_i d^i) = \beta - \zeta_k$ due to the quadratic structure of $h(\cdot)$.

In order to solve Problem (2) we used the approximations as follows:

$$\mathcal{R}_1 = \{y^i = \mu_i e^i, \;\; y^{i+n} = -y^i \mid i = 1, \ldots, n\},$$

$$\mathcal{R}_2 = \{y^i = z^k + \mu_i e^i, \;\; y^{i+n} = z^k - \mu_i e^i \mid i = 1, \ldots, n\},$$

where z^k is the current iteration point.

Besides, we also applied the third approximation using the form (37), where $d^i (i = 1, \ldots, n)$ have been produced as the solutions of the linear programs

$$\langle e^i, x \rangle \downarrow \min_x, \;\; x \geq 0, \;\; Mx + q \geq 0, \;\; i = 1, \ldots, n \tag{38}$$

and d^{n+1} is the solution of the similar problem

$$\langle e, x \rangle \downarrow \min_x, \;\; x \geq 0, \;\; Mx + q \geq 0, \tag{39}$$

with $e = (1, \ldots, 1)^T \in \mathbb{R}^n$.

The results of computational testing of the developed GSA have first been published in [34] and turned out to be rather promising for the test problems of dimension till 400.

Now we are having the software which is able to solve LCP (2) till the dimension 10^3 in 10–12 min and till the dimension 10^4 in 90–150 min, remember, by means of (only one) almost the same computer as it was used in [34], without applying any parallel technology.

In addition, in order to compare the efficiency of GSA with existing software the same series of randomly generated problems have been solved using the solver PATH [34], which was especially designed for solving the LCP (1). Note that all computational simulations have been carried out by students and postgraduate students.

So, the qualities of the programs implemented may vary significantly.

Nevertheless, we can conclude that the developed GSA proved to be rather effective for solving (nonconvex) LCPs with indefinite matrix M.

5.2 Bimatrix Games

Here we present the principal points of the numerical search for Nash equilibrium (NE) defined in (5) in the two-person game stated in (4). The computational algorithm has been developed on the foundation of the following result of Mills [21].

Theorem 4. *([21]) (i) A situation (x^*, y^*) is a Nash equilibrium in the bimatrix game $G(A,B)$ (4) if and only if (x^*, y^*) is a part of a global solution $(x^*, y^*, \alpha_*, \beta_*) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^2$ to Problem (6).*

(ii) Moreover, α_* and β_* are the payoffs of the first and the second players, respectively:

$$\langle x^*, Ay^* \rangle = \alpha_*, \quad \langle x^*, By^* \rangle = \beta_*.$$

(iii) Finally, the optimization value of the goal function in Problem (6) is equal to zero

$$\mathcal{V}(6) = F(x^*, y^*, \alpha_*, \beta_*) = 0.$$

□

5.2.1 Classification

In order to develop a numerical method for solving Problem (6) we have, first, to classify it as a nonconvex problem. Because all the constraints in (6) are linear, we have to decide about the features of the cost function of Problem (6). It can be readily seen that this function has the d.c. decomposition as follows:

$$F(x,y,\alpha,\beta) = h(x,y) - g(x,y,\alpha,\beta), \tag{40}$$

where

$$\left.\begin{aligned} h(x,y) &= \tfrac{1}{4}(\|x+Ay\|^2 + \|x+By\|^2) \\ g(x,y,\alpha,\beta) &= \tfrac{1}{4}(\|x-Ay\|^2 + \|x-By\|^2) + \alpha + \beta \end{aligned}\right\} \tag{41}$$

are convex functions ($h(\cdot)$ on $\mathbb{R}^{m+n}$, $g(\cdot)$ on $\mathbb{R}^{m+n+2}$). In other words, we have the d.c. minimization problem $(\mathcal{P}_{\mathrm{BM}})$ as follows:

$$(\mathcal{P}_{\mathrm{BM}}):\quad \left.\begin{array}{c} F_0(x,y,\alpha,\beta) = -F(x,y,\alpha,\beta) = g(x,y,\alpha,\beta) - h(x,y) \downarrow \min\limits_{x,y,\alpha,\beta}, \\ (x,\beta) \in X = \{\, x \in S_m, \quad \beta \in \mathbb{R} \mid x^\top B \le \beta e_n \,\}, \\ (y,\alpha) \in Y = \{\, y \in S_n, \quad \alpha \in \mathbb{R} \mid Ay \le \alpha e_m \,\}. \end{array}\right\} \qquad (6')$$

It is easy to see that the cost function $F_0(x,y,\alpha,\beta)$ is nonnegative ($F_0(\cdot) \ge 0$, $F(\cdot) \le 0$) on the feasible set of Problem $(\mathcal{P}_{\mathrm{BM}})$–$(6')$.

In addition, if we denote

$$\alpha(y) := \max_{1 \le i \le m} (Ay)_i, \quad \beta(x) := \max_{1 \le j \le n} (x^\top B)_j, \qquad (42)$$

then due to the necessity proof of Theorem 2 we can reformulate Theorem 2 in the contrapositive form as follows.

Theorem 5. *([28]) If a feasible tuple $(\hat{x},\hat{y},\hat{\alpha},\hat{\beta})$ is not a global solution to Problem $(\mathcal{P}_{BM})$–$(6')$, then there exist some vectors $(u,v) \in S_m \times S_n$ and $(\bar{x},\bar{y}) \in S_m \times S_n$, and a number γ such that*

$$\gamma - h(u,v) = \zeta := F_0(\hat{x},\hat{y},\hat{\alpha},\hat{\beta}) > 0, \qquad (43)$$

$$g(u,v,\alpha(v),\beta(u)) \le \gamma \le \sup(g,D), \qquad (44)$$

$$g(\bar{x},\bar{y},\alpha(\bar{y}),\beta(\bar{x})) - \gamma < \langle \nabla_x h(u,v), \bar{x} - u \rangle + \langle \nabla_y h(u,v), \bar{y} - v \rangle. \qquad (45)$$

□

Applying just this result we will develop a GSM for finding a global solution to $(\mathcal{P}_{\mathrm{BM}})$–$(6')$. The first step of this GSM is a local search algorithm which takes into account the bilinear structure of the cost function $\alpha + \beta - \langle x, (A+B)y \rangle$ of the Problem $(\mathcal{P}_{\mathrm{BM}})$–$(6')$.

5.2.2 Local Search

First, let us repeat that LSMs play the important role in the processes of search for a global solution to nonconvex problems, since it provides for the so-called critical (stationary) points which may be considerably better than a simple feasible point. Moreover, if a starting point occurs rather closed to a global solution (as in the case of Newton method for solving systems of nonlinear equations), then an LSM is able to provide for the global solution.

Therefore, we have to pay our attention and considerable efforts to a creation (a design or a choice) and the substantiation of local search procedures.

For instance, for the case of Problem $(6')$ it might be possible to apply the LSM (15)–(16) taking into account the d.c. representation (40)–(41) and applying the corresponding methods of quadratic programming.

However, in this case the bilinear nature of Problem $(\mathcal{P}_{\mathrm{BM}})$–$(6')$, the specific character of the cost function $\langle x, (A+B)y \rangle$, namely, its bilinearity, would be lost. Therefore, we propose to follow another way, more natural in the case, taking

into account the bilinear structure of the goal function $F_0(x,y,\alpha,\beta) = \alpha + \beta - \langle x,(A+B)y\rangle$. Combining the linearization idea and the separation of the variables into groups according to the statement of Problem $(\mathcal{P}_{\mathrm{BM}})$–(6′), we obtain without alternative a procedure of consecutive solving the following two (linearized at a point $(u,v) \in \mathbb{R}^m \times \mathbb{R}^n$) problems

$$(\mathcal{PL}_x): \quad \begin{cases} \beta - \langle (A+B)v, x\rangle \downarrow \min\limits_{(x,\beta)}, \\ (x,\beta) \in X = \{(x,\beta) \in S_m \times \mathbb{R} \mid x^\top B \le \beta e_n\}, \end{cases} \tag{46}$$

$$(\mathcal{PL}_y): \quad \begin{cases} \alpha - \langle u^\top (A+B), y\rangle \downarrow \min\limits_{(y,\alpha)}, \\ (y,\alpha) \in Y = \{(y,\alpha) \in S_n \times \mathbb{R} \mid Ay \le \alpha e_m\}. \end{cases} \tag{47}$$

Unexpectedly enough, the procedure of consecutive solving the Problem $(\mathcal{PL}_x)$ and $(\mathcal{PL}_y)$ converges in the following sense.

Theorem 6. *([33]) The sequence of the tuples $(x^s, y^s, \alpha_s, \beta_s)$ generated by the LSM consisting in the consecutive fulfilling of the following inequalities*

$$\alpha_{s+1} - \langle x^s(A+B), y^{s+1}\rangle - \frac{\rho_s}{2} \le \inf_{(y,\alpha)} \{\alpha - \langle x^s(A+B), y\rangle \mid (y,\alpha) \in Y\}, \tag{48}$$

$$\beta_{s+1} - \langle x^{s+1}(A+B), y^{s+1}\rangle - \frac{\rho_s}{2} \le \inf_{(x,\beta)} \{\beta - \langle x,(A+B)y^{s+1}\rangle \mid (x,\beta) \in X\}, \tag{49}$$

converges to a quadruple $(\hat{x},\hat{y},\hat{\alpha},\hat{\beta})$ satisfying the conditions as follows

$$\left.\begin{aligned} F_0(\hat{x},\hat{y},\hat{\alpha},\hat{\beta}) \le F_0(\hat{x},y,\alpha,\hat{\beta}) \quad \forall (y,\alpha) \in Y, \\ F_0(\hat{x},\hat{y},\hat{\alpha},\hat{\beta}) \le F_0(x,\hat{y},\hat{\alpha},\beta) \quad \forall (x,\beta) \in X, \end{aligned}\right\} \tag{50}$$

provided that $\rho_s > 0$, $s = 0,1,2,\ldots$, $\sum\limits_{s=0}^{\infty} \rho_s < +\infty$. □

We will call, henceforth, such a point satisfying (50) a critical point of Problem $(\mathcal{P}_{\mathrm{BM}})$–(6′). The LSM (46)–(49) has been tested on a rather large field of well-known test problems [33], and also on the various test problems especially constructed with the help of the idea from [5], by beginning the known games of small dimension (2×2, 3×3) and until the test-games of rather high size (say, $m = n = 1{,}000$).

Computational simulations certify unexpected effectiveness of the developed LSM that naturally depends on the method or a package of applied software (CPLEX) that was used for solving the linear problems (46), (47). Now we are able to perform LSM with the data $m = n = 10^6$ rather easily and effectively.

5.2.3 Global Search Algorithm

Recall that, in addition to local search, the basic stages of a GSM include an approximation of the level surface of the convex function $h(\cdot)$ (which creates the basic

nonconvexity in Problem $(\mathcal{P}_{\mathrm{BM}})$–(6′)), a solution of a linearized problem $(\mathcal{PL}_s)$, the verification of the VI (30) with $w^i \in U(h, \gamma - \zeta) = \{(x,y) \in \mathbb{R}^{m+n} \mid h(x,y) = \gamma - \zeta\}$, and finally a line search along the variable $\gamma \in \mathbb{R}$.

Taking into account the particularities of Problem (6′) the following modifications have been introduced into the general scheme of Global Search on the bases of Theorem 5.

1. Due to the properties of the cost function $F_0(x,y,\alpha,\beta)$ (see Theorem 4) the supplementary stopping criterion was introduced.
2. Two new parameters (q and ν) have been introduced in order to control the speed and the accuracy of the algorithm [23].

Let us now describe the GSM for solving Problem $(\mathcal{P}_{\mathrm{BM}})$–(6′) in a more algorithmic form.

Assume, we are given a starting feasible point $(x^0, y^0, \alpha_0, \beta_0) \in D = X \times Y$; number sequences $\{\tau_k\}$ and $\{\delta_k\}$, $k = 0,1,2,\ldots$, $\tau_k \downarrow 0$, $\delta_k \downarrow 0$ $(k \to \infty)$; a set of directions $Dir = \{(\bar{u}^1, \bar{v}^1), \ldots, (\bar{u}^N, \bar{v}^N) \in \mathbb{R}^{m+n}\}$ the bounds $\gamma_- \approx \inf(g, D), \gamma_+ \approx \sup(g, D)$; and parameters $\nu \in]0,1[$ and q.

Global Search Methods for $(\mathcal{P}_{\mathrm{BM}})$–(6′).

Step 0. Set $k := 0$, $(\bar{x}^k, \bar{y}^k, \bar{\alpha}_k, \bar{\beta}_k) := (x^0, y^0, \alpha_0, \beta_0)$, $s := 0$, $p := 1$, $\gamma := \gamma_-$, $\Delta\gamma = (\gamma_+ - \gamma_-)/q$.

Step 1. Starting from $(\bar{x}^k, \bar{y}^k, \bar{\alpha}_k, \bar{\beta}_k) \in D$, move to τ_k—critical point $(x^k, y^k, \alpha^k, \beta^k) \in D$ by means of the special LSM (48)–(49).

Set $\xi_k := F_0(x^k, y^k, \alpha_k, \beta_k) \le F_0(\bar{x}^k, \bar{y}^k, \bar{\alpha}_k, \bar{\beta}_k)$.

Step 2. (*Stopping criterion*). If $\xi_k \le \varepsilon$, where ε is the prescribed accuracy, then STOP: $(x^k, y^k) \in NE(G, \varepsilon)$.

Step 3. With the help of the point $(\bar{u}^p, \bar{v}^p) \in Dir$ $(p = 1,..,N)$ construct a point (u^p, v^p) such that $h(u^p, v^p) = \gamma - \xi_k$. Compute the numbers $\alpha_p := \max\limits_{1 \le i \le m} (Av^p)_i, \beta_p := \max\limits_{1 \le j \le n} (u^p B)_j$.

Step 4. If $g(u^p, v^p, \alpha_p, \beta_p) > \gamma + \nu\gamma$, then set $p := p + 1$ and return to Step 3. Else go to Step 5.

Step 5. Starting at the point $(u^p, v^p, \alpha_p, \beta_p)$ find a $2\tau_k$-critical point $(\hat{x}^p, \hat{y}^p, \hat{\alpha}_p, \hat{\beta}_p) \in D$ of Problem (6′) by means of special LSM.

Step 6. (*Stopping criterion*). If $F_0((\hat{x}^p, \hat{y}^p, \hat{\alpha}_p, \hat{\beta}_p) \le \varepsilon$, then STOP. $(\hat{x}^p, \hat{y}^p) \in NE(G, \varepsilon)$.

Step 7. Find a δ_k-solution (x_0^p, y_0^p) to the level problem or, what is equivalent,

$$\begin{aligned} &\langle \nabla_x h(x_0^p, y_0^p), \hat{x}^p - x_0^p \rangle + \langle \nabla_y h(x_0^p, y_0^p), \hat{y}^p - y_0^p \rangle + \delta_k \\ &\ge \sup_{(x,y)} \{ \langle \nabla_x h(x_0^p, y_0^p), \hat{x}^p - x \rangle + \langle \nabla_y h(x_0^p, y_0^p), \hat{y}^p - y \rangle \mid h(x,y) = \gamma - \zeta_k \}, \end{aligned} \tag{51}$$

where $h(x_0^p, y_0^p) = \gamma - \zeta_k$.

Step 8. Compute

$$\eta_k(\gamma) = g(\hat{x}^p, \hat{y}^p, \hat{\alpha}_p, \hat{\beta}_p) - \gamma - \langle \nabla_x h(x_0^p, y_0^p), \hat{x}^p - \hat{x}_0^p \rangle - \langle \nabla_y h(x_0^p, y_0^p), \hat{y}^p - \hat{y}_0^p \rangle.$$

Step 9. If $\eta_k(\gamma) \geq 0$ and $p < N$, then set $p := p+1$ and loop to Step 3.
Step 10. If $\eta_k(\gamma) \geq 0$ and $p = N$, then set $\gamma := \gamma + \triangle\gamma$ and $p := 1$ and return to Step 3.
Step 11. If $\eta_k(\gamma) < 0$, then $k := k+1, (\bar{x}^{k+1}, \bar{y}^{k+1}, \bar{\alpha}_{k+1}, \bar{\beta}_{k+1}) := (\hat{x}^p, \hat{y}^p, \hat{\alpha}^p, \hat{\beta}^p)$, and return to Step 1.
Step 12. If $p = N$ and $\eta_k(\gamma) \geq 0 \forall \gamma \in [\gamma_-, \gamma_+]$ (i.e., the line search with respect to $\gamma \in [\gamma_-, \gamma_+]$ is finished), then stop.

The GSM presented above is not an algorithm, since some of its steps have not been described clearly, and must be precised. For instance, it is not clear how to find the pair (x_0^p, y_0^p) on Step 7, besides, how to construct a point (u^p, v^p) on the level surface of $h(\cdot)$: $h(u^p, v^p) = \gamma - \zeta_k$ with the help of a given direction $(\bar{u}^p, \bar{v}^p)$ on Step 3. As to the first problem of Step 7, it can be solved analytically for the quadratic function

$$h(x,y) = \frac{1}{4}(\|x+Ay\|^2 + \|x+By\|^2), \tag{52}$$

more precisely, the exact solution is given by the formula [32, 33]

$$(x_0^p, y_0^p) = t(\hat{x}^p, \hat{y}^p),\ t = \left[\frac{\gamma - \zeta_k}{h(\hat{x}^p, \hat{y}^p)}\right]^{\frac{1}{2}}.$$

The construction of point (u^p, v^p) satisfying $h(u^p, v^p) = \gamma - \zeta_k$ can be done in the similar way:

$$(u^p, v^p) = \lambda_p(\bar{u}^p, \bar{v}^p), p = 1, \ldots, N,$$
$$\lambda_p = \lambda_p(\zeta_k, \gamma) = \pm[\gamma - \zeta_k / h(\bar{u}^p, \bar{v}^p)]^{\frac{1}{2}}.$$

To the end of the solving Problem $(\mathcal{P}_{\mathrm{BM}})$–(6′) it was used the approximations of the level surface $U(h, \gamma - \zeta) = \{(x,y) |\ h(x,y) = \gamma - \zeta\}$ (see Step 3) constructed with the help of the following sets of directions

$$Dir1 = \{(e^i, e^j) \in \mathbb{R}^{m+n} |\ i = 1, \ldots, m,\ j = 1, \ldots, n\},$$

where $\{e^i\}$ is the Euclidian basis in $\mathbb{R}^m$ and $\{e^j\}$ is the basis in $\mathbb{R}^n$, respectively;

$$Dir2 = \{(e^i + x, e^j + y) |\ i = 1, \ldots, m,\ j = 1, \ldots, n\},$$

where (x,y) is a critical point provided by the special LSM;

$$Dir3 = \{(a^j + e_m, b^i + e_n) |\ i = 1, \ldots, m,\ j = 1, \ldots, n\},$$

where $a^j \in \mathbb{R}^n$ are the columns in A and $b^i \in \mathbb{R}^n$ are the rows in B, and $e_p = (1, \ldots, 1) \in \mathbb{R}^p$, $p = m, n$.

Note that the sets $Dir1, Dir2, Dir3$ have been selected as the most efficient ones after comparative computational experiments. But on the other hand, it is easy to

see that the number of points in the constructed approximations strongly depends on the size of the problem, i.e. is equal to $m \times n$.

So, the number of points in the approximations grows as q^2, where $q = \min\{m;n\}$.

It is clear that this moment makes it prohibited the numerical solution of Problem $(\mathcal{P}_{\mathrm{BM}})$–(6) of high dimension due to the excessive solution time.

In order to avoid this drawback it was employed some reducing procedure of the sets $Dir1,2,3$ to sets with the number of points equal to $2(m+n)$ [23, 33].

5.2.4 Computational Simulations

The numerical experiments were conducted applying software programs implementing the GSAs described above. For all the problems, a starting point was chosen as follows:

$$x_i^0 = \frac{1}{m},\ i = 1,2,\ldots,m;\ y_j^0 = \frac{1}{n},\ j = 1,2,\ldots,n,$$
$$\alpha_0 = \max_i (Ay^0)_i,\ \beta_0 = \max_j (x^0 B)_j.$$

The computational simulations have been separated into several stages, and the first results of these experiments have been published in [23].

Further, the analysis of the results allowed us to conclude about some shortcomings of the software program developed. First of all, it was the solving method for linearized problems (46)–(47). Recall that to the end the simplex method program or the supporting cone method program was employed, which showed itself very excessive from the viewpoint of the solution time of problems of high dimension.

As a consequence, for solving the BM games of rather high dimension (up to $1{,}000 \times 1{,}000$) we decided to apply ILOG CPLEX 9.1(http://www-01.ibm.com/soft-ware/commerce/optimization/cplex-optimizer/index.html) especially oriented to LP problems. In addition, in order to create the worst conditions for the global search software the entries of matrices A and B have randomly been generated from the interval $[-n,n]$, where $n = m$.

The software programs of global search were run on Pentium 4, CPU 3 GHz with 512 Mb of RAM and have been implemented by post-graduated students without a long computational experience.

Nevertheless, the results of computational solving of BM games $(m = n)$ can be viewed as rather promising from the point of view of analysis of numeric results of Table 1.

In Table 1, $m = n$ is the number of pure strategies of players 1 and 2, F_0 stands for the value of the goal function at the starting points, F_k is the corresponding value at the best obtained point, *st* is the number of iterations of GSAs (or, what is the same, the number of critical (stationary) points passed by GS algorithms), *LP* and *Loc* represent the number of linearized problems solved and the number of local search algorithm's applications, respectively.

Table 1 Computational results

$m=n$	F_0	F_k	st	LP	Loc	$Time$
200	40.3811	0	2	972	65	00:43.66
250	51.1617	0	4	4,843	321	5:36.30
300	60.1209	0	2	28	2	00:44.17
400	75.0987	0	6	16,168	978	59:05.49
500	75.3494	0	2	158	20	05:57.88
600	84.1025	0	2	82	7	07:33.92
700	89.0439	0	2	54	6	11:50.34
800	99.8335	0	2	48	3	15:28.70
900	100.0419	0	3	136	11	29:31.38
1,000	106.8368	0	3	178	18	45:04.34

It can readily be seen that in the cases of $m=n=250$ and 400 it happened to randomly generate very difficult problems.

Despite these difficulties, all test-problems have successfully been solved that certifies on the computational effectiveness of the software program created on the basis of the GSA and the Global Search Theory.

Now, we are preparing to attack the bimatrix game of dimension $m=n=10^4$ and the similar three-person-game of dimension $m=n=l=5$, and 10.

5.3 Quadratic-Linear Bilevel Optimization

In this subsection we will consider the following problem of bilevel programming

$$(\mathcal{BP}): \qquad F(x,y) := \frac{1}{2}\langle x,Cx\rangle + \langle c,x\rangle + \frac{1}{2}\langle y,Cy\rangle + \langle c_1,y\rangle \downarrow \min_{x,y}, \tag{53}$$

$$(x,y) \in X := \{(x,y) \in \mathbb{R}^m \times \mathbb{R}^n \mid Ax+By \le a, \quad x \ge 0\}, \tag{54}$$

$$y \in Y_*(x) := \operatorname{Arg}\min_y \{\langle d,y\rangle \mid y \in Y(x)\} \tag{55}$$

$$Y(x) := \{y \in \mathbb{R}^n \mid A_1x+B_1y \le b, \quad y \ge 0\}, \tag{56}$$

where we are seeking an optimistic solution [6, 8], i.e. the upper level (x, leader) and the lower level (y, follower) are searching together (in cooperation) a common solution (x_*,y_*). Here, $c \in \mathbb{R}^m$, $d,c_1 \in \mathbb{R}^n$, $a \in \mathbb{R}^p$, $b \in \mathbb{R}^q$, and matrices C, C_1, A, B, A_1, B_1 are of corresponding dimensions. In addition, $C=C^\top > 0$, $C_1=C_1^\top > 0$, so that the leader cost function is a convex quadratic function, while the follower goal function is linear.

Assume that

$(\mathcal{H})$: (i) the function $F(x,y)$ is bounded from below on X,

(ii) the function $\langle d,y\rangle$ is bounded from below on $Y(x)\ \forall x \in Pr(X)$.

It is clear, that, from the first glance, any nonconvexity is not visible in the formulation $(\mathcal{BP})$–(53)–(56). In order to put it explicit we apply the KKT-conditions for the follower problem (55) –(56):

$$\left.\begin{array}{c} d+vB_1 \geq 0, \quad v \geq 0, \quad A_1x+B_1y \leq b, \\ \langle d,y\rangle - \langle A_1x-b,v\rangle = 0, \end{array}\right\} \tag{57}$$

where v is the Lagrangian multipliers. Since the follower problem is also convex, the relations (57) are equivalent to the statement (55)–(56).

Let us replace, now, in $(\mathcal{BP})$ the follower problem by (57). It yields us the new problem

$$\left.\begin{array}{ll} (\mathcal{P}): & F(x,y) \downarrow \min\limits_{x,y,v}, \\ Ax+By \leq a, & A_1x+B_1y \leq b, \quad d+vB_1 \geq 0, \\ \langle d,y\rangle = \langle A_1x-b,v\rangle, & x \geq 0, \quad y \geq 0, \quad v \geq 0. \end{array}\right\} \tag{58}$$

The following result establishes the relation between the $(\mathcal{BP})$ and Problem $(\mathcal{P})$.

Theorem 7. *([8]) For the pair (x^*,y^*) to be a global solution to Problem $(\mathcal{BP})$–(53)–(56), it is necessary and sufficient that there exists a vector $v^* \in \mathbb{R}^q$ such that the triple (x^*,y^*,v^*) is a global solution to Problem $(\mathcal{P})$–(58).* □

Note, that, first, the relation exists only between the global solutions of Problems $(\mathcal{P})$ and $(\mathcal{BP})$, but it does not take place between local solutions or between local and global ones.

Second, it is easy to see that the feasible set of Problem $(\mathcal{P})$–(58) is nonconvex because of the presence of the bilinear equality-constraint in (58). Thus, Problem $(\mathcal{P})$–(58) turns out to be nonconvex.

Let us denote

$$H(x,y,v) := \langle d,y\rangle - \langle A_1x-b,v\rangle \tag{59}$$

and introduce a μ-parametric family of problems as follows:

$$\left.\begin{array}{ll} (\mathcal{P}(\mu)): & F_1(x,y,v,\mu) = F(x,y)+\mu H(x,y,v) \downarrow \min\limits_{x,y,v}, \\ (x,y,v) \in D := \{(x,y,v) \mid Ax+By \leq a, & A_1x+B_1y \leq b, \\ & d+vB_1 \geq 0, \quad x \geq 0, \quad y \geq 0, v \geq 0\}, \end{array}\right\} \tag{60}$$

where $\mu > 0$ is a penalty parameter. If we rewrite the function $H(\cdot)$ in the form

$$H(x,y,v) = \langle d+vB_1,y\rangle - \langle A_1x+B_1y-b,v\rangle, \tag{59$'$}$$

then it becomes clear that

$$H(x,y,v) \geq 0 \quad \forall (x,y,v) \in D. \tag{61}$$

Furthermore, it can be readily seen that for a fixed value of μ Problem $(\mathcal{P}(\mu))$ is convex and quadratic with respect to the variables (x,y), and, besides, bilinear with respect to the variables x and v. So, Problem $(\mathcal{P}(\mu))$ can be called quadratic-bilinear, but anyway it stays to be nonconvex with the convex feasible set D (defined in (60)) and the nonconvex objective function $F_1(\cdot)$. Below we will show that $F_1(x,y,v,\mu)$ is a d.c. function.

Let us suppose $(x(\mu),y(\mu),v(\mu))$ be a solution to Problem $(\mathcal{P}(\mu))$–(60) for a given $\mu \in \mathbb{R}$. Further, denote $H[\mu] = H(x(\mu),y(\mu),v(\mu))$. Then the following relations between Problems $(\mathcal{P})$–(58) and $(\mathcal{P}(\mu))$–(60) take place.

(i) If the equality $H[\hat{\mu}] = 0$ holds for some value $\hat{\mu} \in \mathbb{R}$ and $(\hat{x},\hat{y},\hat{v}) = (\hat{x} = x(\hat{\mu}),\hat{y} = y(\hat{\mu}),\hat{v} = v(\hat{\mu}))$ is a solution to Problem $(\mathcal{P}(\hat{\mu}))$, then the triple $(\hat{x},\hat{y},\hat{v})$ is a solution to Problem $(\mathcal{P})$.

(ii) Moreover, for all $\mu > \hat{\mu}$ the equality $H[\mu] = H(x(\mu),y(\mu),v(\mu)) = 0$ holds, and, in addition, $(x(\mu),y(\mu),v(\mu))$ is a solution to Problem $(\mathcal{P})$.

In connection with these assertions we have results more suitable for computational uses.

Proposition 1. *([36]) Let $(x(\mu),y(\mu),v(\mu)) \in D$ be a τ_1-solution to Problem $(\mathcal{P}(\mu))$, and, besides,*

$$H(x(\mu),y(\mu),v(\mu)) \le \tau_2.$$

Then

(i) $y(\mu)$ is a τ_2-solution to the follower problem (55)–(56) with parameter $x = x(\mu)$;

(ii) $(x(\mu),y(\mu))$ is an approximate τ_1-solution to Problem $(\mathcal{BP})$–(53)–(56). □

The above assertions allow us to apply the global search methodology developed in Sect. 4 for solving Problem $(\mathcal{P}(\mu))$–(60) and, as a consequence, for finding an approximate global solution to Problem $(\mathcal{BP})$–(53)–(56).

5.3.1 Local Search

It can be readily seen that Problem $(\mathcal{P}(\mu))$ can be rewritten in the following form

$$(\mathcal{P}(\mu)): \qquad F_1(x,y,v) := \frac{1}{2}\langle x,Cx\rangle + \langle c,x\rangle + \frac{1}{2}\langle y,C_1y\rangle + \langle c_1,y\rangle + \mu\left[\langle d,y\rangle - \langle A_1x-b,v\rangle\right] \downarrow \min_{x,y,v}, \quad (62)$$

$$(x,y) \in Z := \{(x,y) \mid Ax+By \le a, \quad A_1x+B_1y \le b, \quad x \ge 0, \quad y \ge 0\}, \quad (63)$$

$$v \in V := \{v \mid d+vB_1 \ge 0, \quad v \ge 0\}. \quad (64)$$

On account of the assumptions $(\mathcal{H})$, it is easy to see that the cost function $F_1(\cdot)$ is bounded from below on the set $D = Z \times V$. Further, the statement (62)–(64) of Problem $(\mathcal{P}(\mu))$ suggests the idea of local search consisting in a consecutive solution of

the problem (62)–(64) with respect to the groups of variables; more precisely, in the case (62)–(64) , first, with respect to the pair (x,y) and, after that, with respect to the variables v, or in the inverse order.

Note that Problem $(\mathcal{P}(\mu))$ with a fixed value of the variable v becomes a convex quadratic optimization problem. On the other hand, for a fixed pair (x,y) we obtain a linear programming (LP) problem with respect to v. So, these auxiliary problems can be solved by standard software packages (CPLEX, X-Press, etc.)

Therefore, we can produce local search as it was done for the Bimatrix games.

Given some starting point $v_0 \in V$, we describe a so-called V-procedure as follows:

Step 0. Set $s := 0$, $v^s := v_0$.

Step 1. Find a $\frac{\rho_s}{2}$-solution (x^{s+1}, y^{s+1}) of the problem

$$\left.\begin{array}{c} \frac{1}{2}\langle x, Cx\rangle + \langle c, x\rangle + \frac{1}{2}\langle y, C_1 y\rangle + \langle c_1, y\rangle \\ +\mu\left[\langle d, y\rangle - \langle A_1 x - b, v^s\rangle\right] \downarrow \min\limits_{x,y}, \quad (x,y) \in Z, \end{array}\right\} \qquad (\mathcal{PL}_s)$$

so that the following inequality holds

$$F_1(x^{s+1}, y^{s+1}, v^s) \leq \inf\limits_{(x,y)}\{F_1(x,y,v^s) \mid (x,y) \in Z\} + \frac{\rho_2}{2}. \qquad (65)$$

Step 2. Find a $\frac{\rho_s}{2}$-solution v^{s+1} of LP problem

$$\langle d - A_1 x^{s+1}, v\rangle \downarrow \min\limits_{v}, \quad v \in V, \qquad (\mathcal{LP}_s)$$

so that the following inequality is satisfied

$$F_1(x^{s+1}, y^{s+1}, v^{s+1}) \leq \inf\limits_{v}\{F_1(x^{s+1}, y^{s+1}, v) \mid v \in V\} + \frac{\rho_2}{2}. \qquad (66)$$

Step 3. Set $s := s+1$ and loop to Step 1.

Under the condition

$$\rho_s > 0, \quad s = 0, 1, 2, \ldots, \quad \sum_{s=0}^{\infty} \rho_s < +\infty,$$

we can prove, as it was in the Bimatrix games, that the numerical sequence $\{F_{1s} = F_1(x^s, y^s, v^s)\}$ generated by the V-procedure from above is converging.

Moreover, if $(x^s, y^s, v^s) \to (\hat{x}, \hat{y}, \hat{v})$, then the point $(\hat{x}, \hat{y}, \hat{v})$ turns out to be a critical point of Problem $(\mathcal{P}(\mu))$–(62)–(64) [35] or partially global solution to $(\mathcal{P}(\mu))$, i.e.

$$\left.\begin{array}{l} F_1(\hat{x}, \hat{y}, \hat{v}) \leq F_1(x, y, \hat{v}) \quad \forall (x,y) \in Z, \\ F_1(\hat{x}, \hat{y}, \hat{v}) \leq F_1(\hat{x}, \hat{y}, v) \quad \forall v \in V. \end{array}\right\} \qquad (67)$$

Note that if a point $(\bar{x}, \bar{y}, \bar{v})$ is a local solution to Problem $(\mathcal{P}(\mu))$–(62)–(64), then $(\bar{x}, \bar{y}, \bar{v})$ turns out to be a critical point of Problem $(\mathcal{P}(\mu))$. Thus, the notion of critical point just introduced is really substantiated by and connected with the common notion of local solution. The similar to V-procedure so-called XY-procedure (starting at a point $(x_0, y_0) \in Z$) has also been studied and substantiated.

In order to test the developed LSM a rather large field of test-problems of the form $(\mathcal{BP})$–(53)–(56) has been constructed with the help of the idea of Calamai and Vicente [5], which provides for the bilevel problems with well-known properties, local and global solutions (even the numbers of which is known).

Now, a few words about the numerical testing of LSM.

First, the computational simulation was threefold:

(a) to choose a suitable value of the penalty parameter μ that provides for the equality $H(x(\mu), y(\mu), v(\mu)) = 0$ (the exact penalty [3, 22, 40]);
(b) to find starting points suitable for Global Search, i.e. from which the LSM was not able to reach a global solution;
(c) and finally, to compare two versions of LSM (with V- or XY-procedures).

Analyzing the testing results, one concluded that the computational time was rather short (less than 0.1 s), when the stopping criterion was satisfied at the accuracy $\tau = 10^{-4}$.

Furthermore, the value $\mu = 10$ of penalty parameter μ turned out to be sufficient to reach the equality $H(x(\mu), y(\mu), v(\mu)) = 0$ at a τ-critical point $(x(\mu), y(\mu), v(\mu))$. The targets (b) and (c) have been also reached.

Moreover, it should be specially noted the high rate of convergence of the XY- and V-procedures on the considered series of randomly generated problems, only two iterations were needed (starting from arbitrary feasible point) in order to get a critical point. So, the results of computational testing of the LSM were rather promising [35, 37].

5.3.2 Global Search

Let us repeat that the numerical test results showed that the special LSMs (V- and XY procedures) do not, in general, yield a global solution, even in problems of small sizes.

According to the methodology of Sect. 4, first we need to derive an explicit d.c. decomposition (if possible) of the cost function of the problem under scrutiny.

It is not hard to see that the goal function $F_1(x, y, v)$ of the problem $(\mathcal{P}(\mu))$ can be represented as a difference of two convex functions, for instance, as follows:

$$F_1(x, y, v) = g(x, y, v) - h(x, v), \tag{68}$$

where

$$\left.\begin{array}{c} g(x,y,v) = \frac{1}{2}\langle x, Cx\rangle + \langle c,x\rangle + \frac{1}{2}\langle y, C_1 y\rangle + \langle c_1, y\rangle \\ + \mu\big(\langle v,b\rangle + \langle y,d\rangle + \frac{1}{4}\|v - A_1 x\|^2\big), \\ h(x,v) = \frac{\mu}{4}\|v + A_1 x\|^2 \end{array}\right\} \tag{69}$$

are convex functions. Note that this d.c. decomposition is different with respect to these ones that was used in [35–37].

As it was noted above, the procedures of escaping critical points are based on GOC of Theorem 2 (see (25)) and employing the constructive (algorithmic) property of GOC. In the case of Problem $(\mathcal{P}(\mu))$ these GOCs take the form as follows:

$$(x_*, y_*, v_*) \in Sol(\mathcal{P}(\mu)), \quad \zeta := F_1(x_*, y_*, v_*) \Longrightarrow$$
$$\forall (z,w,\gamma) \in \mathbb{R}^{m+q+1}: \quad h(z,w) = \gamma - \zeta, \tag{70}$$
$$g(x,y,v) - \gamma \geq \langle \nabla_{xv} h(z,w), (x,v) - (z,w)\rangle \quad \forall (x,y,v) \in D. \tag{71}$$

Besides, if for some $(\hat{z}, \hat{w}, \hat{\gamma})$ in (70) and $(\hat{x}, \hat{y}, \hat{v}) \in D$

$$g(\hat{x}, \hat{y}, \hat{v}) < \hat{\gamma} + \langle \nabla_{xv} h(\hat{z}, \hat{w}), (\hat{x}, \hat{v}) - (\hat{z}, \hat{w})\rangle,$$

i.e. the VI (71) is violated, then due to the convexity of $h(\cdot)$ it follows

$$F_1(\hat{x}, \hat{y}, \hat{v}) < F_1(x_*, y_*, v_*).$$

In other words, $(\hat{x}, \hat{y}, \hat{v}) \in D$ is "better" than (x_*, y_*, v_*).

Similarly to Sect. 4 and according to the methodology presented in Sect. 3 Problem $(\mathcal{P}(\mu))$ is decomposed into several simpler problems as follows:

(a) one-dimensional search along the variable γ;
(b) constructing the level surface approximation of the convex function $h(x,v)$, which does not depend on y, as it was in our earlier papers [35–37]. It is clear that in this case the approximations must be easier to construct.

On the other hand, we have to pay attention to the fact that in view of the different d.c. representation (68)–(69) the global search has to be changed and becomes different with respect to [35, 36].

Assume, we are given a point $(x_0, y_0, v_0) \in \mathbb{R}^{m+n+q}$, numerical sequences $\{\tau_k\}$, $\{\delta_k\}$, $\tau_k, \delta_k > 0$, $k = 0, 1, \ldots$, $\tau_k \downarrow 0$, $\delta_k \downarrow 0$ $(k \to \infty)$, numbers $\gamma_- \approx \inf\limits_{(x,y,v)} (g, D)$ and $\gamma_+ \approx \sup\limits_{(x,y,v)} (g, D)$, an algorithm's parameter M and a direction's set of the form

$$Dir = \Big\{ (a^l, c^l) \in \mathbb{R}^{m+q} \mid (a^l, c^l) \neq 0, \ \ l = 1, \ldots, N \Big\}.$$

The GS algorithm used here can be represented as follows:

Step 0. Set $k := 0$, $(\bar{x}^k, \bar{y}^k, \bar{v}^k) := (x_0, y_0, v_0)$, $l := 1$. $\gamma := \gamma_-; \Delta\gamma := \gamma_+ - \gamma_-/M$.
Step 1. Starting at the point $(\bar{x}^k, \bar{y}^k, \bar{v}^k)$ construct a τ_k-critical point $(x_k, y_k, v_k) \in D$ in Problem $(\mathcal{P}(\mu))$ by applying V- or XY-procedure. Set $\zeta_k := F_1(x_k, y_k, v_k)$.

Step 2. Given a point $(a^l, c^l) \in Dir$, construct a point (z^l, w^l) such that $h(z^l, w^l) = \gamma - \zeta_k$.

Step 3. Solve the linearized problem as follows:

$$(\mathcal{PL}_l): \quad g(x,y,v) - \langle \nabla_{xv} h(z^l, w^l), (x,v) \rangle \downarrow \min_{(x,y,v)}, \quad (x,y,v) \in D.$$

Let the point $(\hat{x}, \hat{y}, \hat{v})$ be a solution to $(\mathcal{PL}_l)$.

Step 4. Starting at the point $(\hat{x}, \hat{y}, \hat{v})$ construct a δ_k-critical point $(\hat{x}_l, \hat{y}_l, \hat{v}_l)$.

Step 5. If $F_1(\hat{x}_l, \hat{y}_l, \hat{v}_l) < \zeta_k \stackrel{\triangle}{=} F_1(x_k, y_k, v_k)$, then set $(\bar{x}^{k+1}, \bar{y}^{k+1}, \bar{v}^{k+1}) := (\hat{x}_l, \hat{y}_l, \hat{v}_l)$, $k := k+1$, $l := 1$, $\gamma := \gamma_-$ and loop to Step 1.

Step 6. If $F_1(\hat{x}_l, \hat{y}_l, \hat{v}_l) \geq \zeta_k$ and $l < N$, then set $l := l+1$ and return to Step 2.

Step 7. If $F_1(\hat{x}_l, \hat{y}_l, \hat{v}_l) \geq \zeta_k$ and $l = N$, then set $\gamma := \gamma + \Delta\gamma, l := 1$ and come back to Step 2.

Step 8. If $l = N$, $F_1(\hat{x}_l, \hat{y}_l, \hat{v}_l) \geq \zeta_k \ \ \forall \gamma \in [\gamma_-, \gamma_+]$ (i.e., one-dimensional search along γ over the interval $[\gamma_-, \gamma_+]$ is terminated), then STOP; (x_k, y_k, v_k) is a critical point provided by Algorithm of global search.

Remark 1. It is clear that different values of the parameter M are responsible for the partitioning of the interval $[\gamma_-, \gamma_+]$ into a suitable number of parts to implement a passive one-dimensional search along γ. On the other hand, it is necessary to precise how to construct a direction's set Dir and, furthermore, an approximation of the level surface $h(z,w) = \gamma - \zeta_k$.

Taking into account that in contrast to the earlier papers [35–37] here due to (69)

$$h(x,v) \stackrel{\triangle}{=} \frac{\mu}{4}\|v + A_1 x\|^2, \tag{69'}$$

we have to choose $\gamma \geq \zeta_k$ so that γ_- can always be chosen as follows: $\gamma_- := \zeta_k$. Other points of the implementation of the algorithm were similar to [35–37].

Let us focus now on the construction of approximation of the level surface

$$U(\gamma) = \{(z,w) |\ h(z,w) = \gamma - \zeta_k\}.$$

Recall that, on the one hand, such an approximation should be representative enough to escape a critical point (if possible). On the other hand, if we are rather far from a global solution, then the approximation must allow us to "jump out" the critical point where we are.

Let us show how to construct an approximation. Given a set of directions

$$Dir = \{(a^l, c^l) \in \mathbb{R}^{m+q} | (a^l, c^l) \neq 0, l = 1, \ldots, N\},$$

we construct a point of an approximation $\mathcal{A}_n$ in a rather simple manner as follows:

$$(z^l, w^l) = \lambda_l (a^l, c^l),\ h(z^l, w^l) = \gamma - \zeta_k, l = 1, \ldots, N. \tag{72}$$

Due to (69 $'$) the corresponding equation

$$\frac{\mu}{4}\|c^l + A_1 a^l\|\lambda_l^2 = \gamma - \zeta_k \qquad (72\,')$$

leads us to very simple computing in order to calculate λ_l.

As the sets of directions, one can consider, for example, the set

$$Dir1 = \{(x^k + e^i, v^k + e^j), (x^k - e^i, v^k - e^j) |\ i = 1, \ldots, m, j = 1, \ldots, q\}$$

where (x^k, v^k) is the part of the current critical point (x^k, y^k, v^k); $e^i \in \mathbb{R}^m$, $e^j \in \mathbb{R}^q$ are the Euclidean basis vectors. Further, it has been employed the set

$$Dir0 = \{(a^i, b^j) | (a^i, b^j) \neq 0,\ i = 1, \ldots, m, j = 1, \ldots, q\}.$$

where a^i and b^j are, respectively, rows and columns of the matrix A_1, which specifies nonconvexity in the goal function of Problem $(\mathcal{P}(\mu))$–(60).

Note that the numbers of points in the approximations constructed are equal to $2qm$ and grow rapidly with the dimension. Therefore, we have also made the reduction of the approximations as described in [35, 36]. The first stages of computational testings of the developed GSA have been presented in [36, 37].

Here, we will show the preliminary results of further computational experiments with improved GSA described above (see Tables 2 and 3).

In particular we see in Table 2 the comparison of the results of computational solving the test-problems (generated, as above, with the help of the methodology from [5]) by GSA described above and by means of very popular package of applied software KNITRO (www.ziena.com/knitro.htm).

Since KNITRO is not able to solve bilevel problems directly, the equivalent formulation $(\mathcal{P}(\mu))$ (with $\mu = 10, 15, 20$) was used to this end.

On the other hand, GSA has been run on a computer with the processor Intel Core 2 Duo 2.0 GHz, while KNITRO used a computer with more powerful processor Intel Core 2 Quad 2.8 GHz. In Table 2 F_* is the known optimal values of the test-problems, F_{Kms} and T are the best values of the goal function and the corresponding solving time provided by KNITRO, while F_{XY}, F_V, and T stand for the best values of the cost function and the solution time obtained by GSA (using XY- or V-procedure as LSM). The bold values in Table 2 denote the successful cases when the known global solutions to the test-problems have been reached by the used algorithms.

Analyzing results of Table 2, it is easy to note that the KNITRO (multistart) was successful to find the global solutions only in 61 % of the test-problems of the middle dimension with the accuracy $\varepsilon = 10^{-2}$. Meanwhile, applying the programs implementing GSA, all considered test-problems have been solved at the same precision.

Moreover, it is not hard to see the big difference in solution time between GSA and KNITRO for the problems of middle dimension more than 10. For example, for $m = n = 30$, KNITRO worked about 1.5–2 h without reaching a global solution, meanwhile GSA provided for a global solution in 2 min approximately.

Table 2 Comparison of global search algorithm (GSA) with KNITRO

Name	F_*	KNITRO multistart		Global search			
		F_{Kms}	T	F_{XY}	T	F_V	T
5×5_1	−21	**−21**	7.7	**−21**	11.4	**−21**	7.3
5×5_2	−9	**−9**	8.7	**−9**	11.3	**−9**	4.7
5×5_3	−5	**−5**	9.9	**−5**	11.7	**−5**	4.6
10×10_1	−38	−30	1:17.2	**−38**	33.1	**−38**	23.7
10×10_2	−26	**−26**	1:17.0	**−26**	29.2	**−26**	12.4
10×10_3	−14	**−14**	1:22.4	**−14**	24.6	**−14**	16.0
15×15_1	−19	**−19**	11:08.5	**−19**	20.3	**−19**	19.0
15×15_2	−27	−19	5:43.6	**−27**	38.6	**−27**	31.5
15×15_3	−43	−35	7:01.2	**−43**	48.1	**−43**	35.5
20×20_1	−24	**−24**	19:13.7	**−24**	26.9	**−24**	45.9
20×20_2	−48	**−48**	30:20.5	**−47.999**	1:12.0	**−47.999**	1:10.9
20×20_3	−52	−32	29:54.4	**−52**	1:13.1	**−52**	52.0
30×30_1	−142	−134	1:34:34.9	**−141.997**	4:30.3	**−141.997**	1:46.8
30×30_2	−58	−38	1:31:23.5	**−58**	1:26.2	**−58**	1:51.3
30×30_3	−42	−29.999	2:39:15.4	**−42**	51.5	**−42**	1:08.9

Table 3 Testing of global search algorithm (GSA) on problems of high dimension

$m+n$	N	$LocSol_{\text{avg}}$	Loc_{avg}	St_{avg}	T_{avg}
20	1,000	146.2	2,012.7	1.9	8.57
40	1,000	13, 1284.1	3,436.6	2.1	20.69
60	100	$1.34 \cdot 10^8$	4,601.5	2.0	34.36
80	100	$1.17 \cdot 10^{11}$	6,485.1	2.1	59.51
100	100	$1.20 \cdot 10^{14}$	9,352.5	2.1	1:40.29
150	10	$3.78 \cdot 10^{21}$	8,050.3	3.0	1:52.38
200	10	$1.27 \cdot 10^{29}$	12,263.8	2.8	4:00.69
250	10	$4.27 \cdot 10^{36}$	17,704.3	2.7	8:04.28
300	10	$3.93 \cdot 10^{44}$	72,245.6	17.9	49:53.48
350	10	$2.12 \cdot 10^{52}$	216,721.1	25.2	3:56:09.12
400	10	$8.64 \cdot 10^{59}$	318,448.7	27.6	9:10:24.43

Now, let us look at Table 3 where presented the results of computational solution of the test-problems of high dimension (until $m = n = 200$) provided by the software program implemented in a computer with the processor Intel Core i5-2400 3.1 GHz. In Table 3 N is the number of test-problems in series, $LocSol_{\text{avg}}$ is an average number of local solutions which are not global in one problem of the series (this is very important difficulty index of the problem); Loc_{avg} stands for the average number of switching on of the LSM in conducting GSA; St_{avg} is the average number of iterations of GSA or critical points passed by GSA; T_{avg} is the average working time of the program implementing the GSA.

From the results of computation testing we can see, firstly, that all 2,350 randomly generated problems have been successfully solved so that, regardless the fantastic difficulty of the test-problems ($m+n=300$ and $LocSol_{\text{avg}} = 3.93 \cdot 10^{44}$, $m+n=400$ and $LocSol_{\text{avg}} = 8.64 \cdot 10^{59}$), GSA has found a global solution in every considered test-problem.

However, this version of program is characterized by the rapid increase in computing time with the growth of the dimension: for $m+n=400$ it takes more than 9 h. On the other hand, it can be explained by the number Loc_{avg} of local search applications which is more than 310 thousands.

Moreover, it is not hard to note that the number St_{avg} of approximately critical points, at which it happened an improvement of the cost function, turned out to be rather moderate with respect to the number $LocSol_{\text{avg}}$ of local solutions (different from global ones) which is varying from rather big ($m+n=60$ and $LocSol_{\text{avg}} = 1.34 \cdot 10^{8}$) until incalculable ($m+n=300$, $LocSol_{\text{avg}} = 3.93 \cdot 10^{44}$; $m+n=400$, $LocSol_{\text{avg}} = 8.64 \cdot 10^{59}$).

So, we conclude that the new results of computational solving the bilevel problem can be viewed as rather promising and competitive. Moreover, we did not be successful to find, at present, the solution's results of similar problems of such dimensions in the existing literature.

6 Concluding Remarks

In the present paper, new procedures of finding the solution to the linear complementarity problem with indefinite matrices, the Nash equilibrium in bimatrix games, and optimistic solution in quadratic-linear bilevel optimization problems have been proposed, discussed, and illustrated.

Further, a new approach based on GOCs, LSMs and GSMs, was applied in order to solve all three problems. In addition, the new results of computational solutions were presented in the paper. According to these results, the new approach has shown itself rather promising and competitive.

References

1. Alexandrov, A.D.: On surfaces represented a difference of convex functions. Proc. Acad. Sci. KSSR. Ser. Math. Mech. **3**, 3–20 (1949) (in Russian)
2. Alexandrov, A.D.: The surfaces that can be represented by a difference of convex functions. Proc. Acad. Sci. USSR. Ser. Math. Mech. **72**(4), 613–616 (1950)
3. Bonnans, J.F., Gilbert, J.C., Lemaréchal, C., Sagastizábal, C.A.: Numerical Optimization: Theoretical and Practical Aspects, 2nd edn. Springer, Berlin, Heidelberg (2006)
4. Bulatov, V.P.: The approximation method for solving some problems of mathematical programming. Applied Mathematics. Irkutsk State University, pp. 82–88 (1969)
5. Calamai, P.H., Vicente, L.N.: Generating quadratic bilevel programming test problems. ACM Trans. Math. Softw. **20**, 103–119 (1994)

6. Colson, B., Marcotte, P., Savard, C.: An overview of bilevel optimization. Ann. Oper. Res. **153**(1), 235–256 (2007)
7. Cottle, R.W., Pang, J.-S., Stone, R.E.: The linear complementarity problem. SIAM, Philadelphia (2009) [Originally published by Academic Press, Boston (1992)]
8. Dempe, S.: Foundations of Bilevel Programming. Kluwer, Dordrecht (2002)
9. Floudas, C.A., Pardalos, P.M. (eds.): Frontiers in Global Optimization. Kluwer, New York (2004)
10. Grötschel, M. (ed.): Optimization Stories. Documenta Mathematica, Bielefeld (2012)
11. Gruzdeva, T.V., Strekalovsky, A.S., Orlov, A.V., Druzhinina, O.V.: Nonsmooth minimization problems for the difference of two convex functions. Numer. Methods Program. **12**(2), 139–151 (2011) (in Russian)
12. Hiriart-Urruty, J.-B.: Generalized Differentiability, Duality and optimization for Problems dealing with Diffrence of Convex Functions. Lecture Notes in Economics and Mathematical Systems, vol. 256, pp. 37–69. Springer, Berlin (1985)
13. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Springer, Berlin (1993)
14. Hoai An, L.T., Tao, P.D.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. Ann. Oper. Res. **133**, 23–46 (2005)
15. Hoai An, L.T., Tao, P.D.: On solving linear complemetarity problems by DC programming and DCA. J. Comput. Optim. Theory Appl. **50**(3), 507–524 (2011)
16. Hoai An, L.T., Tao, P.D., Van Thoai, N., Canh, N.N.: DC programming techniques for solving a class of nonlinear bilevel programs. J. Glob. Optim. **44**(3), 313–337 (2009)
17. Horst R., Tuy H.: Global Optimization: Deterministic Approaches. Springer, Berlin (1993)
18. Malyshev, A.V., Strekalovsky, A.S.: Connection of some bilevel and nonlinear optimization problems. Russian Math. **55**(4), 83–86 (2011)
19. Malyshev, A.V., Strekalovsky, A.S.: On global search for pessimistic solution in bilevel problems. Int. J. Biomed. Softw Comput. Hum. Sci. (Special Issue on Variational Inequality and Combinatorial Problems) **18**(1), 57–61 (2011)
20. Mangasarian, O.L.: Equilibrium points in bimatrix games. J. Soc. Ind. Appl. Math. **12**, 778–780 (1964)
21. Mills, H.: Equilibrium points in finite games. J. Soc. Ind. Appl. Math. **8**(4), 397–402 (1960)
22. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Berlin (2006)
23. Orlov, A.V., Strekalovsky A.S.: Numerical search for equilibria in bimatrix games. Comput. Math. Math. Phys. **45**(6), 947–960 (2005)
24. Pang, J.-S.: Three modelling paradigms in mathematical programming. Math. Program Ser B **125**(2), 297–323 (2010)
25. Pardalos, P.M., Resende, M.G.C. (eds.): Handbook of Applied Optimization. Oxford Univesity Press, New York (2002)
26. Petrova, E.G., Strekalovsky, A.S.: The quadratic-linear bilevel problems solving via nonconvex constraint problems. Int. J. Biomed. Softw Comput. Hum. Sci. (Special Issue on Variational Inequality and Combinatorial Problems) **18**(1), 63–67 (2011)
27. Strekalovsky, A.S.: On problem of global extremum. Proc. USSR Acad. Sci. **292**(5), 1062–1066 (1987)
28. Strekalovsky, A.S.: Elements of Nonconvex Optimization. Nauka Publication, Novosibirsk (2003) (in Russian)
29. Strekalovsky, A.S.: On the minimization of the difference of convex functions on a feasible set. Comput. Math. Math. Phys. **43**(3), 399–409 (2003)
30. Strekalovsky, A.S.: Minimizing sequences in problems with D.C. constraints. Comput. Math. Math. Phys. **45**(3), 418–429 (2005)
31. Strekalovsky, A.S., Gruzdeva, T.V.: Local search in problems with nonconvex constraints. Comput. Math. Math. Phys. **47**(3), 397–413 (2007)
32. Strekalovsky, A.S., Orlov, A.V.: A new approach to nonconvex optimization. Numer. Methods Program. **8**(2), 11–27 (2007) (in Russian)

33. Strekalovsky, A.S., Orlov, A.V.: Bimatrix Games and Bilinear Programming. Physmathlit, Moscow (2007) (in Russian)
34. Strekalovsky A.S., Petrova, E.G., Mazurkevich, E.O.: On numerical solving a linear problem of complementarity. Comput. Math. Math. Phys. **49**(8), 1318–1331 (2009)
35. Strekalovsky, A.S., Orlov, A.V., Malyshev, A.V.: Local search in a quadratic-linear bilevel programming problem. Numer. Anal. Appl. **3**(1), 59–70 (2010)
36. Strekalovsky, A.S., Orlov, A.V., Malyshev, A.V.: Numerical solution of a class of bilevel programming problems. Numer. Anal. Appl. **3**(2), 165–173 (2010)
37. Strekalovsky, A.S., Orlov, A.V., Malyshev, A.V.: On computational search for optimistic solutions in bilevel problems. J. Glob. Optim. **48**, 159–172 (2010)
38. Trefethen, L.N., Bau, D.: Numerical Linear Algebra. SIAM, Philadelphia (1997)
39. Tuy, H.: D.C. optimization: theory, methods and algorithms. In: Horst, R., Pardalos, P.M. (eds.) Handbook of Global Optimization, pp. 149–216. Kluwer, Dordrecht (1995)
40. Vasilyev, F.P.: Optimization Methods. Factorial Press, Moscow (2002) (in Russian)

Variational Principles in Gauge Spaces

Mihai Turinici

1 Cârjă–Ursescu Principles

1.1 Preliminaries

Throughout this exposition, the axiomatic system to be used is Zermelo–Fraenkel's (abbreviated: (ZF)), as described in Cohen [16, Chap. 2, Sect. 3]. The notations and basic facts about these are more or less usual. Some important ones are given below.

(A) Let X be a nonempty set. By a *relation* over it, we mean any (nonempty) part $\mathcal{R}$ of $X \times X$; in this case, $(X, \mathcal{R})$ is called a *relational structure*. As usual, we may regard $\mathcal{R}$ as a mapping from X to $\mathcal{P}(X)$ (=the class of all subsets in X). Precisely, for each $x \in X$, denote $X(x, \mathcal{R}) = \{y \in X; x\mathcal{R}y\}$ (the *section* of $\mathcal{R}$ through x); then, the mapping in question is $[\mathcal{R}(x) = X(x, \mathcal{R}), x \in X]$. Call $\mathcal{R}$, *proper* when $\mathcal{R}(x) \neq \emptyset$, for all $x \in X$; note that, in such a case, $\mathcal{R}$ appears as a mapping between X and $\mathcal{P}_0(X)$ (=the class of all nonempty parts in X). This will be also referred to as: $(X, \mathcal{R})$ is a *proper* relational structure.

Call the relation $(\leq)$ over X, *quasi-order*, provided it is *reflexive* $[x \leq x, \forall x \in X]$ and *transitive* $[x \leq y$ and $y \leq z$ imply $x \leq z]$. If, in addition, $(\leq)$ is *antisymmetric* $[x \leq y$ and $y \leq x$ imply $x = y]$, then it is called a *(partial) order* on X. Let $(X, \leq)$ be a partially ordered structure. By a *chain* in X we mean any totally ordered part C of it (in the sense: for each $x, y \in C$, either $x \leq y$ or $y \leq x$). Given the subset Y of X, call $u \in X$ an *upper bound* of Y, provided $y \leq u$, for all $y \in Y$; when such elements exist, we say that Y is *bounded above*. Further, call $z \in X$, *maximal* (modulo $(\leq)$) provided $X(z, \leq) = \{z\}$; i.e.: $[z \leq w \in X$ implies $z = w]$. Finally, let us say that $(\leq)$ is a *Zorn order* when, for each starting $u \in X$ there exists a $(\leq)$-maximal element $v \in X$ with $u \leq v$.

M. Turinici (✉)
"A. Myller" Mathematical Seminar, "A. I. Cuza" University, 700506 Iaşi, Romania
e-mail: mturi@uaic.ro

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_24,

For each couple A,B of nonempty sets, let $\mathcal{F}(A,B)$ stand for the class of all functions from A to B. In particular, if $A=B$, we write $\mathcal{F}(A)$ in place of $\mathcal{F}(A,A)$. On the other hand, when $A=N:=\{0,1,\ldots\}$ (=the set of all natural numbers), we denote $\mathcal{F}(N,B)$ as $\mathcal{S}(B)$. Each element $x\in\mathcal{S}(B)$ is referred to as a *sequence* in B; and denoted as $(x(n);n\geq 0)$ or $(x_n;n\geq 0)$; when no confusion can arise, we simplify this notation as $(x(n))$ or (x_n), respectively.

(B) By a *pseudometric* on X we mean any map $d:X\times X\to R_+:=[0,\infty[$. If, in addition, d is *symmetric* $[d(x,y)=d(y,x),\forall x,y\in X]$, *triangular* $[d(x,z)\leq d(x,y)+d(y,z),\forall x,y,z\in X]$, and *reflexive* $[d(x,x)=0,\forall x\in X]$, then it is called a *semimetric* on X; and (X,d) is termed a *semimetric space*. Moreover, if in addition to this, d is *sufficient* $[x,y\in X,\ d(x,y)=0\Longrightarrow x=y]$, we say that it is a *metric* on X; and (X,d) is referred to as a *metric space*.

Let (X,d) be a semimetric space. Denote, for each subset Y of X

$$\operatorname{diam}(Y)=\sup\{d(x_1,x_2);x_1,x_2\in Y\}\ \text{(the diameter of } Y);$$

when $\operatorname{diam}(Y)<\infty$, we say that Y is *d-bounded*. We introduce a d-convergence and a d-Cauchy structure on X as follows. Given the sequence (x_n) in X and the point $x\in X$, we say that (x_n), *d-converges* to x (written as: $x_n\xrightarrow{d}x$), if $d(x_n,x)\to 0$ as $n\to\infty$; i.e.,

$$\forall\varepsilon>0,\exists i=i(\varepsilon):\ n\geq i\Longrightarrow d(x_n,x)<\varepsilon.$$

The set of all such points x will be denoted $\lim_n(x_n)$; if it is nonempty, then (x_n) is called *d-convergent*. Note that, in this case, $\lim_n(x_n)$ may be not a singleton; but, when $d(.,.)$ is a metric, this is retainable. Further, call the sequence (x_n), *d-Cauchy* when $d(x_m,x_n)\to 0$ as $m,n\to\infty$, $m\leq n$; i.e.,

$$\forall\varepsilon>0,\exists j=j(\varepsilon):\ j\leq m\leq n\Longrightarrow d(x_m,x_n)<\varepsilon;$$

or, equivalently:

$$\forall\eta>0,\exists k=k(\eta):\ \operatorname{diam}(\{x_n;n\geq k\})<\eta.$$

Note that, as d is semimetric, any d-convergent sequence is d-Cauchy. The reciprocal is not in general true; when it holds, we say that (X,d) is *complete*.

(C) Remember that an outstanding part of (ZF) is the *Axiom of Choice* (abbreviated: (AC)) which, in a convenient manner, may be written as

> For each nonempty set X, there exists a (selective) function $f:\mathcal{P}_0(X)\to X$, with $f(Y)\in Y,\forall Y\in\mathcal{P}_0(X)$.

There are many logical equivalents of (AC); see, for instance, Moore [38, Appendix 2]. A basic one is the *Zorn Maximal Principle* (in short: (ZMP)), expressed as (cf. Bourbaki [7]):

> Let the partially ordered structure $(X,\leq)$ be inductive. Then, $(\leq)$ is a Zorn order.

[Here, *inductive* means: any totally ordered part C of X is bounded above]. Sometimes, when the ambient set X is endowed with denumerable type structures, the existence of maximal elements may be determined by using a weaker form of (AC), called: *Dependent Choice Principle*. 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 result, referred to as the *Finite Dependent Choice property* (in short: (DC-fin)) is available in the strongly reduced Zermelo–Fraenkel system (ZF-AC). Given $a \in X$, let us say that the k-sequence $F : N(k,>) \to X$ (where $k \geq 2$) is $(a,\mathcal{R})$-*iterative* provided $F(0) = a$ and $F(i)\mathcal{R}F(i+1)$, for all $i \in N(k-1,>)$.

Lemma 1. *Let the relational structure $(X,\mathcal{R})$ be proper. Then, for each $k \geq 2$, the following property holds:*

$$\text{for each } a \in X, \text{ there exists an } (a,\mathcal{R})\text{-iterative } k\text{-sequence.} \tag{1}$$

Proof. Denote by $\pi(k)$ the conclusion above. Clearly, $\pi(2)$ is true; just take $b \in \mathcal{R}(a)$ and define $F : N(2,>) \to X$ as: $F(0) = a$, $F(1) = b$. Assume that $\pi(k)$ is true, 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 $(a,\mathcal{R})$-iterative k-sequence, assured by hypothesis. As $\mathcal{R}$ is proper, $\mathcal{R}(F(k-1))$ 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 $(a,\mathcal{R})$-iterative $(k+1)$-sequence; and then, we are done.

Now, it is natural to see what happens when k "tends to infinity." At a first glance, the following *Dependent Choice Principle* (in short: (DC)) is obtainable in (ZF-AC) from this "limit" process. Given $a \in X$, let us say that the sequence $(x_n; n \geq 0)$ in X is $(a;\mathcal{R})$-*iterative* provided $[x_0 = a; x_{n+1} \in \mathcal{R}(x_n), \forall n]$.

Proposition 1. *Let the relational structure $(X,\mathcal{R})$ be proper. Then, for each $a \in X$ there is an $(a,\mathcal{R})$-iterative sequence in X.*

Concerning this aspect, we stress that—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 [6] and Tarski [42]—is deductible from (AC), but not conversely; cf. Wolk [52]. Moreover, by the developments in Moskhovakis [39, Chap. 8], and Schechter [41, Chap. 6], the reduced system (ZF-AC+DC) it large enough so as to cover the "usual" mathematics; see also Moore [38, Appendix 2, Table 4].

(D) Let $(\mathcal{R}_n; n \geq 0)$ 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 $(a;(\mathcal{R}_n; n \geq 0))$-*iterative* provided $[x_0 = a$, $x_{n+1} \in \mathcal{R}_n(x_n), \forall n]$. The following *Diagonal Dependent Choice Principle* (in short: (DDC)) is also taken into consideration.

Proposition 2. *Let $(\mathcal{R}_n; n \geq 0)$ be a sequence of proper relations on X. Then, for each $a \in X$, there exists at least one $(a;(\mathcal{R}_n; n \geq 0))$-iterative sequence in X.*

Clearly, (DDC) includes (DC) to which it reduces when $(\mathcal{R}_n; n \geq 0)$ is constant. The reciprocal of this is also true. In fact, letting the premises of (DDC) hold, put $P = N \times X$; and let $\mathcal{S}$ be the relation over P introduced as

$$\mathcal{S}(i,x) = \{i+1\} \times \mathcal{R}_i(x), \ \ (i,x) \in P.$$

It will suffice applying (DC) to $(P, \mathcal{S})$ and $b := (0,a) \in P$ to get the conclusion in the statement; we do not give details.

Summing up, (DDC) is provable in (ZF-AC+DC). This is valid as well for its variant, referred to as: the *Selected Dependent Choice Principle* (in short: (SDC)).

Proposition 3. *Let the map $F : N \to \mathcal{P}_0(X)$ and the relation $\mathcal{R}$ over X fulfill*

$$(\forall n \in N)\text{: } \mathcal{R}(x) \cap F(n+1) \neq \emptyset, \ \ \forall x \in F(n) \ \ \textit{[F is $\mathcal{R}$-chainable].}$$

Then, for each $a \in F(0)$ there exists a sequence $(x(n); n \geq 0)$ in X with

$$x(0) = a; \ x(n) \in F(n), \ \forall n; \ x(n) \mathcal{R} x(n+1), \ \forall n. \tag{2}$$

As before, (SDC) $\Longrightarrow$ (DC) ($\Longleftrightarrow$ (DDC)); just take $F(n) = X$, $n \geq 0$. But, the reciprocal is also true, in the sense: (DDC) $\Longrightarrow$ (SDC). This follows from

Proof. **(Proposition 3)** Let the premises of (SDC) be admitted. Define a sequence of relations $(\mathcal{R}_n; n \geq 0)$ over X as: for each $n \geq 0$,

$$\mathcal{R}_n(x) = \mathcal{R}(x) \cap F(n+1), \ \text{ if } x \in F(n); \ \mathcal{R}_n(x) = \{x\}, \text{ if } x \in X \setminus F(n).$$

Clearly, $\mathcal{R}_n$ is proper, for all $n \geq 0$. So, by (DDC), it follows that, for the starting $a \in F(0)$, there exists an $(a; (\mathcal{R}_n; n \geq 0))$-iterative sequence $(x(n); n \geq 0)$ in X. Combining with the very definition of $(\mathcal{R}_n; n \geq 0)$, yields the desired conclusion.

In particular, when $\mathcal{R} = X \times X$, F is $\mathcal{R}$-chainable. The corresponding variant of (SDC) is just the Denumerable Axiom of Choice (in short: (AC-N)):

Proposition 4. *Let $F : N \to \mathcal{P}_0(X)$ be a function. Then, for each $a \in F(0)$ there exists a function $f : N \to X$ with $f(0) = a$ and $f(n) \in F(n)$, $\forall n \geq 0$.*

Remark 1. Note that, as a consequence of the above facts, (DC) $\Longrightarrow$ (AC-N) in (ZF-AC). A direct verification of this is obtainable by taking $P = N \times X$ and introducing the relation over it: $[\mathcal{R}(n,x) = \{n+1\} \times F(n+1), \ n \geq 0, \ x \in X]$; we do not give details. The reciprocal of the written inclusion is not true; see Moskhovakis [39, Chap. 8, Sect. 8.25] for details.

(E) A direct application of these facts is the following. Let X be a nonempty set. Given some property π involving $\mathcal{P}_0(X)$, denote by (π) the subclass of all $Y \in \mathcal{P}_0(X)$ fulfilling it. In this case, let us say that π is *countably inductive* provided:

$$(Y_i \in (\pi), \forall i \geq 0) \text{ implies } Y := \cap \{Y_i; i \geq 0\} \in (\pi) \text{ (hence, } Y \in \mathcal{P}_0(X)).$$

A basic example of this type is the following. Let $(X,\leq)$ be a quasi-ordered structure. Call $Z\in\mathcal{P}_0(X)$, $(\leq)$-*cofinal* in X when $[X(u,\leq)\cap Z\neq\emptyset,\ \forall u\in X]$. In addition, let us say that $Z\in\mathcal{P}_0(X)$ is $(\leq)$-*invariant* provided $w\in Z$ implies $X(w,\leq)\subseteq Z$. The intersection of these properties will be referred to as: Z is $(\leq)$-cofinal-invariant; in short: $(\leq)$-cof-inv. The following *Cof-inv statement* is available in (ZF-AC+DC). Call $(X,\leq)$, *sequentially inductive* provided each ascending sequence in X has an upper bound (modulo $(\leq)$).

Proposition 5. *Assume that $(X,\leq)$ is sequentially inductive. Then, the $(\leq)$-cof-inv property is countably inductive.*

Proof. Let $(F(i);i\geq 0)$ be a sequence in $\mathcal{P}_0(X)$ such that: $F(i)$ is $(\leq)$-cof-inv, for each $i\geq 0$. We intend to show that $Y:=\cap\{F(i);i\geq 0\}$ is endowed with the same property. Clearly, Y is $(\leq)$-invariant; but, for the moment, $Y=\emptyset$ cannot be avoided. We show that Y is $(\leq)$-cofinal too; hence, nonempty. Let $u\in X$ be arbitrary fixed. Further, let the relation $\mathcal{R}$ over X be introduced as $[\mathcal{R}(x)=X(x,\leq),\ x\in X]$; i.e.: $\mathcal{R}$ is the *graph* of $(\leq)$. By the $(\leq)$-cofinal property,

$$\mathcal{R}(x)\cap F(i)=X(x\leq)\cap F(i)\neq\emptyset,\ \forall i\geq 0,\forall x\in X. \tag{3}$$

In particular, this tells us that $X(u,\leq)\cap F(0)\neq\emptyset$; let a be one of its elements. From the SDC it follows that, for this starting element, there exists a sequence $(x_n;n\geq 0)$ in X with

$$x_0=a;\ x_n\in F(n),\forall n;\ x_n\leq x_{n+1},\forall n. \tag{4}$$

As $(X,\leq)$ is sequentially inductive, there exists some $v\in X$ with $x_n\leq v$, $\forall n$. In particular, from $u\leq a=x_0\leq v$, one has $u\leq v$. Moreover, by the $(\leq)$-invariance properties of $(F(n);n\geq 0)$, we have $v\in F(n)$, $\forall n$; hence $v\in Y$. Putting these together, one gets the desired fact.

(F) Concerning the metrical structures to be considered, some basic examples are constructed below.

Let P be a nonempty set. The simplest metric over P is:

$$(s,t\in P)\text{: }[d(s,t)=0,\text{ if }s=t]\text{ and }[d(s,t)=1,\text{ if }s\neq t];$$

it will be referred to as: the *discrete metric* on P. A "sequential" version of it is the following. Remember that $\mathcal{S}(P)$ stands for the class of all sequences in P. Fix some $a\in P$; and put $X=\{x\in\mathcal{S}(P);x(0)=a\}$. Define a mapping $d_\infty:X\times X\to R_+$ as

$$d_\infty(x,y)=\sum_n 2^{-n}d(x(n),y(n)),\ \text{ for all } x=(x(n)),\ y=(y(n))\text{ in } X.$$

It is not hard to see that d_∞ is a metric on X. A natural question to be discussed here is the completeness property. In this direction, we have

Proposition 6. *Under the above conventions, the metrical structure (X,d_∞) is complete: each d_∞-Cauchy sequence in X is d_∞-convergent.*

Proof. Let $(x^n; n \geq 0)$ be a sequence in X; it may be written as

$$(x^n = (x^n(0), x^n(1), \ldots) = (a, x^n(1), \ldots); n \geq 0).$$

Assume that $(x^n; n \geq 0)$ is d_∞-Cauchy. This may be also characterized as:

$$\forall \varepsilon > 0 \colon C(\varepsilon) := \{n \in N; n \leq p \leq q \Longrightarrow d_\infty(x^p, x^q) < \varepsilon\} \neq \emptyset.$$

As a consequence, the map $\varepsilon \mapsto C(\varepsilon)$ is increasing on $R_+^0 :=]0, \infty[$, in the sense: $\varepsilon_* < \varepsilon^*$ implies $C(\varepsilon_*) \subseteq C(\varepsilon^*)$; so that, the map $\varepsilon \mapsto \Gamma(\varepsilon) := \min[C(\varepsilon)]$ is decreasing on R_+^0: $\varepsilon_* < \varepsilon^*$ implies $\Gamma(\varepsilon_*) \geq \Gamma(\varepsilon^*)$. Let $(\varepsilon_n; n \geq 0)$ be a strictly descending sequence in R_+^0 with $\varepsilon_n < 2^{-n}$, $\forall n$ (hence $\varepsilon_n \to 0$). Denote for simplicity

$$m(k) = \Gamma(\varepsilon_k),\ n(k) = m(k) + k,\ k \geq 0.$$

By the properties above, the map $k \mapsto m(k)$ is increasing; hence, the map $k \mapsto n(k)$ is strictly increasing. For the moment, it is clear that $x^{n(0)}(0) = x^p(0) = a$, $\forall p \geq n(0)$. Further, by the very definition of these maps, $n(1) \leq p \leq q \Longrightarrow d_\infty(x^p, x^q) < \varepsilon_1$. Combining with the definition of d_∞ gives $2^{-1} d(x^p(1), x^q(1)) < \varepsilon_1$, if $n(1) \leq p \leq q$; so that (as $\varepsilon_1 < 2^{-1}$), $x^{n(1)}(1) = x^p(1)$, for all $p \geq n(1)$. The procedure may continue indefinitely; it gives us a strictly ascending sequence of ranks $(n(i); i \geq 0)$ with

$$x^{n(i)}(i) = x^p(i), \text{ for all } p \geq n(i) \text{ and all } i \geq 0. \tag{5}$$

Let $y = (y(i); i \geq 0)$ be the "diagonal" sequence $(y(i) = x^{n(i)}(i); i \geq 0)$; clearly, it is an element of X. We claim that our initial sequence $(x^n; n \geq 0)$ is convergent (modulo d_∞) to y. In fact, let $\varepsilon > 0$ be arbitrary fixed; and $h = h(\varepsilon)$ be such that $[2^{-j} < \varepsilon, \forall j \geq h]$. For each $n \geq n(h)$ we have (by the above properties)

$$d(x^n, y) = \sum_{i \leq h} 2^{-i} d(x^n(i), x^{n(i)}(i)) + \sum_{i > h} 2^{-i} d(x^n(i), x^{n(i)}(i)) =$$
$$\sum_{i > h} 2^{-i} d(x^n(i), x^{n(i)}(i)) \leq \sum_{i > h} 2^{-i} = 2^{-h} < \varepsilon;$$

and, from this, we are done.

1.2 (DC) $\Longrightarrow$ (CU) $\Longleftrightarrow$ (BB)

Let X be a nonempty set. Take a *quasi-order* $(\leq)$ over it; and a function $\varphi : X \to R \cup \{-\infty, \infty\}$. Call the point $z \in X$, $(\leq, \varphi)$-*maximal* when: $z \leq w \in X$ implies $\varphi(z) = \varphi(w)$; the set of all these will be denoted as $\max(X; \leq; \varphi)$. Sufficient conditions for existence of such elements are to be written in terms of the function φ belonging to certain subclasses of $\mathcal{F}(X, R \cup \{-\infty, \infty\})$. The basic ones are listed below:

(P0) general case ($\varphi(X) \cap \{-\infty, \infty\} \neq \emptyset$ cannot be avoided)
(P1) $\varphi(X) \subseteq R \cup \{\infty\}$ and φ is bounded below ($\inf \varphi(X) > -\infty$)
(P2) $\varphi(X) \subseteq R \cup \{\infty\}$ and φ is positive ($\inf \varphi(X) \geq 0$)
(P3) $\varphi(X) \subseteq R$ and φ is bounded below ($\inf \varphi(X) > -\infty$)
(P4) $\varphi(X) \subseteq R$ and φ is positive ($\inf \varphi(X) \geq 0$)

(P5) $\varphi(X) \subseteq R$ and φ is bounded ($-\infty < \inf \varphi(X) \leq \sup \varphi(X) < \infty$)
(P6) $\varphi(X) \subseteq R$ and φ is bounded positive ($0 \leq \inf \varphi(X) \leq \sup \varphi(X) < \infty$).

The following "multiple" (global) ordering principle is now considered:

Theorem 1. *Suppose that*

(b01) $(X, \leq)$ *is sequentially inductive:*
each ascending sequence has an upper bound (modulo $(\leq)$*)*
(b02) φ *is* $(\leq)$*-decreasing* ($x_1 \leq x_2 \Longrightarrow \varphi(x_1) \geq \varphi(x_2)$)
(b03) φ *belongs to the subclass (Pj), for some* $j \in \{0,1,2,3,4,5,6\}$.

Then, $\max(X; \leq; \varphi)$ *is*
(i) $(\leq)$*-cofinal in X [for each* $u \in X$ *there exists* $v \in \max(X; \leq; \varphi)$ *with* $u \leq v$*]*
(ii) $(\leq)$*-invariant in X* [$z \in \max(X; \leq; \varphi) \Longrightarrow X(z, \leq) \subseteq \max(X; \leq; \varphi)$].

Denote the obtained (global) results as (CU;Pj), where $j \in \{0,1,2,3,4,5,6\}$; these will be referred to as (global) Cârjă–Ursescu ordering principles. The relationships between them are clarified in the (global) Equivalence statement below:

Lemma 2. *We have [in (ZF-AC)]:*
(i) *(CU;P(j))* $\Longleftrightarrow$ *(CU;P(j+1)),* $\forall j \in \{1,3,5\}$
(ii) *(CU;P5)* $\Longrightarrow$ *(CU;P0)* $\Longrightarrow$ *(CU;P1)* $\Longrightarrow$ *(CU;P3)* $\Longrightarrow$ *(CU;P5).*
Hence, all these principles are equivalent in (ZF-AC).

Proof. **(j)** The inclusions (CU;P(j+1)) $\Longrightarrow$ (CU;P(j)) for $j \in \{1,3,5\}$ are deductible from the following remark: if the function φ is like in (CU;P(j)), then its translate $[\psi(.) = \varphi(.) - \inf \varphi(X)]$ fulfills the requirements of (CU;P(j+1)). This, and the reciprocal inclusions being fulfilled, proves the first part.

(jj) All inclusions in the second part, with the exception of the first one are clear.

(jjj) It remains to verify the quoted relation. Let the premises of (CU;P0) hold. Define the function $\chi : X \to [0, \pi]$ as $[\chi(x) = A(\varphi(x)), x \in X]$; where

$$A(t) = \pi/2 + \operatorname{arctg}(t) \text{ if } t \in R;\ A(-\infty) = 0;\ A(\infty) = \pi.$$

Clearly, χ is $(\leq)$-decreasing and belongs to the subclass (P5). Therefore, by the conclusion of (CU;P5), for each $u \in X$ there exists a $(\leq, \chi)$-maximal $v \in X$ with $u \leq v$. This, along with $\max(X; \leq; \varphi) = \max(X; \leq; \chi)$, gives the desired conclusion.

Note that the obtained relations cannot assure us that these principles are deductible in (ZF-AC+DC). This, however, holds; as results from

Proposition 7. *We have [in (ZF-AC)] (DC)* $\Longrightarrow$ *(CU;P5); hence (by the above) (DC)* $\Longrightarrow$ *(CU;Pj), for each* $j \in \{0,1,2,3,4,5,6\}$.

Proof. Assume that $(X, \leq)$ is sequentially inductive and φ is $(\leq)$-decreasing; in addition, let φ belong to the subclass (P5). Define the function $\beta : X \to R$ as: $\beta(v) := \inf \varphi(X(v, \leq))$, $v \in X$. Clearly, β is $(\leq)$-increasing, bounded, and

$$\sup \varphi(X) \geq \varphi(v) \geq \beta(v) \geq \inf \varphi(X),\ \forall v \in X. \tag{6}$$

Moreover, the $(\leq)$-decreasing property of φ gives a characterization like

$$v \text{ is } (\leq, \varphi)\text{-maximal iff } \varphi(v) = \beta(v). \tag{7}$$

Assume by contradiction that the conclusion of (CU;P5) would be false; i.e. (by the preceding observation) there must be some $u \in X$ such that:

(b04) for each $v \in X_u := X(u, \leq)$, one has $\varphi(v) > \beta(v)$.

Consequently (for all such v), $\varphi(v) > (1/2)(\varphi(v) + \beta(v)) > \beta(v)$; hence

$$v \leq w \text{ and } (1/2)(\varphi(v) + \beta(v)) > \varphi(w), \tag{8}$$

for at least one w (belonging to X_u). The relation $\mathcal{R}$ over X_u given in this way is proper. So, by (DC), there must be a sequence (u_n) in X_u with $u_0 = u$ and

$$u_n \leq u_{n+1},\ (1/2)(\varphi(u_n) + \beta(u_n)) > \varphi(u_{n+1}), \text{ for all } n. \tag{9}$$

We have thus constructed an ascending sequence (u_n) in X_u for which the real sequence $(\varphi(u_n))$ is strictly descending and bounded below; hence, $\lambda := \lim_n \varphi(u_n)$ exists in R. As $(X, \leq)$ is sequentially inductive, there exists $v \in M$ such that $u_n \leq v$, for all n. Clearly, $\varphi(u_n) \geq \varphi(v)$, $\forall n$; and (by the properties of β) $\varphi(v) \geq \beta(v) \geq \beta(u_n)$, $\forall n$. The former of these relations gives $\lambda \geq \varphi(v)$. On the other hand, the latter of these relations yields (by the definition of (u_n)), $(1/2)(\varphi(u_n) + \beta(v)) > \varphi(u_{n+1})$, for all n. Passing to limit as $n \to \infty$ gives $(\varphi(v) \geq)\beta(v) \geq \lambda$; so, combining with the preceding one, $\varphi(v) = \beta(v)(= \lambda)$, contradiction. Hence, the working assumption above cannot hold; and conclusion follows.

Finally, note that (CU;P0) is (the global variant of) the 1993 Cârjă–Ursescu variational principle [13] (in short: (CU)). Moreover, (CU;P3) is just (the global variant of) the 1976 Brezis–Browder ordering principle [9] (abbreviated as: (BB)). In a "local" formulation, this last result may be expressed as follows:

Theorem 2. *Suppose that the quasi-ordered structure $(X, \leq)$ is sequentially inductive and the function $\varphi : X \to R$ is $(\leq)$-decreasing, bounded from below. Then, for each $u \in X$ there exists $v \in X$, with*

(a) *$u \leq v$;* ***(aa)*** *$v \leq w \in X$ implies $\varphi(v) = \varphi(w)$.*

Finally, note that, a slightly different argument for getting the same conclusion may be found in Cârjă et al. [14, Chap. 2, Sect. 2.1]. Further metrical aspects of these questions were discussed in Turinici [43].

1.3 (BB) $\Longrightarrow$ (KP)

In the following, the relationships between (BB) [or, equivalently, (CU)] and some other maximal results in the area are discussed.

Let $(X,\leq)$ be a quasi-ordered structure; and $\varphi : X \to R_+ \cup \{\infty\}$ be a function. The following *almost regular* version of (CU) (in short: (CU-areg)) is available:

Theorem 3. *Assume that $(X,\leq)$ is sequentially inductive, φ is $(\leq)$-decreasing, and*

(c01) $(X,\leq)$ is almost regular (modulo φ):
$\forall x \in X$, $\forall \varepsilon > 0$, $\exists y = y(x,\varepsilon) \geq x$ such that $\varphi(y) \leq \varepsilon$.

Then, for each $u \in X$ there exists $v \in X$ with $u \leq v$ and $\varphi(v) = 0$ (hence, necessarily, v is $(\leq,\varphi)$-maximal).

Proof. By the almost regular condition, there must be some $z \geq u$ with $\varphi(z) < \infty$. Clearly, (BB) applies to $X(z,\leq)$ and $(\leq,\varphi)$. So, for $z \in X(z,\leq)$, there exists $v \in X(z,\leq)$ with
(i) $z \leq v$ (hence $u \leq v$); **(ii)** v is $(\leq,\varphi)$-maximal in $X(z,\leq)$.
Suppose by contradiction that $\gamma := \varphi(v) > 0$; and fix some β in $]0,\gamma[$. Again by the almost regular condition, there exists $y = y(v,\beta) \geq v$ (hence $y \in X(z,\leq)$) with $\varphi(y) \leq \beta < \gamma (= \varphi(v))$; impossible, by the second conclusion above. Hence, $\varphi(v) = 0$; and the proof is complete.

By this reasoning, (CU-areg) is deductible from (BB). The converse inclusion is also true; to verify it, we need some conventions. By a (generalized) *pseudometric* over X we shall mean any map $d : X \times X \to R_+ \cup \{\infty\}$. Fix such an object; supposed to be *reflexive* $[d(x,x) = 0, \forall x \in X]$. Call $z \in X$, $(\leq,d)$-*maximal*, if: $u,v \in X$ and $z \leq u \leq v$ imply $d(u,v) = 0$. Note that, if d is (in addition) *sufficient* $[d(x,y) = 0 \Longrightarrow x = y]$, the $(\leq,d)$-maximal property becomes: $w \in X, z \leq w \Longrightarrow z = w$ (and reads: z is *strongly* $(\leq)$-*maximal*). So, existence results involving such points may be viewed as "metrical" versions of the ZMP we just encountered; cf. Moore [38, Chap. 4, Sect. 4]. A natural way of deriving them is to start from the fact that, in terms of the associated function $\varphi_d(x) = \sup\{d(u,v); x \leq u \leq v\}, x \in X$, this property may be characterized as: $\varphi_d(z) = 0$. So, a basic source for determining such elements is (CU-areg) above (applied to the underlying function). To do this, note that φ_d is $(\leq)$-decreasing. On the other hand, the almost regularity (modulo φ_d) condition may be written as:

(c02) $(X,\leq)$ is weakly regular (modulo d): $\forall x \in X, \forall \varepsilon > 0$,
$\exists y = y(x,\varepsilon) \geq x$ such that $y \leq u \leq v \Longrightarrow d(u,v) \leq \varepsilon$.

Putting these together, it results (by the preceding ordering principle) the maximal statement due to Kang and Park [34] (in short: (KP)):

Theorem 4. *Assume that the quasi-ordered reflexive pseudometric space $(X,\leq,d)$ is such that $(X,\leq)$ is sequentially inductive and weakly regular (modulo d). Then, for each $u \in X$ there exists a $(\leq,d)$-maximal $v \in X$ with $u \leq v$.*

Clearly, (BB) $\Longrightarrow$ (KP). The reciprocal implication holds too; as results from

Proposition 8. *We have [in (ZF-AC)] (KP) $\Longrightarrow$ (BB); hence, (KP) $\Longleftrightarrow$ (BB).*

Proof. Let $\varphi : X \to R$ be as in the premise of (BB). Denote $d(x,y) = |\varphi(x) - \varphi(y)|, x,y \in X$; this map is a semimetric on X. Further, let $\beta(.)$ stand for the associated (to φ) function $[\beta(v) := \inf \varphi(X(v,\leq)),\ v \in X]$. Assume that the conclusion in (BB) is false: there must be some $u \in X$ such that (cf. a previous argument):

for each $v \in X_u := X(u,\leq)$, one has $\varphi(v) > \beta(v)$.

Clearly, $(X_u,\leq)$ is sequentially inductive. Moreover, let $v \in X_u$ be arbitrary fixed; hence, $\varphi(v) > \beta(v)$. Given ε in $]0, \varphi(v) - \beta(v)[$, there exists $y \in X(v,\leq) \subseteq X_u$ with $\beta(v) \leq \varphi(y) < \beta(v) + \varepsilon < \varphi(v)$. This tells us that, if $y \leq s \leq t$, then $s,t \in X_u$ and

$$\beta(v) + \varepsilon > \varphi(y) \geq \varphi(s) \geq \varphi(t) \geq \beta(v);$$

whence $d(s,t) = \varphi(s) - \varphi(t) < \varepsilon$; so that, $(X_u,\leq)$ is weakly regular (modulo d). Summing up, (KP) applies to $(X_u,\leq,d)$; so that, for the given $u \in X_u$ there exists a $(\leq,d)$-maximal $v \in X_u$ with $u \leq v$. By the very definition of $d(.,.)$, the obtained element is $(\leq,\varphi)$-maximal in X; contradiction. This ends the argument.

Summing up, we have established the inclusion chain: BB $\Longrightarrow$ (CU-areg) $\Longrightarrow$ (KP) $\Longrightarrow$ (BB). Hence, all these ordering principles are nothing but logical equivalents of (BB) or (CU). It is natural to ask whether the maximal principles in Altman [1] and Turinici [45] enter in this scheme. A (positive) answer to this is available with the "diagonal" version of (DC); we do not give details. This conclusion comprises as well some further extensions of these results to sequential convergence structures (as in Kasahara [35]) or pseudo-uniform structures (constructed under the model in Nachbin [40, Chap. 2, Sect. 2]); we refer to the paper by Turinici [50] for details.

1.4 Ekeland Variational Principles

A basic application of these facts is to Ekeland variational principles.

(A) Let (X,d) be a metric space; and $\varphi : X \to R \cup \{\infty\}$ be a function, with

(d01) φ is proper: $\mathrm{Dom}(\varphi) := \{x \in X; \varphi(x) < \infty\} \neq \emptyset$.

The quasi-order $(\preceq_{(d,\varphi)})$ over X introduced as

(d02) $(x_1,x_2 \in X)$: $x_1 \preceq_{(d,\varphi)} x_2$ iff $d(x_1,x_2) + \varphi(x_2) \leq \varphi(x_1)$

is antisymmetric—hence, an ordering—on $\mathrm{Dom}(\varphi)$. It is our objective in the following to determine sufficient conditions under which $(\preceq_{(d,\varphi)})$ be a Zorn order on $\mathrm{Dom}(\varphi)$. Precisely, these consist in

(I) Boundedness properties of the objective function:
the classes [(Pj); $j \in \{1,2,3,4,5,6\}$] we just encountered

(II) Boundedness properties of the ambient metric space:

(L1) general case: [$\mathrm{diam}(X) = \infty$] cannot be avoided
(L2) (X,d) is bounded: $\mathrm{diam}(X) < \infty$

(III) Global completeness property of the whole family of objects:

(gdc) (X,d,φ) is globally descending complete: each d-Cauchy sequence (x_n) in X with $(\varphi(x_n))=$descending, is d-convergent to some $x\in X$; with, in addition, $[\varphi(x_n)\geq\varphi(x),\forall n]$.

Our main result is

Theorem 5. *Assume that the proper function φ belongs to the class (Pj), for some $j\in\{1,2,3,4,5,6\}$, the metric space (X,d) belongs to the class (Lm), for some $m\in\{1,2\}$, and the triple (X,d,φ) has the property (gdc). Then, for each $u\in Dom(\varphi)$ there exists $v\in Dom(\varphi)$ with*

(a) *$d(u,v)\leq\varphi(u)-\varphi(v)$ (hence $\varphi(u)\geq\varphi(v)$)*

(aa) *$d(v,x)>\varphi(v)-\varphi(x)$, for each $x\in X\setminus\{v\}$.*

Denote the obtained statements as (EVP;Pj;Lm;gdc), where $j\in\{1,2,3,4,5,6\}$, $m\in\{1,2\}$; these will be referred to as: "composed" Ekeland variational principles. The relationships between them are given by

Lemma 3. *We have [in (ZF-AC)],*

(i) *(EVP;P(j);Lm;gdc) $\Longleftrightarrow$ (EVP;P(j+1);Lm;gdc), $\forall j\in\{1,3,5\}$*

(ii) *the propositional map $(j,m)\mapsto(EVP;Pj;Lm;gdc)$ is decreasing:*
if $(j,m)\leq(j',m')$, then (EVP;Pj;Lm;gdc) $\Longrightarrow$ (EVP;Pj';Lm';gdc).

Proof. (i) Let the triple (X,d,φ) be as in the premise of (EVP;P(j);Lm;gdc). Then, the associated triple (X,d,ψ), where $[\psi(.):=\varphi(.)-\inf\varphi(X)]$, fulfills conditions of (EVP;P(j+1);Lm;gdc); and, from this, we are done.

(ii) Evident, by the definition of the subclasses in question.

(B) Now, the property (gdc) is obtainable as an intersection of three components:

(III-a) Completeness properties for the whole family of objects

(B1) (X,d,φ) is descending complete: each d-Cauchy sequence (x_n) in X with $(\varphi(x_n))$=descending, is d-convergent

(B2) (X,d,φ) is complete: each d-Cauchy sequence in X is d-convergent

(III-b) Lower semicontinuity properties of the objective function

(V1) φ is descending (X,d)-lsc:
$\lim_n\varphi(x_n)\geq\varphi(x)$, for each sequence (x_n) in X and each $x\in X$ with
$x_n\overset{d}{\longrightarrow}x$ and $(\varphi(x_n))$=descending

(V2) φ is (X,d)-lsc: $\liminf_n\varphi(x_n)\geq\varphi(x)$,
for each sequence (x_n) in X and each $x\in X$ with $x_n\overset{d}{\longrightarrow}x$.

This yields the "factor" Ekeland variational principles (EVP;Pj;Lm;Bh,Vq), where $j\in\{1,2,3,4,5,6\}$, $m,h,q\in\{1,2\}$; with the properties

(am-1) (EVP;Pj;Lm;gdc) $\Longrightarrow$ (EVP;Pj;Lm;Bh,Vq), for all admissible (j,m,h,q)

(am-2) (EVP;P(j);Lm;Bh,Vq) $\Longleftrightarrow$ (EVP;P(j+1);Lm;Bh,Vq), for all $j\in\{1,3,5\}$, $m,h,q\in\{1,2\}$

(am-3) the propositional map $(j,m,h,q) \mapsto (EVP;Pj;Lm;Bh,Vq)$ is decreasing: $(j,m,h,q) \le (j',m',h',q')$ gives (EVP;Pj;Lm;Bh,Vq) $\Longrightarrow$ (EVP;Pj';Lm';Bh',Vq').

Now, to get all these, it will suffice proving that the "weakest" variational principle (EVP;P1;L1;gdc) is deductible in (ZF-AC+DC). This follows from

Proposition 9. *We have [in (ZF-AC)] (DC) $\Longrightarrow$ (BB) $\Longrightarrow$ (EVP;P1;L1;gdc); hence, all principles (EVP;Pj;Lm;gdc), (EVP;Pj;Lm;Bh,Vq), where $j \in \{1,2,3,4,5,6\}$, $m,h,q \in \{1,2\}$, are deductible in (ZF-AC+DC).*

Proof. Let the triple (X,d,φ) be as in the premises of (EVP;P1;L1;gdc). Denote for simplicity $(\preceq) := (\preceq_{(d,\varphi)})$; hence

$$(x,y \in X)\text{: } x \preceq y \text{ iff } d(x,y) + \varphi(y) \le \varphi(x).$$

Remember that $(\preceq)$ is an order on $\mathrm{Dom}(\varphi)$; hence, all the more, on its subset $X_u := X(u,\preceq)$. We claim that (BB) applies to $(X_u,\preceq)$ and φ (restricted to X_u). In fact, let $(x_n; n \ge 0)$ be a $(\preceq)$-ascending sequence in X_u:

$$\text{(d03)}\quad d(x_n,x_m) \le \varphi(x_n) - \varphi(x_m), \text{ if } n \le m.$$

The sequence $(\varphi(x_n))$ is descending and bounded from below; hence, a Cauchy one. This, along with the working hypothesis, tells us that (x_n) is a d-Cauchy sequence in X_u. Putting these facts together, it results via (gdc) that there must be some $y \in X$ with $x_n \xrightarrow{d} y$, and $[\varphi(x_n) \ge \varphi(y), \forall n]$. Passing to limit as $m \to \infty$ in the working hypothesis, one gets $d(x_n,y) \le \varphi(x_n) - \varphi(y)$, $\forall n$; or, equivalently, $x_n \preceq y$, $\forall n$; whence, $y \in X_u$; so that, $(X_u,\preceq)$ is sequentially inductive. On the other hand, φ (restricted to X_u) is $(\preceq)$-decreasing; and this proves our claim. From (BB) it then follows that, for the starting $u \in X_u$ there exists $v \in X_u$ with

(j) $u \preceq v$; **(jj)** $v \preceq w \in X_u$ implies $\varphi(v) = \varphi(w)$.

The former of these is just our first conclusion in the statement. And the latter one gives our second conclusion of the same. In fact, let $x \in X$ be such that $d(v,x) \le \varphi(v) - \varphi(x)$ (hence, $v \preceq x$). As a consequence, $x \in X_u$; so that (by the conclusion **(jj)** above) $\varphi(v) = \varphi(x)$. Combining with the previous metrical relation gives $d(v,x) = 0$; whence $v = x$ (as d is sufficient); and we are done.

Remark 2. We stress that, by the very proof of the "composed" result above, one has, in (ZF-AC),

(am-4) (EVP;P5;L2;gdc) $\Longrightarrow$ (EVP;P1;L1;gdc);

hence, (EVP;P5;L2;gdc) $\Longleftrightarrow$ (EVP;P1;L1;gdc).

In fact, if (X,d,φ) fulfills conditions of (EVP;P1;L1;gdc), then (X_u,d,φ) (where $u \in \mathrm{Dom}(\varphi)$) fulfills conditions of (EVP;P5;L2;gdc); and, from the conclusion of this variational principle, we are done. However, as the statement above shows, the deduction of these principles requires the system (ZF-AC+DC). Similar properties are valid for the families of "factor" variational principles

((EVP;Pj;Lm;B1,V1); $j \in \{1,2,3,4,5,6\}$, $m \in \{1,2\}$),
((EVP;Pj;Lm;B2,V2); $j \in \{1,2,3,4,5,6\}$, $m \in \{1,2\}$);

we do not give details.

Note that (EVP;P1;L1;B2,V2) is just the 1974 Ekeland's variational principle [19]; denoted as (EVP); its complete formulation is as follows:

Theorem 6. *Assume that $\varphi : X \to R \cup \{\infty\}$ is proper, lower semicontinuous, and (X,d) is complete. Then, for each $u \in Dom(\varphi)$, there exists $v \in Dom(\varphi)$, with*
(b) *$d(u,v) \leq \varphi(u) - \varphi(v)$ (hence $\varphi(u) \geq \varphi(v) \geq \inf \varphi(X)$)*
(bb) *$[x \in X,\ d(v,x) \leq \varphi(v) - \varphi(x)] \Longrightarrow v = x$.*

The "strongest" principle in this series (EVP;P5;L2;B2,V2) is called the *bounded finitary* variant of (EVP) and is denoted as: (EVP-bf). Further technical aspects may be found in Bao and Khanh [4]; see also Hamel [26, Chap. 4] and Turinici [46, 49].

1.5 (EVP) $\Longrightarrow$ (DC)

The variational statements we just exposed found some useful applications to control and optimization, generalized differential calculus, critical point theory and global analysis; we refer to the 1979 paper by Ekeland [20] for a survey of these. As a consequence, many extensions of such principles were proposed; for a consistent list of these, we refer to the 1997 monograph by Hyers et al. [30, Chap. 5], and the 2003 monograph by Goepfert et al. [23, Chap. 3]. Note that, for each variational principle of this type (VP, say) one has (DC) $\Longrightarrow$ (VP) $\Longrightarrow$ (EVP); so, it is legitimate asking of to what extent are these logical inclusions effective. At a first glance, a negative answer is highly expectable; because, (DC) is "too general" with respect to (EVP). However, the situation is exactly opposite, in the sense: (EVP) includes (DC); and then, we closed the circle between all such principles. An early result of this type was provided in 1987 by Brunner [12]; for a different answer to the same, we refer to the 1999 paper by Dodu and Morillon [18]. It is our aim in the following to show that a further extension of this last result is possible, in the sense: (DC) is deductible from a certain Lipschitz bounded countable version of (EVP). Putting these together, it then results that any such variational principle (VP) is equivalent with both (DC) and (EVP).

Let $(X,\leq)$ be a partially ordered structure. Remember that $z \in X$ is $(\leq)$-*maximal* if $z \leq w \in X$ implies $z = w$; the class of all these will be denoted as $\max(X,\leq)$. In this case, we say that $(\leq)$ is a *Zorn order* when $\max(X,\leq)$ is (nonempty and) cofinal in X; i.e.: for each $u \in X$, there exists $v \in \max(X,\leq)$ with $u \leq v$. In particular, when $d(.,.)$ is a metric on X and $\varphi : X \to R_+$ is some function, a good example of partial order on X is that introduced by the convention

$$x \leq_{(d,\varphi)} y \text{ iff } d(x,y) \leq \varphi(x) - \varphi(y);$$

referred to as the *Brøndsted order [11]* attached to the couple (d,φ). Further, let us say that φ is d-*Lipschitz*, provided $|\varphi(x) - \varphi(y)| \leq Ld(x,y)$, $\forall x,y \in X$, for some $L > 0$; note that, any such function is uniformly continuous on X.

The following stronger variant of (EVP) enters in our discussion.

Theorem 7. *Let the metric space (X,d) and the function $\varphi : X \to R_+$ satisfy*

(e01) (X,d) *is bounded and complete*
(e02) φ *is d-Lipschitz (hence, bounded)*
(e03) $\varphi(X)$ *is (at most) countable.*

Then, $(\leq_{(d,\varphi)})$ *is a Zorn order.*

We call this the Lipschitz bounded countable version of EVP (in short: (EVP-Lbc)). By the above developments, (DC) $\Longrightarrow$ (EVP-bf) $\Longrightarrow$ (EVP-Lbc). The remarkable fact to be added is that this last result implies (DC).

Proposition 10. *We have [in (ZF-AC)]: (EVP-Lbc)* $\Longrightarrow$ *(DC). Hence, (DC) and (EVP-Lbc) are equivalent to each other in (ZF-AC); so that, all ordering/variational principles above are equivalent with both (DC) and (EVP-Lbc).*

Proof. The argument will be divided in several steps.

Part 1. Let M be a nonempty set; and $\mathcal{R}$ be a proper relation over M. Fix $a \in M$; and take some other point α, that does not belong to M. Put $P = M \cup \{\alpha\}$; and let $d(.,.)$ stand for the discrete metric on P:

$$d(s,t) = 0, \text{ if } s = t;\ \ d(s,t) = 1, \text{ if } s \neq t.$$

Remember that, $\mathcal{S}(P)$ is the class of all sequences $x = (x(n); n \geq 0)$ with elements in P. Denote $X = \{x \in \mathcal{S}(P); x(0) = a\}$; and let us introduce the metric

$$d_\infty(x,y) = \sum_n 2^{-n} d(x(n),y(n)),\ \text{ for } x = (x(n)) \text{ and } y = (y(n)) \text{ in } X.$$

Clearly, (X,d_∞) is bounded. Moreover, by a previous auxiliary statement, (X,d_∞) is complete: each d_∞-Cauchy sequence in X is d_∞-convergent.

Part 2. Let Y stand for the class of all sequences $x = (x(n); n \geq 0)$ in X with

$$(\forall n):\ x(n), x(n+1) \in M \Longrightarrow x(n) \mathcal{R} x(n+1).$$

Note that $Y \neq \emptyset$; for, given $b \in \mathcal{R}(a)$, the sequence $y = (y(n); n \geq 0)$ in X introduced as $(y(0) = a,\ y(1) \in b,\ y(n) = \alpha,\ n \geq 2)$ is an element of it.

Lemma 4. *The subset Y is* d_∞*-closed; hence,* d_∞*-complete as well.*

Proof. Let $(x^n := (x^n(0) = a, x^n(1), \ldots); n \geq 0)$ be a sequence in Y, and $y = (y(n); n \geq 0)$ be an element of X with $x^n \to y$ (modulo d_∞); that is,

$$d_\infty(x^n,y) := \sum_i 2^{-i} d(x^n(i), y(i)) \to 0,\ \text{as } n \to \infty.$$

Note that, as a direct consequence of this,

$$x^n(i) \xrightarrow{d} y(i) \text{ as } n \to \infty,\ \forall i \geq 0. \tag{10}$$

Further, as d_∞ is metric, $(x^n; n \geq 0)$ is d_∞-Cauchy; so, by a preliminary statement, there exists a strictly ascending sequence of ranks $(n(i); i \geq 0)$, with

$$(\forall i \geq 0):\ \ x^{n(i)}(i) = x^p(i),\ \ \forall p \geq n(i). \tag{11}$$

In this case, the d_∞-limit, $y=(y(n);n\geq 0)$, of our sequence must have the form

$$y(i)=x^{n(i)}(i),\ \text{for all } i\geq 0. \tag{12}$$

We now claim that the representation of Y gives us the desired conclusion: $y\in Y$. In fact, let $i\geq 0$ be such that $y(i),y(i+1)\in M$. Note that, by the previous relations,

$$y(i)=x^{n(i)}(i)=x^{n(i+1)}(i)\in M;\ \ y(i+1)=x^{n(i+1)}(i+1)\in M.$$

This, along with $x^{n+1}\in Y$, yields

$$x^{n(i+1)}(i)\mathcal{R}x^{n(i+1)}(i+1);\ \text{that is, } y(i)\mathcal{R}y(i+1).$$

The argument is thereby complete.

Part 3. Now, let us note that, conclusion of our statement is equivalent with $Y\cap \mathcal{S}(M)\neq\emptyset$. For, taking some sequence $y=(y(n);n\geq 0)$ in this intersection, we have $y(n),y(n+1)\in M$, $\forall n$; so that, by definition, $y(n)\mathcal{R}y(n+1)$, $\forall n$; whence, $(y(n);n\geq 0)$ is $(a,\mathcal{R})$-iterative. Assume by contradiction that this would be not true:

(e04) $Y\cap\mathcal{S}(M)=\emptyset$; i.e.: $\forall y=(y(n);n\geq 0)\in Y$, $\exists k=k(y)\geq 1:\ y(k)=\alpha$.

As a consequence, the functions below are well defined:

$$g(y)=\min\{k\geq 1;y(k)=\alpha\},\ \varphi(y)=2^{2-g(y)},\ \ y\in Y.$$

Some basic properties of these are described in

Lemma 5. *The following are valid:*

(i) *the functions g, φ are continuous on Y; precisely,*

$$\forall y\in Y,\exists\beta=\beta(y)>0:\ z\in Y,d_\infty(z,y)<\beta\Longrightarrow g(z)=g(y),\ \varphi(z)=\varphi(y) \tag{13}$$

(ii) *the function φ is d_∞-Lipschitz, in the sense:*

$$|\varphi(x)-\varphi(y)|\leq 4d_\infty(x,y),\ \ \forall x,y\in Y \tag{14}$$

(iii) *$g(Y)$ is countable; hence, so is $\varphi(Y)$.*

Proof. (i) Fix $y=(y(n);n\geq 0)\in Y$, and put $r=g(y)$; we therefore have

$$r\geq 1,y(r)=\alpha,y(k)\in M,\forall k\in N(r,>).$$

Take some $\beta\in]0,2^{-r}[$; and let $z=(z(n);n\geq 0)\in Y$ be such that $d_\infty(y,z)<\beta$. By the definition of our metric,

$$2^{-k}d(y(k),z(k))<\beta<2^{-r},\ \ \forall k\in N(r,\geq);$$

and this yields $[z(k)=y(k),\ \forall k\in N(r,\geq)]$. In particular, we must have

$$z(k)\in M,\ \ \forall k\in N(r,>);\ z(r)=\alpha;$$

so that $g(z)=\alpha=g(y)$.

(ii) Let $x=(x(n);n\geq 0)$ and $y=(y(n);n\geq 0)$ be two points in Y. Denote, for simplicity $r=g(x)$, $s=g(y)$. If $r=s$, all is clear; so, it remains the opposite case $r\neq s$; without loss, one may assume that $r<s$. As a consequence,

$$\begin{array}{l} x=(x(0),\ldots,x(r-1),\alpha,\ldots,x(s-1),x(s),\ldots),\\ y=(y(0),\ldots,y(r-1),y(r),\ldots y(s-1),\alpha,\ldots). \end{array}$$

In particular, $y(r)\in M$; hence $y(r)\neq\alpha$; and then,

$$d_\infty(x,y)\geq 2^{-r}\geq 2^{-r}-2^{-s}=|2^{-r}-2^{-s}|.$$

This gives the conclusion we need.

(iii) Evident.

Part 4. We show that, under the introduced conventions,

$$\text{for each } v\in Y \text{ there exists } y\in Y\setminus\{v\} \text{ such that } d_\infty(v,y)\leq\varphi(v)-\varphi(y); \qquad (15)$$

or, in other words: each element of Y is non-maximal with respect to the Brøndsted ordering attached to d_∞ and φ:

$(x_1,x_2\in Y)$: $x_1\leq x_2$ iff $d_\infty(x_1,x_2)\leq\varphi(x_1)-\varphi(x_2)$.

In fact, let $v=(v(n);n\geq 0)$ be the representation of this $v\in Y$. Put $g(v)=r$; hence

$$r\geq 1;\ v(0),\ldots,v(r-1)\in M;\ v(r)=\alpha.$$

Note that, by the definition of Y, one gets the relations

$$v(i)\mathcal{R}v(i+1),\ \text{ whenever } (r\geq 2 \text{ and}) \ i\leq r-2.$$

Take $y=(y(n);n\geq 0)$ in $Y\setminus\{v\}$ according to

(e05) $y(k)=v(k),\ \forall k\in N(r,>);\ y(h)=\alpha,\ \forall h\in N(r+1,<)$;
(e06) $y(r),y(r+1)\in M;\ y(i)\mathcal{R}y(i+1),\ \forall i\in\{r-1,r\}$.

(The last relation is possible, by the Finite Dependent Choice property). As a consequence of this, $g(y)=r+2$. Now, the desired relation above becomes

$$d_\infty(v,y)\leq 2^{2-r}-2^{-r}=3.2^{-r}.$$

According to the representation of $y\in Y\setminus\{v\}$, this means

$$\sum_{i\geq r}2^{-i}d(v(i),y(i))\leq 3.2^{-r}. \qquad (16)$$

But then, the last relation is clear, in view of

$$\sum_{i\geq r} 2^{-i} d(v(i),y(i)) \leq \sum_{i\geq r} 2^{-i} = 2^{1-r} < 3.2^{-r}.$$

Part 5. We may now pass to the final part of the argument. By the above facts, (EVP-Lbc) is applicable to the metric space (Y,d_∞) and the function $\varphi : Y \to R_+$ (introduced as before). Hence, the associated Brøndsted order $(\leq)$ (see above) is a Zorn one. As a consequence, there exists, for the starting point (in Y)

$$u = (u(n); n \geq 0):\ u(0) = a,\ u(n) = \alpha,\ \forall n \geq 1,$$

some other point $v = (v(n); n \geq 0)$ in Y with

$$u \leq v :\ d_\infty(u,v) \leq \varphi(u) - \varphi(v) \tag{17}$$

$$v \text{ is } (\leq)\text{-maximal} :\ d_\infty(v,y) > \varphi(v) - \varphi(y),\ \ \forall y \in Y \setminus \{v\}. \tag{18}$$

This, however, contradicts the preceding step and shows that $Y \cap \mathcal{S}(M) \neq \emptyset$. But then (by the very definition of Y) there must be some sequence $y = (y(n); n \geq 0)$ in M with $y(0) = a$ and $y(n) \mathcal{R} y(n+1)$, $\forall n$. The proof is complete.

In particular, when the boundedness and Lipschitz properties are ignored, this result is just the one in Dodu and Morillon [18]. So, it is natural to call it as: Dodu–Morillon statement; in short: (DM).

Let X be a nonempty set; and $(\leq)$ be an order on it. We say that $(\leq)$ has the *inf-lattice* property, provided: $x \wedge y := \inf(x,y)$ exists, for all $x,y \in X$. Further, let $d : X \times X \to R_+$ be a metric over X; and $\varphi : X \to R_+$ be some function. 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 an assumption, any function $\psi : X \to R$ is continuous over X. However, the *Lipschitz property* ($|\psi(x) - \psi(y)| \leq L d(x,y)$, for all $x,y \in X$, and some $L > 0$) cannot be assured, in general.

Now, the result below is a particular case of (EVP):

Theorem 8. *Let the metric space (X,d) and the function $\varphi : X \to R_+$ satisfy*

(e07) (X,d) is discrete bounded and complete
(e08) $(\leq_{(d,\varphi)})$ has the inf-lattice property
(e09) φ is d-nonexpansive and $\varphi(X)$ is countable.

Then, $(\leq_{(d,\varphi)})$ is a Zorn order.

We shall refer to it as: the discrete Lipschitz countable version of EVP (in short: (EVP-dLc)). Clearly, (EVP) $\Longrightarrow$ (EVP-dLc). The remarkable fact to be added is that this last principle yields (DC); so, it completes the circle between all these.

Proposition 11. *We have [in (ZF-AC)]: (EVP-dLc) $\Longrightarrow$ (DC). Hence, (EVP), (DC) and (EVP-dLc) are equivalent to each other in (ZF-AC).*

For a complete proof, see Turinici [51]. In particular, when the discrete, bounded, inf-lattice and nonexpansive properties are ignored in (EVP-dLc), the last result above reduces to the one in Brunner [12]; so, it is natural to name it as: Brunner statement; in short: (Bru).

Finally, note that, by the conclusions of (DM) and (Bru), we have (EVP-Lbc) $\Longleftrightarrow$ (EVP-dLc); moreover, both these statements are equivalent with (DC) and/or (EVP). It would be interesting to give a direct proof of this equivalence; in fact, of (EVP-Lbc) $\Longrightarrow$ (EVP-dLc). Further aspects may be found in Schechter [41, Chap. 19, Sect. 19.51].

2 Variational Principles in Fang Spaces

2.1 Conical Gauge Functions

Let Y be a (real) *vector space*. Take a *convex cone* H of Y ($\alpha H+\beta H\subseteq H$, for each α,β in R_+); which in addition is *non-degenerate* ($H\neq\{0\}$), *proper* ($H\neq Y$); and let $(\leq)$ stand for its induced quasi-order [$x\leq y$ iff $y-x\in H$]. Further, take some point $k^0\in H\setminus(-H)$; and put (for $y\in Y$)

(a01) $\Gamma(H;k^0;y)=\{s\in R_+;k^0s\leq y\}$, $\gamma(H;k^0;y)=\sup\Gamma(H;k^0;y)$.

(Here, by convention, $\sup(\emptyset)=-\infty$). We therefore defined a multivalued function $\Gamma(.):=\Gamma(H;k^0;.)$ from Y to $\mathcal{P}(R_+)$, and a function $\gamma(.):=\gamma(H;k^0;.)$ from Y to $R_+\cup\{-\infty,\infty\}$ with

$$\Gamma(y)\neq\emptyset \text{ (hence, } 0\leq\gamma(y)\leq\infty\text{), iff } y\in H; \tag{19}$$

the latter of these will be referred to as the *gauge* function attached to $(H;k^0)$. Note that, for each $y\in H$,

$$\Gamma(y) \text{ is hereditary } (s\in\Gamma(y)\Longrightarrow[0,s]\subseteq\Gamma(y)); \tag{20}$$

so, $\Gamma(y)$ is an interval of R_+, having $0\in R_+$ as left endpoint. In addition, the couple (Γ,γ) is positively homogeneous and increasing

$$\Gamma(ty)=t\Gamma(y),\gamma(ty)=t\gamma(y),\ \ \forall t>0,\ \forall y\in Y \tag{21}$$

$$y_1,y_2\in Y,\ y_1\leq y_2 \text{ implies } \Gamma(y_1)\subseteq\Gamma(y_2),\ \gamma(y_1)\leq\gamma(y_2). \tag{22}$$

(Here, $t\emptyset=\emptyset,\forall t\in R_+^0$). An important question to be solved is that of Γ being *proper* [$\Gamma(y)\neq R_+,\forall y\in H$]. According to Cristescu [17, Chap. 5, Sect. 1], we say that H is *Archimedean*, provided

(a02) $[v\in Y,h\in H,\Gamma(H;v;h)=R_+]$ imply $v\in -H$.

Likewise, let us call H, *semi-Archimedean*, if

(a03) $\Gamma(H;k;y)$ is closed, $\forall k\in H\setminus(-H),\forall y\in H$.

Lemma 6. *The following are valid:*

(i) *If H is Archimedean, then $\Gamma(.)$ is proper, in the sense: $0 \leq \gamma(y) < \infty$ and $\Gamma(y) = [0, \gamma(y)]$, for all $y \in H$; so, H is semi-Archimedean too*

(ii) *Let H be semi-Archimedean; and $\alpha \in R_+$, $y \in H$, $(\beta_n; n \geq 0) \subseteq R_+$ be such that $[k^0\alpha \leq y + k^0\beta_n, \forall n]$ and $\beta_n \to 0$. Then, $k^0\alpha \leq y$.*

Proof. **(i)** Let $y \in H$ be arbitrary fixed. If $(\Gamma(H; k^0; y) =)\Gamma(y) = R_+$ then, by the Archimedean property of H, one gets $k^0 \in -H$; contradiction. Consequently, $R_+ \setminus \Gamma(y) \neq \emptyset$; so that, by the hereditary property, $\Gamma(y)$ is bounded [whence, $0 \leq \gamma(y) < \infty$]. Further, again by this property, $k^0\gamma(y) - y \leq k^0 t$, for all $t > 0$; wherefrom $\Gamma(H; k^0\gamma(y) - y; k^0) = R_+$. This, again combined with the Archimedean property of H, gives $\gamma(y) \in \Gamma(y)$; i.e., $\Gamma(y) = [0, \gamma(y)]$.

(ii) If $\alpha = 0$ or $[\beta_n = 0$, for some $n \geq 0]$, we are done; so, without loss, one may assume that $\alpha > 0$ and $\beta_n > 0, \forall n$. As $\beta_n \to 0 < \alpha$, there must be some $n(\alpha) \geq 0$ in such a way that $0 < \alpha - \beta_n < \alpha$, $\forall n \geq n(\alpha)$. The imposed hypothesis now gives: $\alpha - \beta_n \in \Gamma(y)$, $\forall n \geq n(\alpha)$. Passing to limit as $n \to \infty$ yields (by the semi-Archimedean property of H), $\alpha \in \Gamma(y)$; and the assertion follows.

The following couple of properties will be useful in the sequel:

Lemma 7. *The gauge function γ is super-additive and subtractive:*

$$\gamma(y_1 + y_2) \geq \gamma(y_1) + \gamma(y_2), \text{ if the right member exists} \tag{23}$$

$$\gamma(y_1 - y_2) \leq \gamma(y_1) - \gamma(y_2), \text{ if the right member exists.} \tag{24}$$

Proof. Without loss, one may assume that $y_1, y_2 \in H$ and $\gamma(y_1) > 0$, $\gamma(y_2) > 0$. By the hereditary property, $y_1 \geq k^0 t_1, y_2 \geq k^0 t_2$, whenever $0 \leq t_1 < \gamma(y_1), 0 \leq t_2 < \gamma(y_2)$; and this yields (for all such (t_1, t_2)) $y_1 + y_2 \geq k^0[t_1 + t_2]$ (i.e.: $\gamma(y_1 + y_2) \geq t_1 + t_2$). This, and the arbitrariness of the precise couple, ends the argument. The second part is directly obtainable from the first one, in a standard way.

In particular, when Y is locally convex, the Archimedean property of H is holding whenever H is closed. Then, our developments reduce to the ones in Goepfert et al. [22]. Note that an axiomatic approach of these facts is possible, under the lines in Artzner et al. [2]; we do not give details.

2.2 Zhu–Li Vectorial Principles

Let in the following Y stand for a (real) vector space.

(A) Take a (non-degenerate, proper) Archimedean convex cone H of Y; and let $(\leq_H)$ stand for its induced quasi-order. Further, let K be some (non-degenerate, proper) semi-Archimedean convex cone of Y, with $K \subseteq H$; and denote by $(\leq_K)$ the induced quasi-order.

(B) Further, let X be a nonempty set. By a *pseudometric* over X we mean any map $d : X \times X \to R_+$; if, in addition, d is *reflexive* $[d(x,x) = 0, \forall x \in X]$ and *symmetric* $[d(x,y) = d(y,x), \forall x,y \in X]$, we say that it is a *rs-pseudometric*. Let $(\Lambda, \leq)$ be some directed quasi-ordered structure. Take a family $D = (d_\lambda; \lambda \in \Lambda)$ of rs-pseudometrics over X, with the properties: Λ-*sufficient* $[d_\lambda(x,y) = 0, \forall \lambda \in \Lambda \Longrightarrow x = y]$, Λ-*monotone* $[\lambda \leq \mu$ implies $d_\lambda(.,.) \leq d_\mu(.,.)]$ and Λ-*triangular* $[\forall \lambda \in \Lambda, \exists \mu \in \Lambda(\lambda, \leq)$, with $d_\lambda(x,z) \leq d_\mu(x,y) + d_\mu(y,z), \forall x,y,z \in X]$. By definition, D will be referred to as a *Fang metric*; and (X,D), as a *Fang space*.

The introduced Fang metric D may now generate a conv-Cauchy structure, in the following way. Take an arbitrary sequence $(x_n; n \geq 0)$ in X. Given $\lambda \in \Lambda$, the d_λ-*convergence* of this sequence towards an $x \in X$ [depicted as: $x_n \xrightarrow{d_\lambda} x$], means: $d_\lambda(x_p, x) \to 0$ as $p \to \infty$; [i.e.: $\forall \varepsilon > 0, \exists i = i(\varepsilon)$, such that $i \leq p \Longrightarrow d_\lambda(x_p, x) < \varepsilon$]. If this holds for all $\lambda \in \Lambda$, then $(x_n; n \geq 0)$ is said to D-*converge* towards x [written as: $x_n \xrightarrow{D} x$]; moreover, if $x \in X$ is generic in such a convention, $(x_n; n \geq 0)$ is called D-*convergent*. On the other hand, given $\lambda \in \Lambda$, the d_λ-Cauchy property of $(x_n; n \geq 0)$ means: $d_\lambda(x_p, x_q) \to 0$ as $p,q \to \infty$, $p \leq q$ [i.e.: $\forall \varepsilon > 0, \exists j := j(\lambda, \varepsilon)$, such that $j \leq p \leq q \Longrightarrow d_\lambda(x_p, x_q) < \varepsilon$]. If this holds for each $\lambda \in \Lambda$, we say that $(x_n; n \geq 0)$ is D-*Cauchy*. Note that any D-convergent sequence is D-Cauchy too; the reciprocal is not in general valid.

(C) Now, let (Y, H, K) be as above and (X,D) be a Fang space. Let us complete Y with an element $\infty \notin Y$; the algebraic/order conventions introduced over the completion $Y \cup \{\infty\}$ are

$$\infty = b + \infty = \infty + b, \ \forall b \in Y \cup \{\infty\}; \ \infty = \lambda\infty, \ \forall \lambda \in R_+^0$$
$$b \leq_H \infty, \ b \leq_K \infty, \ \forall b \in Y \cup \{\infty\}; \ \neg(\infty \leq_H b), \ \neg(\infty \leq_K b), \ \forall b \in Y.$$

Take a couple of functions $F : X \to Y \cup \{\infty\}$ and $k : \Lambda \to K$, with

(b01) F is proper: $\mathrm{Dom}(F) := \{x \in X; F(x) \neq \infty\} \neq \emptyset$;
(b02) k is K-increasing $[\lambda \leq \mu \Longrightarrow k(\lambda) \leq_K k(\mu)]$.

The relation $(\preceq_{(D,k,F)})$ over X introduced as

(b03) $(x_1, x_2 \in X)$: $x_1 \preceq_{(D,k,F)} x_2$ iff $k(\lambda) d_\lambda(x_1, x_2) + F(x_2) \leq_K F(x_1), \forall \lambda \in \Lambda$

is a quasi-order. For a number of both practical and theoretical reasons, it would be useful to determine sufficient conditions under which $(\preceq_{(D,k,F)})$ is a Zorn order with respect to $\mathrm{Dom}(F)$. To reach this objective, some regularity conditions about our data must be imposed. Precisely, these consist in

(I) Boundedness properties of the (vectorial) objective function:

(Q1) F is H-bounded below: $F(x) \geq_H b, \forall x \in X$, for some $b \in Y$
(Q2) F is H-positive: $F(x) \geq_H 0, \forall x \in X$
(Q3) $F(X) \subseteq Y$ and F is H-bounded below: $F(x) \geq_H b, \forall x \in X$, for some $b \in Y$
(Q4) $F(X) \subseteq Y$ and F is H-positive: $F(x) \geq_H 0, \forall x \in X$
(Q5) $F(X) \subseteq Y$ and F is H-bounded: $a \geq_H F(x) \geq_H b, \forall x \in X$, for some $a, b \in Y$
(Q6) $F(X) \subseteq Y$ and F is H-bounded positive:
$a \geq_H F(x) \geq_H 0, \forall x \in X$, for some $a \in Y$

(II) Strict positivity properties of the coefficient function

(M1) $k(\Lambda) \subseteq K \setminus (-H)$ (hence, $k(\Lambda) \subseteq K \setminus (-K)$)
(M2) $k(\Lambda) \subseteq K \setminus (-H)$ and k is constant
($k(\lambda) = k^0, \forall \lambda \in \Lambda$, for some $k^0 \in K \setminus (-H)$)

(III) Global completeness property of the whole objects family:

(gsdc) (X, D, K, F) is globally sequentially descending complete: each D-Cauchy sequence (x_n) in X for which $(F(x_n))$ is K-descending, D-converges to some $x \in X$; with, in addition, $[F(x_n) \geq_K F(x), \forall n]$.

Our main result is

Theorem 9. *Let the proper function $F : X \to Y \cup \{\infty\}$ and the K-increasing coefficient function $k : \Lambda \to K$ be such that F is in the subclass (Qj) for some $j \in \{1,2,3,4,5,6\}$, $k(.)$ has the property (Mp), for some $p \in \{1,2\}$, and (X, D, K, F) has the property (gsdc). Then, for each $u \in Dom(F)$, there exists $v \in Dom(F)$, with*
(a) *$k(\lambda) d_\lambda(u, v) \leq_K F(u) - F(v)$, $\forall \lambda \in \Lambda$ (hence, $F(u) \geq_K F(v)$)*
(aa) *$x \in X$, $[k(\lambda) d_\lambda(v, x) \leq_K F(v) - F(x), \forall \lambda \in \Lambda] \Longrightarrow v = x$.*

Denote the obtained statements as (ZL;Qj;Mp;gsdc), where $j \in \{1,2,3,4,5,6\}$, $p \in \{1,2\}$; these will be referred to as: "composed" Zhu–Li variational principles. The relationships between them are given by

Lemma 8. *We have [in (ZF-AC)],*
(i) *(ZL;Q(j);Mp;gsdc) $\Longleftrightarrow$ (ZL;Q(j+1);Mp;gsdc), $\forall j \in \{1,3,5\}$, $\forall p \in \{1,2\}$*
(ii) *the propositional map $(j, p) \mapsto (ZL; Qj; Mp; gsdc)$ is decreasing:*
if $(j, p) \leq (j', p')$, then (ZL;Pj;Mp;gsdc) $\Longrightarrow$ (ZL;Pj';Mp';gsdc).

Proof. (i): Let (X, D, K, F) be as in the premise of the principle (ZL;Q(j);Mp;gsdc). Then, the quadruple (X, D, K, G), where $G(.) := F(.) - b$, fulfills conditions of the principle (ZL;Q(j+1);Mp;gsdc); and, from this, we are done.
(ii) Evident, by the definition of the subclasses in question.

Now, the property (gsdc) is obtainable as an intersection of two components:
(III-a) Completeness properties involving the Fang structure:

(C1) (X, D, K, F) is sequentially descending complete: each D-Cauchy sequence (x_n) in X with $(F(x_n))$=K-descending is D-convergent
(C2) (X, D, K, F) is sequentially complete:
each D-Cauchy sequence in X is D-convergent.

(III-b) Lower semi-continuity properties of the objective function:

(W) F is sequentially K-descending (X, D)-lsc:
for each sequence (x_n) in X and each $x \in X$ with
$x_n \xrightarrow{D} x$ and $(F(x_n))$=K-descending, we have $[F(x_n) \geq_K F(x), \forall n]$.

This gives the "factor" Zhu–Li variational principles (ZL;Qj;Mp;Ch,W), where $j \in \{1,2,3,4,5,6\}$, $p,h \in \{1,2\}$; with the properties

(v-1) (ZL;Qj;Mp;gsdc) $\Longrightarrow$ (ZL;Qj;Mp;Ch,W), for all admissible (j,p,h)

(v-2) (ZL;Q(j);Mp;Ah,Ch,W) $\Longleftrightarrow$ (ZL;Q(j+1);Mp;Ch,W),
for all $j \in \{1,3,5\}$, $p,h \in \{1,2\}$

(v-3) the propositional map $(j,p,h) \mapsto (ZL;Qj;Mp;Ch,W)$ is decreasing:
if $(j,p,h) \leq (j',p',h')$, then (EVP;Qj;Mp;Ch,W) $\Longrightarrow$ (EVP;Qj';Mp';Ch',W).

Now, to get all these, it will suffice proving that the "weakest" variational principle (ZL;Q1;M1;gsdc)—or, equivalently (see above), (ZL;Q2;M1;gsdc)—is deductible in (ZF-AC+DC). This follows from

Proposition 12. *We have [in (ZF-AC)] ((DC) $\Longrightarrow$) (BB) $\Longrightarrow$ (ZL;Q2;M1;gsdc); hence, all these principles are deductible in (ZF-AC+DC).*

Proof. Let the triple (Y,H,K), the Fang space (X,D), the proper function $F : X \to Y \cup \{\infty\}$, and the K-increasing coefficient function $k : \Lambda \to K$ be as in the premises of (ZL;Q2;M1;gsdc). By (M1), the quasi-order $(\preceq_{(D,k,F)})$ is antisymmetric—hence an order—on $\mathrm{Dom}(F)$; we do not give details. Given $u \in \mathrm{Dom}(F)$, denote $X_u = X(u, \preceq_{(D,k,F)})$. Note that, by the very definition of this subset,

$$0 \leq_H F(x) \leq_K F(u) \text{ (hence, } 0 \leq_H F(x) \leq_H F(u)), \forall x \in X_u. \tag{25}$$

So, if we denote again by F the restriction of the initial function F to the subset X_u, one has $F(X_u) \subseteq H$: and, moreover (see a preceding observation),

$$X_u = \{x \in X; k(\lambda) d_\lambda(u,x) \leq_K F(u) - F(x),\ \forall \lambda \in \Lambda\}. \tag{26}$$

The argument will be divided in a number of steps.

Part 1. Fix $\theta \in \Lambda$ and put $\Theta = \Lambda(\theta, \leq)$; note that (as $(\Lambda, \leq)$ is directed),

$$\Theta \text{ is cofinal in } \Lambda\text{: for each } \lambda \in \Lambda \text{ there exists } \mu \in \Theta \text{ with } \lambda \leq \mu. \tag{27}$$

Let $\delta(.) := \gamma(H; k(\theta); .)$ stand for the gauge function attached to $(H; k(\theta))$. As H is Archimedean, $\delta(H) \subseteq R_+$. Put also $\delta(\infty) = \infty$; then, the function $[\psi(x) = \delta(F(x)),\ x \in X]$ is an element of $\mathcal{F}(X, R_+ \cup \{-\infty, \infty\})$. Let again ψ stand for the restriction to X_u of this function; by the relations above, one has $\psi(X_u) \subseteq R_+$.

Part 2. Let $(\sqsubseteq_{(D,\psi)})$ stand for the relation over X_u:

(b04) $(x_1, x_2 \in X_u)$: $x_1 \sqsubseteq_{(D,\psi)} x_2$ iff $d_\lambda(x_1,x_2) \leq \psi(x_1) - \psi(x_2), \forall \lambda \in \Lambda$;

it is an order on X_u, as it can be directly seen. We claim that the following double inclusion holds:

$$(\forall x_1, x_2 \in X_u):\ x_1 \preceq_{(D,k,F)} x_2 \Longrightarrow x_1 \sqsubseteq_{(D,\psi)} x_2 \Longrightarrow\ \psi(x_1) \geq \psi(x_2). \tag{28}$$

The second part is clear; so, it remains to verify the first part. Let $x_1, x_2 \in X_u$ be such that $x_1 \preceq_{(D,k,F)} x_2$; that is:

$$k(\lambda) d_\lambda(x_1,x_2) \leq_K F(x_1) - F(x_2),\ \forall \lambda \in \Lambda.$$

As $k(.)$ is K-increasing, this yields

$$k(\theta)d_\lambda(x_1,x_2) \leq_K F(x_1)-F(x_2),\ \forall \lambda \in \Theta;$$

so that (by the subtractive property of gauge function $\delta(.)$)

$$d_\lambda(x_1,x_2) \leq \delta(F(x_1)-F(x_2)) \leq \psi(x_1)-\psi(x_2),\ \forall \lambda \in \Theta.$$

This, in turn, yields (as $\lambda \mapsto d_\lambda(.,.)$ is increasing and Θ is cofinal in $(\Lambda,\leq)$)

$$d_\lambda(x_1,x_2) \leq \psi(x_1)-\psi(x_2),\ \forall \lambda \in \Lambda;$$

that is: $x_1 \sqsubseteq_{(D,\psi)} x_2$; hence, the assertion.

Part 3. We show that (BB) is applicable to $(X_u,\preceq_{(D,k,F)})$ and ψ (restricted to X_u). Firstly, by the double inclusion above, ψ is decreasing (modulo $(\preceq_{(D,k,F)})$). Secondly, let the sequence $(x_n;n\geq 0)$ in X_u be $(\preceq_{(D,k,F)})$-ascending:

(b05) $(\forall \lambda \in \Lambda)$: $k(\lambda)d_\lambda(x_n,x_m) \leq_K F(x_n)-F(x_m)$, if $n \leq m$.

note that, in such a case, $(F(x_n);n\geq 0)$ is K-descending. Again by the quoted double inclusion, $(x_n;n\geq 0)$ is $(\sqsubseteq_{(D,\psi)})$-ascending:

$$(\forall \lambda \in \Lambda):\ d_\lambda(x_n,x_m) \leq \psi(x_n)-\psi(x_m),\ \text{if } n \leq m; \tag{29}$$

and, from this, $(x_n;n\geq 0)$ is D-Cauchy. Combining with (gsdc) yields $x_n \xrightarrow{D} x$, for some (uniquely determined) $x \in X$; with, in addition, $F(x_n) \geq_K F(x)$, $\forall n$. Taking into account the working condition, we therefore get

$$(\forall \lambda \in \Lambda):\ k(\lambda)d_\lambda(x_n,x_m) \leq_K F(x_n)-F(x),\ \text{ if } n \leq m. \tag{30}$$

Fix $\lambda \in \Lambda$ and $n \geq 0$. Let $\mu \in \Lambda(\lambda,\leq)$ be the index assured by the Λ-triangular property of D. From the preceding relation, we have (as $k(.)$ is K-increasing)

$$\begin{array}{l} k(\lambda)d_\lambda(x_n,x) \leq_K k(\mu)d_\mu(x_n,x_m)+k(\lambda)d_\mu(x_m,x) \\ \leq_K F(x_n)-F(x)+k(\lambda)d_\mu(x_m,x),\ \text{for all } m \geq n; \end{array} \tag{31}$$

This, along with (M1) and the semi-Archimedean property of K, yields (by a preceding auxiliary fact)

$$(\forall \lambda \in \Lambda):\ k(\lambda)d_\lambda(x_n,x) \leq_K F(x_n)-F(x);\ \text{i.e.}:\ x_n \preceq_{(D,k,F)} x. \tag{32}$$

As $n \geq 0$ was arbitrarily fixed, we thus get that the limit point x is an upper bound of $(x_n;n\geq 0)$ [modulo $(\preceq_{(D,k,F)})$]. This gives $x \in X_u$; so that (by the arbitrariness of our sequence), $(X_u,\preceq_{(D,k,F)})$ is sequentially inductive; hence, the claim.

Part 4. Applying (BB) to these data, one gets that, for the starting $u \in X_u$, there exists a point $v \in X_u$, with

(j) $u \preceq_{(D,k,F)} v$; **(jj)** $v \preceq_{(D,k,F)} w \in X_u \Longrightarrow \psi(v)=\psi(w)$.

The former of these is just the first conclusion of the statement. Moreover, by the latter of these, one gets the second conclusion of the same. For, let $x \in X$ be such that $v \preceq_{(D,k,F)} x$. By **(j)**, this yields $u \preceq_{(D,k,F)} x$; whence, $x \in X_u$. This, combined with the double inclusion we already quoted, gives $v \sqsubseteq_{(D,\psi)} x$; and, by **jj)** above, $\psi(v) = \psi(x)$. Combining these, one gets $d_\lambda(v,x) = 0$, $\forall \lambda \in \Lambda$; wherefrom, $v = x$ (as D is sufficient); and the claim follows.

Remark 3. We stress that, by the very proof of the "composed" result above, one has, in (ZF-AC),

(v-4) $(\forall p \in \{1,2\})$: (ZL;Q5;Mp;gsdc) $\Longrightarrow$ (ZL;Q1;Mp;gsdc);
hence, (EVP;Q5;Mp;gsdc) $\Longleftrightarrow$ (EVP;Q1;Mp;gsdc).
In fact, if $(X,D,F,k(.))$ fulfills conditions of (ZL;Q1;Mq;gsdc), then $(X_u,D,F,k(.))$ (where $u \in \mathrm{Dom}(F)$) fulfills conditions of (ZL;Q5;Mq;gsdc); and, from the conclusion of this gauge variational principle, we are done. However, as the statement above shows, the deduction of these gauge principles cannot be reached in (ZF-AC); because it requires (BB) (or, equivalently, (DC)).

In particular, when Y is a locally convex space, the Archimedean property of H is assured when $H = \mathrm{cl}(K)$. The corresponding "factor" variational principle (ZL;Q3;M1;C2,W) is just the main result in Zhu and Li [53] proved via rather different methods; this also explains the conventions we just introduced. On the other hand, the "factor" variational principle (ZL;Q3;M2;C2,W) yields the main result in Turinici [48] which includes the ones in Goepfert et al. [22]. But, as precise by the quoted authors, their statements include (EVP); hence, summing up: ((DC) $\Longrightarrow$) (BB) $\Longrightarrow$ (ZL;Q1;M1;gsdc) $\Longrightarrow$ (ZL;Q1;M1;C1,W) $\Longrightarrow$ (ZL;Q5;M2;C2,W) $\Longrightarrow$ (EVP-bf). This, by the Dodu–Morillon statement, tells us that any of the variational principles (ZL;Qj,Mq;gsdc) and (ZL;Qj;Mq;Ch,W), where $j \in \{1,2,3,4,5,6\}$ and $q,h \in \{1,2\}$, is equivalent with any of the principles (DC), (BB), and/or (EVP).

Note, finally, that these equivalence properties are no longer valid beyond the Fang setting; some results in this direction may be found in Turinici [47].

2.3 Hamel Variational Principles

Let Y be a (real) vector space. By the properties of conical gauge functions,

$$H = K = \text{Archimedean (non-degenerate, proper) (convex) cone of } Y$$

is allowed in the Zhu–Li vector variational principles above. This, in the case of $Y = R$, $H = K = R_+$, yields a lot of "scalar" variational principles over Fang spaces, including Hamel's [27]. It is our aim in the following to state these principle as well as to discuss certain related facts.

(A) Let X be a nonempty set; and $(\Lambda, \leq)$ be a directed quasi-ordered structure. Take a family $D = (d_\lambda; \lambda \in \Lambda)$ of rs-pseudometrics over X; supposed to be Λ-sufficient, Λ-monotone, and Λ-triangular. By a previous convention, D will be

referred to as a *Fang metric*; and (X,D), as a *Fang space*. Further, let $\varphi : X \to R \cup \{\infty\}$ and $k : \Lambda \to R_+$ be a couple of functions with

(c01) φ is proper: $\mathrm{Dom}(\varphi) := \{x \in X; \varphi(x) < \infty\} \neq \emptyset$
(c02) k is increasing: $\lambda \leq \mu \Longrightarrow k(\lambda) \leq k(\mu)$.

(Note that, among all such objects $k(.)$, we have the unitary function $g \in \mathcal{F}(\Lambda, R^0_+)$, introduced as: $g(\lambda) = 1$, $\lambda \in \Lambda$). The quasi-order $(\preceq_{(D,k,\varphi)})$ over X introduced as

(c03) $(x_1, x_2 \in X)$: $x_1 \preceq_{(D,k,\varphi)} x_2$ iff $k(\lambda) d_\lambda(x_1,x_2) + \varphi(x_2) \leq \varphi(x_1)$, $\forall \lambda \in \Lambda$

is antisymmetric—hence, an ordering—on $\mathrm{Dom}(\varphi)$. As before, we want to determine sufficient conditions under which $(\preceq_{(D,k,\varphi)})$ be a Zorn order on $\mathrm{Dom}(\varphi)$. Precisely, these consist in

(I) Boundedness properties of the objective function:
the classes (Pj), $j \in \{1,2,3,4,5,6\}$, we just encountered

(II) Strict positivity properties of the coefficient function

(M1) $k(\Lambda) \subseteq R^0_+$ (k is strictly positive)
(M2) $k(\Lambda) \subseteq R^0_+$ and k is constant:
$k(\lambda) = k^0 g(\lambda) = k^0$, $\forall \lambda \in \Lambda$, for some $k^0 \in R^0_+$
(M3) $k(\Lambda) \subseteq R^0_+$ and k is unitary constant: $k(\lambda) = g(\lambda) = 1, \forall \lambda \in \Lambda$

(III) Global completeness property of the remaining objects:

(gsdc) (X,D,φ) is globally sequentially descending complete: each D-Cauchy sequence (x_n) in X with $(\varphi(x_n))$ = descending, is D-convergent to some $x \in X$ with $[\varphi(x_n) \geq \varphi(x), \forall n]$.

Our main result is

Theorem 10. *Let the proper function $\varphi : X \to R \cup \{\infty\}$ and the increasing coefficient function $k : \Lambda \to R_+$ be such that φ is in the subclass (Pj) for some $j \in \{1,2,3,4,5,6\}$, $k(.)$ has the property (Mp), for some $p \in \{1,2,3\}$, and (X,D,φ) has the property (gsdc). Then, for each $u \in Dom(\varphi)$, there exists $v \in Dom(\varphi)$, with*
(a) *$k(\lambda) d_\lambda(u,v) \leq \varphi(u) - \varphi(v)$, $\forall \lambda \in \Lambda$*
(aa) *$x \in X$, $[k(\lambda) d_\lambda(v,x) \leq \varphi(v) - \varphi(x)$, $\forall \lambda \in \Lambda] \Longrightarrow v = x$.*

Denote the obtained statements as (HVP;Pj;Mp;gsdc), $j \in \{1,2,3,4,5,6\}$, $p \in \{1,2,3\}$; these will be referred to as: "composed" Hamel variational principles. The relationships between them are given by

Lemma 9. *We have [in (ZF-AC)],*
(i) *(HVP;P(j),Mp;gsdc) $\Longleftrightarrow$ (HVP;P(j+1);Mp;gsdc), $\forall j \in \{1,3,5\}$, $\forall p \in \{1,2,3\}$*
(ii) *the propositional map $(j,p) \mapsto (HVP;Pj;Mp;gsdc)$ is decreasing:*
if $(j,p) \leq (j',p')$, then (HVP;Pj;Mp;gsdc) $\Longrightarrow$ (HVP;Pj';Mp';gsdc)
(iii) *(HVP;Pj,M1;gsdc) $\Longrightarrow$ (HVP;Pj,M2;gsdc) $\Longrightarrow$ (HVP;Pj,M3;gsdc) $\Longrightarrow$ (HVP;Pj,M1;gsdc), $\forall j \in \{1,2,3,4,5,6\}$.*

Proof. (i): Let (X,D,φ) be as in the premise of the principle (HVP;P(j);Mp;gsdc). Then, the triple (X,D,ψ), where $\psi(.) := \varphi(.) - \inf \varphi(X)$, fulfills conditions of the principle (ZL;P(j+1);Mp;gsdc); and, from this, we are done.

(ii): Evident, by the definition of the subclasses in question.

(iii): The first and second inclusions are trivial; so, it remains to verify our third inclusion. Let the Fang space (X,D), the proper function φ, and the increasing function $k : \Lambda \to R_+$ be as in the premises of (HVP;Pj;M1;gsdc). Define another family $E = (e_\lambda ; \lambda \in \Lambda)$ of rs-pseudometrics over X according to

(c04) $e_\lambda(x,y) = k(\lambda) d_\lambda(x,y),\ x,y \in X.$

The Λ-sufficiency of E results at once from that of D; and the Λ-monotonicity of the same is reducible to the increasing property of $k(.)$. Finally, we claim that E is Λ-triangular. Let $\lambda \in \Lambda$ be arbitrarily fixed; and $\mu \in \Lambda(\lambda, \leq)$ be given by the Λ-triangular property of D. Again by the increasing property of $k(.)$,

$$e_\lambda(x,z) \leq k(\lambda)[d_\mu(x,y) + d_\mu(y,z)] \leq e_\mu(x,y) + e_\mu(y,z),\ \forall x,y,z \in X;$$

and the assertion follows. Summing up, E is a Fang metric; it may generate a conv-Cauchy structure on X, by the construction we just precise in a previous place. Concerning its connections with the Fang metric D (and its attached conv-Cauchy structure), one has (for all sequences (x_n) in X, and all $x \in X$)

$$[\forall \lambda \in \Lambda] : (x_n \xrightarrow{d_\lambda} x) \Longleftrightarrow (x_n \xrightarrow{e_\lambda} x);\ \text{hence}\ (x_n \xrightarrow{D} x) \Longleftrightarrow (x_n \xrightarrow{E} x);\ (33)$$

as well as (for a generic sequence (x_n) in X)

$$[\forall \lambda \in \Lambda]: d_\lambda\text{-Cauchy} \Longleftrightarrow e_\lambda\text{-Cauchy; hence } D\text{-Cauchy} \Longleftrightarrow E\text{-Cauchy}. \quad (34)$$

The conv-Cauchy structures attached to the Fang metrics D and E are thus equivalent. As a consequence, the triplet (X,E,φ) has the property (gsdc); so that, the couples (X,E) and (φ,g) fulfill conditions of (HVP;Pj;M3;gsdc). By the conclusion of this principle, we then get all desired facts.

Concerning this last aspect, we stress that the introduction of our coefficient function $k(.)$ in our "scalar" setting is related to a better comparison with the vectorial case. However, by the preceding statement, this procedure has a formal character.

(B) Note that, in our setting, the property (gsdc) is obtainable as an intersection of two component families:

(III-a) Completeness properties involving the Fang structure:

(B1) (X,D,φ) is sequentially descending complete: each D-Cauchy sequence (x_n) in X with $(\varphi(x_n))$=descending is D-convergent

(B2) (X,D,φ) is sequentially complete:
each D-Cauchy sequence in X is D-convergent

(III-b) Lower semi-continuity conditions upon the (scalar) objective function:

(V1) φ is sequentially descending (X,D)-lsc:
for each sequence (x_n) in X and each element $x \in X$ with
$x_n \xrightarrow{D} x$ and $(\varphi(x_n))$=descending, we have $[\varphi(x_n) \geq \varphi(x), \forall n]$.
(V2) φ is sequentially (X,D)-lsc: for each sequence (x_n) in X and each element $x \in X$ with $x_n \xrightarrow{D} x$, we have $\liminf_n \varphi(x_n) \geq \varphi(x)$,

This yields the "factor" Hamel variational principles (HVP;Pj;Mp;Bh,Vq), where $j \in \{1,2,3,4,5,6\}$, $p \in \{1,2,3\}$, $h,q \in \{1,2\}$, with the properties

(s-1) (HVP;Pj;Mp;gsdc) $\Longrightarrow$ (HVP;Pj;Mp;Bh,Vq), for all admissible (j,p,h,q)

(s-2) (HVP;P(j);Mp;Bh,Vq) $\Longleftrightarrow$ (HVP;P(j+1);Mp;Bh,Vq), for all $j \in \{1,3,5\}$, $p \in \{1,2,3\}$, $h,q \in \{1,2\}$

(s-3) the propositional map $(j,p,h,q) \mapsto (HVP;Pj;Mp;Bh,Vq)$ is decreasing: $(j,p,h,q) \leq (j',p',h',q')$ gives (HVP;Pj;Mp;Bh,Vq) $\Longrightarrow$ (HVP;Pj';Mp';Bh',Vq').

(C) Concerning the relationships with the Zhu–Li vector variational principles we already presented, the following inclusions hold:

(s-4) (ZL;Qj;Mp;gsdc) $\Longrightarrow$ (HVP;Pj;Mp;gsdc),
$\forall j \in \{1,2,3,4,5,6\}$, $\forall p \in \{1,2,3\}$

(s-5) (ZL;Qj,Mp;Ch,W) $\Longrightarrow$ (HVP;Pj;Mp;Bh,V1),
$\forall j \in \{1,2,3,4,5,6\}$, $\forall p \in \{1,2,3\}$, $\forall h \in \{1,2\}$.

[Just take $Y = R$, $H = K = R_+$ in the quoted principles]. This, by the preceding auxiliary statement, tells us that all "composed" Hamel variational principles (HVP;Pj;Mp;gsdc) and "factor" Hamel variational principles (HVP;Pj;Mp;Bh,Vq), where $j \in \{1,2,3,4,5,6\}$, $p \in \{1,2,3\}$, $h,q \in \{1,2\}$, are deductible from (BB) in (ZF-AC). On the other hand, the relationships with the (composed or factor) Ekeland variational principles are obtainable from the inclusions

(s-6) (HVP;Pj;M3;gsdc) $\Longrightarrow$ (EVP;Pj;gdc), $\forall j \in \{1,2,3,4,5,6\}$

(s-7) (HVP;Pj;M3;Bh,Vq) $\Longrightarrow$ (EVP;Pj;Bh,Vq),
$\forall j \in \{1,2,3,4,5,6\}$, $\forall h,q \in \{1,2\}$.

[Just take Λ as a singleton and $D = \{d\}$, where $d(.,.)$ is a metric on X]. Consequently, all "composed" Hamel variational principles (HVP;Pj;Mp;gsdc) and all "factor" Hamel variational principles (HVP;Pj;Mp;Bh,Vq) deductible from these, where $j \in \{1,2,3,4,5,6\}$, $p \in \{1,2,3\}$, $h,q \in \{1,2\}$, include (EVP-bf) (=the bounded finitary version of (EVP)). Combining with the Dodu–Morillon statement, we get that these Hamel principles are all equivalent with both (BB) and (EVP-bf); hence, with (DC) and/or (EVP) as well.

In particular, (HVP;P3;M1;B2,V1) is just Hamel's variational principle [27] (in short: (HVP)); this also explains our conventions. Note that (HVP)—based on a maximal principle comparable with Brøndsted's [10]—extends the related statement in Fang [21], obtained *via* Zorn maximal techniques. It also includes the contribution due to Hadžić and Žikić [25], founded on the maximal principle in Hicks [29]; we do not give details. In addition, by the above developments, (HVP) is equivalent with both (BB) and (EVP).

(D) The following completion of these facts is to be noted. Let I be some nonempty set. Take a family $F = (f_i; i \in I)$ of rs-pseudometrics over X, supposed

to be *I-sufficient* [$f_i(x,y)=0$, for all $i\in I$ imply $x=y$], and *I-triangular* [for each $i\in I$, there exist $j=j(i)$ and $k=k(i)$ in I such that $f_i(x,z)\le f_j(x,y)+f_k(y,z)$, $\forall x,y,z\in X$]. In this case, the couple (X,F) will be termed a *BMLO space*; see Benbrik et al. [5]. Clearly, any Fang space is a BMLO space as well. But, the reciprocal inclusion is also true. In fact, let Λ stand for the class of all (nonempty) finite parts of I, endowed with the usual inclusion, $(\subseteq)$; note that, $(\Lambda,\subseteq)$ is a directed ordered structure. For each $\lambda\in\Lambda$ define the rs-pseudometric d_λ over X as: $d_\lambda(x,y)=\sup\{f_i(x,y);i\in\lambda\}$, $x,y\in X$. The family $D=(d_\lambda;\lambda\in\Lambda)$ of all these is easily shown to be Λ-sufficient, Λ-monotone, and Λ-triangular; i.e., (X,D) is a Fang space. In addition, all usual F-concepts (like F-convergence and F-Cauchy) are equivalent to their corresponding D-concepts. Hence, all variational results over BMLO spaces established by these authors are completely reducible to those involving Fang spaces we just presented; see also Hamel and Loehne [28]. In particular, this is retainable for the variational principles in uniform spaces (taken as in Bourbaki [8, Chap. 2, Sect. 1]) due to Mizoguchi [37]; because any such structure is a BMLO space. Further aspects may be found in Hadžić and Ovcin [24]; see also Chang et al. [15]. For interesting applications of these facts to Pareto optimality we refer to the 1996 paper by Isac [32] and the references therein.

2.4 (ZF-AC) Approach

In the following, some technical aspects involving the Hamel variational principles we just presented are considered. Let X be a nonempty set; and $(\Lambda,\le)$ be a directed quasi-ordered structure. Take a family $D=(d_\lambda;\lambda\in\Lambda)$ of rs-pseudometrics over X; supposed to be Λ-sufficient, Λ-monotone, and Λ-triangular; by a previous convention, D will be referred to as a Fang metric, and (X,D), as a Fang space. Further, let $\varphi:X\to R\cup\{\infty\}$ be a proper function; and $k:\Lambda\to R_+$ be some increasing function.

(A) By the developments above, it results that, for the deduction, from (BB), of the "composed" Hamel variational principles (HVP;Pj;Mp;gsdc) and the "factor" Hamel variational principles (HVP;Pj;Mp;Bh,Vq), where $j\in\{1,2,3,4,5,6\}$, $p\in\{1,2,3\}$, $h,q\in\{1,2\}$, the operational model is that involving Zhu–Li vector variational principles. According to the argument presented there, (BB) was applied in a "local" way, by means of the point $\theta\in\Lambda$ and its attached section $\Theta:=\Lambda(\theta,\le)$. However, the presence of a "scalar" objective function φ (in place of the vectorial function F) suggests us that a "global" application of (BB) is highly expectable. Note that, by an auxiliary fact, it will suffice that this deduction process be applicable to the composed Hamel variational principle (HVP;P1;M3;gsdc), involving the Fang space (X,D), the proper function $\varphi:X\to R\cup\{\infty\}$, and the unitary increasing function $g:\Lambda\to R_+^0$ (introduced as: $g(\lambda)=1$, $\lambda\in\Lambda$). The quasi-order associated with these data, $(\preceq_{(D,g,\varphi)})$, will be simply denoted as $(\preceq_{(D,\varphi)})$; hence

(d01) $(x_1,x_2\in X)$: $x_1\preceq_{(D,\varphi)}x_2$ iff $d_\lambda(x_1,x_2)+\varphi(x_2)\le\varphi(x_1)$, $\forall\lambda\in\Lambda$.

A positive answer to the posed question is established in the statement below.

Proposition 13. *We have [in (ZF-AC)], (BB) $\Longrightarrow$ (HVP;P1;M3;gsdc); whence, all Hamel variational principles above are deductible from (BB).*

Proof. Let the Fang space (X,D), the proper function $\varphi : X \to R \cup \{\infty\}$, and the unitary function $g : \Lambda \to R_+$ be as in the premises of (HVP;P1;M3;gsdc). Denote (under the previous convention) $X_u := X(u, \preceq_{(D,\varphi)})$. We have to verify that (BB) is applicable (in a global way) to $(X_u, \preceq_{(D,\varphi)})$ and the (proper) function φ (restricted to X_u). Clearly, φ is descending [modulo $(\preceq_{(D,\varphi)})$]. Moreover, let $(x_n; n \geq 0)$ be an ascending [modulo $(\preceq_{(D,\varphi)})$] sequence in X_u:

(d02) $(\forall \lambda \in \Lambda)$: $d_\lambda(x_n, x_m) \leq \varphi(x_n) - \varphi(x_m)$, if $n \leq m$.

The sequence $(\varphi(x_n))$ is descending and bounded from below; hence, a Cauchy one This, along with the working hypothesis above, tells us that (x_n) is D-Cauchy. Taking (gsdc) into account, it follows that $x_n \xrightarrow{D} x$, for some $x \in X$ with $\varphi(x_n) \geq \varphi(x)$, $\forall n$. Combining with the working hypothesis above gives $x_n \preceq_{(D,\varphi)} x$, $\forall n$ (whence $x \in X_u$); so that, $(X_u, \preceq_{(D,\varphi)})$ is sequentially inductive. From (BB) it then follows that, for the starting $u \in X_u$, there exists an element $v \in X_u$, with

(j) $u \preceq_{(D,\varphi)} v$; **(jj)** $v \preceq_{(D,\varphi)} w \in X_u$ implies $\varphi(v) = \varphi(w)$.

The former of these is just our first conclusion in the statement. And the latter one gives our second conclusion of the same. In fact, let $x \in X$ be such that $d_\lambda(v,x) \leq \varphi(v) - \varphi(x)$, $\forall \lambda \in \Lambda$; hence, $v \preceq_{(D,\varphi)} x$. Taking **(j)** into account gives $x \in X_u$; so that (by the conclusion **(jj)** above) $\varphi(v) = \varphi(x)$. Combining with the previous gauge metrical relation, we get $d_\lambda(v,x) = 0$, $\forall \lambda \in \Lambda$; whence $v = x$ (as D is sufficient). This completes the argument.

Remark 4. We stress that, by the very proof of the "composed" result above, one has, in (ZF-AC),

(mr-1) (HVP;P5;M3;gsdc) $\Longrightarrow$ (HVP;P1;M3;gsdc);

hence, (HVP;P5;M3;gsdc) $\Longleftrightarrow$ (HVP;P1;M3;gsdc).

In fact, if (X,D,φ) fulfills conditions of (HVP;P1;M3;gsdc), then (X_u,D,φ) (where $u \in \mathrm{Dom}(\varphi)$) fulfills conditions of (HVP;P5;M3;gsdc); and, from the conclusion of this gauge variational principle, we are done. However, as the statement above shows, the deduction of these gauge principles is based on (BB) (or, equivalently, (DC)); so, it cannot be reached in (ZF-AC).

(B) Remember that, when Λ as a singleton and $D = \{d\}$, [where $d(.,.)$ is a metric on X], then (by the involved conventions)

(mr-2) (HVP;Pj;M3;gsdc) becomes (EVP;Pj;gdc), for $j \in \{1,2,3,4,5,6\}$,

(mr-3) (HVP;Pj;M3;Bh,Vq) becomes (EVP;Pj;Bh,Vq),

$\forall j \in \{1,2,3,4,5,6\}$, $\forall h,q \in \{1,2\}$.

As a consequence of this (and the previous facts)

(mr-4) (DC) $\Longrightarrow$ (BB) $\Longrightarrow$ (HVP;Pj;M3;gsdc) $\Longrightarrow$ (EVP;Pj;gdc) $\Longrightarrow$ (EVP-bf), $\forall j \in \{1,2,3,4,5,6\}$

(mr-5) (DC) $\Longrightarrow$ (BB) $\Longrightarrow$ (HVP;Pj;M3;Bh,Vq) $\Longrightarrow$ (EVP;Pj;Bh,Vq) $\Longrightarrow$ (EVP-bf), $\forall j \in \{1,2,3,4,5,6\}$, $\forall h,q \in \{1,2\}$;

wherefrom, by the Dodu–Morillon statement (in short: (DM)),

(mr-6) (HVP;Pj;M3;gsdc) $\Longleftrightarrow$ (EVP;Pj;gdc), $\forall j \in \{1,2,3,4,5,6\}$

(mr-7) (HVP;Pj;M3;Bh,Vq) $\Longleftrightarrow$ (EVP;Pj;Bh,Vq),

$\forall j \in \{1,2,3,4,5,6\}$, $\forall h,q \in \{1,2\}$.

Note that, by the very arguments involved in (DM), these equivalence properties are available in (ZF-AC+DC). It is our aim in the following to show that, by the "scalar" nature of our setting, some of these relations may be obtained in (ZF-AC); i.e., (DC) may be avoided in certain equivalence properties of this type.

Let X be a nonempty set; and $(\Lambda,\leq)$ be a directed quasi-ordered structure. Further, take a family $D=(d_\lambda;\lambda\in\Lambda)$ of rs-pseudometrics over X, with the properties: Λ-sufficient, Λ-monotone, and Λ-triangular; by our preceding developments, the Fang metric D generates a conv-Cauchy structure on X. Further, let $\varphi: X\to R\cup\{\infty\}$ be a proper function. Denote by $(\preceq_{(D,\varphi)})$ the quasi-order attached to these elements; remember that it is antisymmetric—hence, an order—on $\mathrm{Dom}(\varphi)$.

Generally, the family of rs-pseudometrics $D=(d_\lambda;\lambda\in\Lambda)$ is non-denumerable. For example, in case of Fang spaces constructed from a probabilistic metric space (cf. Fang [21]) or fuzzy metric spaces (cf. Hadžić and Žikić [25]) we have $(\Lambda,\leq):=(]0,1],\geq)$; here, $(\geq)$ is the usual dual ordering on R. Despite this, its associated quasi-order $(\preceq_{(D,\varphi)})$ may be ultimately viewed as a Brøndsted one, by simply taking the supremum in the left-hand side of the relation that introduces it. So, we may ask about the "metrical" aspects of this procedure. Before passing to the effective part, we need some preliminary facts. Denote

$$\text{(d03)}\quad \Delta(x,y)=\sup\{d_\lambda(x,y);\lambda\in\Lambda\},\ x,y\in X.$$

Since all members of D are rs-pseudometrics, Δ is also endowed with such properties. Moreover Δ is *triangular* $[\Delta(x,z)\leq\Delta(x,y)+\Delta(y,z),\forall x,y,z\in X]$, since D is Λ-triangular; and, finally, Δ is *sufficient* $[\Delta(x,y)=0\Longrightarrow x=y]$; because so is D. Summing up, Δ is a generalized metric on X, in the Luxemburg–Jung sense [33, 36]. It allows us introducing a conv-Cauchy structure on X as follows. Letting $(x_n;n\geq 0)$ in X and the point $x\in X$, let us say that this sequence Δ-*converges* towards x (written as: $x_n\xrightarrow{\Delta}x$), provided $\Delta(x_n,x)\to 0$ as $n\to\infty$; or, equivalently,

$$\forall\varepsilon>0,\ \exists i=i(\varepsilon):\ n\geq i\Longrightarrow\Delta(x_n,x)<\varepsilon.$$

The set of all such x is denoted $\lim_n(x_n)$; when it is nonempty, we say that $(x_n;n\geq 0)$ is Δ-*convergent*; note that, in such a case, $\lim_n(x_n)$ is a singleton. Further, call the sequence $(x_n;n\geq 0)$, Δ-*Cauchy*, when $\Delta(x_m,x_n)\to 0$ as $m,n\to\infty$, $m\leq n$; i.e.,

$$\forall\varepsilon>0,\ \exists j=j(\varepsilon):\ j\leq m\leq n\Longrightarrow\Delta(x_m,x_n)<\varepsilon.$$

By the metrical properties of Δ, any Δ-convergent sequence in X is Δ-Cauchy; the reciprocal is not in general valid.

Having these precise, the natural question to be posed is that of clarifying the relationships between the conv-Cauchy attached to Δ and the one attached to the Fang metric $D=(d_\lambda;\lambda\in\Lambda)$. First, by the introduced conventions, we have the following Relative statement:

Lemma 10. *The generic local inclusions hold:*

$$(\forall(x_n),\forall x):\ [x_n \xrightarrow{\Delta} x] \Longrightarrow [x_n \xrightarrow{D} x] \tag{35}$$

$$\textit{(for each sequence): } \Delta\textit{-Cauchy} \Longrightarrow D\textit{-Cauchy.} \tag{36}$$

The reciprocal inclusions are not in general true; because the conv-Cauchy structure attached to D is finer than that induced by the generalized metric Δ. A completion of these facts is contained in the Metrical convergence statement below.

Lemma 11. *Under these notations,*

$$\textit{(}\forall(x_n),\forall x\textit{): } (x_n) \textit{ is } \Delta\textit{-Cauchy, and } x_n \xrightarrow{D} x \textit{ imply } x_n \xrightarrow{\Delta} x. \tag{37}$$

Proof. Let (x_n) be a Δ-Cauchy sequence in X, so as (for some $x \in X$)

$$x_n \xrightarrow{D} x \text{ (hence } d_\lambda(x_n,x) \to 0, \text{ for each } \lambda \in \Lambda).$$

By definition, for each $\beta > 0$ there exists some rank $n(\beta)$ in such a way that $\Delta(x_i,x_j) \leq \beta$ (hence $d_\lambda(x_i,x_j) \leq \beta$, $\forall \lambda \in \Lambda$), whenever $n(\beta) \leq i \leq j$. Let the rank $i \geq n(\beta)$ be arbitrarily fixed; and, for each $\lambda \in \Lambda$, let $\mu \in \Lambda(\lambda,\leq)$ be the index given by the Λ-triangular property of D. We 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 i like before)

$$d_\lambda(x_i,x) \leq \beta,\ \forall \lambda \in \Lambda \text{ (hence } \Delta(x_i,x) \leq \beta).$$

This, by the arbitrariness of β, yields $x_n \xrightarrow{\Delta} x$; as claimed.

(C) Having these precise, we may now pass to the effective part of our developments . Let $(\preceq_{(\Delta,\varphi)})$ stand for the quasi-order on X:

(d04) $(x_1,x_2 \in X)$: $x_1 \preceq_{(\Delta,\varphi)} x_2$ iff $\Delta(x_1,x_2) + \varphi(x_2) \leq \varphi(x_1)$;

it is antisymmetric—hence, an order—on $\mathrm{Dom}(\varphi)$, as it can be directly seen. The relationships between this and the initial quasi-order $(\preceq_{(D,\varphi)})$ are established in the Identity statement:

Lemma 12. *Under the above conventions, we have*

$$\textit{(}\forall x_1,x_2 \in X\textit{): } x_1 \preceq_{(D,\varphi)} x_2 \textit{ iff } x_1 \preceq_{(\Delta,\varphi)} x_2. \tag{38}$$

In other words: these two quasi-orders are identical.

The verification is immediate, by the very definition of our generalized metric $\Delta(.,.)$; so, further details are not necessary.

We are now in position to give the announced result.

Proposition 14. *We have [in (ZF-AC)]:*

(i) *(EVP;Pj;gdc)* $\Longrightarrow$ *(HVP;Pj,M3;gsdc) (hence,*
(EVP;Pj;gdc) $\Longleftrightarrow$ *(HVP;Pj,M3;gsdc)), for all* $j \in \{1,2,3,4,5,6\}$
(ii) *(EVP;Pj;B1,V1)* $\Longrightarrow$ *(HVP;Pj,M3;B1,V1) (hence,*
(EVP;Pj;B1,V1) $\Longleftrightarrow$ *(HVP;Pj,M3;B1,V1)),* $\forall j \in \{1,2,3,4,5,6\}$
(iii) *(EVP;Pj;B2,V2)* $\Longrightarrow$ *(HVP;Pj,M3;B2,V2) (hence,*
(EVP;Pj;B2,V2) $\Longleftrightarrow$ *(HVP;Pj,M3;B2,V2)),* $\forall j \in \{1,2,3,4,5,6\}$.

Proof. Let the Fang space (X,D), the proper function $\varphi : X \to R \cup \{\infty\}$ (and the increasing unitary function $g : \Lambda \to R_+^0$ be given. Take an element $u \in \mathrm{Dom}(\varphi)$ and put $X_u := X(u, \preceq_{(D,\varphi)})$. By the Identity statement, $X_u = X(u, \preceq_{(\Delta,\varphi)})$; or, equivalently,

$$x \in X_u \text{ iff } \Delta(u,x) \le \varphi(u) - \varphi(x); \tag{39}$$

in addition, the restriction of Δ to X_u is a standard metric. There are three main steps in the argument.

(I) Suppose that (X,D,φ) has the property (gsdc). We claim that the triple (X_u,Δ,φ) has the property (gdc). In fact, let $(x_n; n \ge 0)$ be a sequence in X_u; hence

(d05) $\Delta(u,x_n) \le \varphi(u) - \varphi(x_n)$, $\forall n$.

Assume that (x_n) is Δ-Cauchy, and $(\varphi(x_n))$ is descending. By the Relative statement, (x_n) is a D-Cauchy sequence in X_u. This, along with (gsdc), tells us that $x_n \xrightarrow{D} x$, for some $x \in X$; with, in addition, $\varphi(x_n) \ge \varphi(x)$, $\forall n$. On the other hand, by the Metrical convergence statement, $x_n \xrightarrow{\Delta} x$; so that, passing to limit in the working hypothesis above,

$$\Delta(u,x) \le \varphi(u) - \varphi(x); \text{ hence } x \in X_u.$$

This proves our claim. But then, by the conclusions of (EVP;Pj;gdc) relative to (X_u,Δ,φ), we are done.

(II) Assume that (X,D,φ) has the property (B1), and φ has the property (V1) relative to (X,D). We claim that (X_u,Δ,φ) has the property (B1) and φ has the property (V1) relative to (X_u,Δ). In fact, assume that (x_n) is a Δ-Cauchy sequence in X_u such that $(\varphi(x_n))$ is descending. By the Relative statement, (x_n) is D-Cauchy; so that, from (B1), $x_n \xrightarrow{D} x$, for some $x \in X$; with, in addition, $[\varphi(x_n) \ge \varphi(x), \forall n]$. Combining with the Metrical convergence statement yields $x_n \xrightarrow{\Delta} x$; this, along with the relation involving $(\varphi(x_n))$, gives [passing to limit in the metrical relation concerning our sequence], $x \in X_u$; so that, (B1) holds for our data. Moreover, let the sequence (x_n) in X_u be such that $x_n \xrightarrow{\Delta} x$ for some $x \in X_u$; and $(\varphi(x_n))$ is descending. By the Relative statement, $x_n \xrightarrow{D} x$; so that, by (V1), $\varphi(x_n) \ge \varphi(x)$, $\forall n$; whence, (V1) holds for the same data. Summing up, our claim follows. But then, an application of (EVP;Pj;B1,V1) to (X_u,Δ,φ) gives us all conclusions we need.

(III) Assume that (X,D,φ) has the property (B2), and φ has the property (V2) relative to (X,D). We claim that (X_u,Δ,φ) has the property (B2) and φ has the

property (V2) relative to (X_u, Δ). In fact, assume that (x_n) is a Δ-Cauchy sequence in X_u. By the Relative statement, (x_n) is D-Cauchy; so that, by (B2), $x_n \xrightarrow{D} x$, for some $x \in X$. Combining with the Metrical convergence statement yields $x_n \xrightarrow{\Delta} x$; and this, in combination of (V2), gives $\liminf_n \varphi(x_n) \geq \varphi(x)$. Passing to $\liminf_n$ in the metrical relation involving our sequence, we necessarily have $x \in X_u$; so that, (B2) holds for our data. Moreover, let (x_n) be a sequence in X_u fulfilling $x_n \xrightarrow{\Delta} x$, for some $x \in X_u$. By the Relative statement, $x_n \xrightarrow{D} x$; so that, by (V2), $\liminf_n \varphi(x_n) \geq \varphi(x)$; whence, (V2) holds for our data. Summing up, our claim follows. But then, an application of (EVP;Pj;B2,V2) to (X_u, Δ, φ) gives us all desired conclusions.

As a consequence of the conclusion above, (DC) is not needed to establish the logical equivalence between the Hamel variational principles

(HVP;Pj;M3;gsdc), (HVP;Pj;M3;B1,V1), (HVP;Pj;M3;B2,V2),

and their corresponding Ekeland variational principles

(EVP;Pj;gdc), (EVP;Pj;B1,V1), (EVP;Pj;B2,V2),

where $j \in \{1,2,3,4,5,6\}$. However, as already shown, all these principles are deductible with the aid of (BB); or, equivalently, (DC). Further aspects may be found in Hamel and Loehne [28].

3 Sequential Gauge Variational Statements

3.1 Gauge Ordering Principles

Let $(X, \leq)$ be a quasi-ordered structure. Further, let $\Phi = (\varphi_i; i \geq 0)$ be a sequence of maps in $\mathcal{F}(X, R \cup \{-\infty, \infty\})$; referred to as a *gauge function* over $\mathcal{F}(X, R \cup \{-\infty, \infty\})$. Call $z \in X$, $(\leq, \Phi)$-*maximal*, provided z is $(\leq, \varphi_i)$-maximal, for each $i \geq 0$. The class of all these will be denoted as $\max(X; \leq; \Phi)$; hence

$$\max(X; \leq; \Phi) = \cap\{\max(X; \leq; \varphi_i); i \geq 0\}.$$

To get such points, assume that $(X, \leq)$ is sequentially inductive and

(a01) Φ is $(\leq)$-decreasing: φ_i is $(\leq)$-decreasing, $\forall i \geq 0$.

Further, for each $j \in \{0,1,2,3,4,5,6\}$, let [Pj] stand for the [attached to (Pj)] subclass of all gauge functions over $\mathcal{F}(X, R \cup \{-\infty, \infty\})$ introduced as:

(a02) Φ belongs to the subclass [Pj] iff φ_i belongs to the subclass (Pj), $\forall i \geq 0$.

The following "multiple" gauge ordering principle enters into discussion:

Theorem 11. *Let $(X,\leq)$ be sequentially inductive, Φ be $(\leq)$-decreasing, and*

(a03) Φ belongs to the subclass [Pj], for some $j \in \{0,1,2,3,4,5,6\}$.

Then, $\max(X;\leq;\Phi)$ *is*

(i) *$(\leq)$-cofinal in X [$\forall u \in X$, $\exists v \in \max(X;\leq;\Phi)$: $u \leq v$]*
(ii) *$(\leq)$-invariant in X [$v \geq u \in \max(X;\leq;\Phi) \Longrightarrow v \in \max(X;\leq;\Phi)$].*

For simplicity, we indicate these gauge ordering principles as (CUg;Pj), where $j \in \{0,1,2,3,4,5,6\}$. Note that (CUg;P0) is the gauge variant of the (global) ordering principle (CU); so that, it will be written as (CUg). On the other hand, (CUg;P3) is nothing else than the gauge variant of (BB), obtained in Turinici [44]; denoted as (BBg).

Concerning the relationships between these, the following Gauge Equivalence statement is available:

Lemma 13. *We have [in (ZF-AC)]:*

(i) *(CUg;P(j)) $\Longleftrightarrow$ (CUg;P(j+1)), $\forall j \in \{1,3,5\}$*
(ii) *(CUg;P5) $\Longrightarrow$ (CUg;P0) $\Longrightarrow$ (CUg;P1) $\Longrightarrow$ (CUg;P3) $\Longrightarrow$ (CUg;P5)*

Hence, all these principles are equivalent in (ZF-AC).

The proof of this mimics the one of its corresponding non-gauge (global) Equivalence statement; so, it will be omitted.

Note that the obtained relations cannot assure us that these principles are deductible in (ZF-AC+DC). This, however, holds; as results from

Proposition 15. *We have [in (ZF-AC)], (DC) $\Longrightarrow$ (CUg;P5); hence (by the above) (DC) $\Longrightarrow$ (CUg;Pj), for each $j \in \{0,1,2,3,4,5,6\}$.*

Proof. Let the premises of (CUg;P5) be accepted. From (CU;P5) [valid in (ZF-AC+DC)], the subset $Y_i := \max(X;\leq;\varphi_i)$ is nonempty $(\leq)$-cof-inv, for each $i \geq 0$. This, along with Cof-inv statement [valid in (ZF-AC+DC) too], tells us that the intersection of these, $\cap\{Y_i; i \geq 0\} = \max(X;\leq;\Phi)$, has the same properties.

Remark 5. By the very arguments above, one gets, in (ZF-AC+DC):

$$\text{(CU;Pj)} \Longrightarrow \text{(CUg;Pj)} \Longrightarrow \text{(CU;Pj)},\ j \in \{0,1,2,3,4,5,6\}.$$

Hence, for each $j \in \{0,1,2,3,4,5,6\}$, the ordering principle (CU;Pj) is equivalent with its gauge version (CUg;Pj). This, however, cannot be established on (ZF-AC); because of Cof-inv statement.

Finally, an interesting question to be posed is that of such inclusion chains being retainable beyond the countable case. Unfortunately, this is not in general possible; see Isac [31] for details.

3.2 Gauge Variational Principles

Let X be a nonempty set. By a *pseudometric* over X we shall mean any map $d : X \times X \to R_+$. If, in addition, d is *symmetric* $[d(x,y) = d(y,x), \forall x,y \in X]$, *triangular* $[d(x,z) \leq d(x,y) + d(y,z), \forall x,y,z \in X]$, and *reflexive* $[d(x,x) = 0, \forall x \in X]$, we say that it is a *semimetric* over X; and (X,d) is a *semimetric space*. The sequential convergence $(\overset{d}{\longrightarrow})$ attached to d is introduced as: the sequence (x_n) in X, *d-converges* to $x \in X$ (and we write: $x_n \overset{d}{\longrightarrow} x$), iff $d(x_n,x) \to 0$ as $n \to \infty$. This also reads: x is a *d-limit* of (x_n); when x is generically taken, we say that (x_n) is *d-convergent*. Further, the d-Cauchy property of a sequence (x_n) in X means: $d(x_m,x_n) \to 0$ as $m,n \to \infty$, $m \leq n$. By the imposed upon d properties, each d-convergent sequence is d-Cauchy too; the reciprocal is not in general valid.

Let $D = (d_i; i \geq 0)$ be a denumerable family of semimetrics on X; supposed to be *sufficient* $[d_i(x,y) = 0, \forall i \geq 0$, implies $x = y]$. Then, D is called a *gauge metric* on X; and (X,D) will be referred to as a *gauge space*. We say that the sequence (x_n) in X, *D-converges* to $x \in X$ (and we write $x_n \overset{D}{\longrightarrow} x$), when it d_i-converges to x, for each $i \geq 0$. The set of all such points x will be denoted $\lim_n(x_n)$; if it is nonempty, then (x_n) is called *D-convergent*. Note that, in this case, $\lim_n(x_n)$ is a singleton, because the gauge metric D is sufficient. Likewise, the sequence (x_n) in X is called *D-Cauchy*, when it is d_i-Cauchy, for each $i \geq 0$. By the remark above, any D-convergent sequence is D-Cauchy; the reciprocal is not in general true.

(A) Having these precise, let (X,D) be a gauge space, Further, let $\Phi = (\varphi_i; i \geq 0)$ be a gauge function over $\mathcal{F}(X, R \cup \{\infty\})$, with

(b01) Φ is *strongly proper*: φ_i is proper, for all $i \geq 0$, and $\mathrm{Dom}(\Phi) := \cap\{\mathrm{Dom}(\varphi_i); i \geq 0]\}$ is nonempty.

Define a quasi-order $(\preceq_{(D,\Phi)})$ on X as

(b02) $x_1 \preceq_{(D,\Phi)} x_2$ iff $d_i(x_1,x_2) + \varphi_i(x_2) \leq \varphi_i(x_1), \forall i$;

note that it is antisymmetric—hence, an order—on $\mathrm{Dom}(\Phi)$. It is our objective in the following to determine sufficient conditions under which $(\preceq_{(D,\Phi)})$ be a Zorn order on $\mathrm{Dom}(\Phi)$. Precisely, these consist in

(I) Boundedness properties of the objective gauge function:
the classes [Pj], $j \in \{1,2,3,4,5,6\}$, we just encountered

(II) Global completeness property of the objects family:

(gsdc) (X,D,Φ) is globally sequentially descending complete: each D-Cauchy sequence (x_n) in X with $(\Phi(x_n))$=descending, is D-convergent to some $x \in X$ with $[\Phi(x_n) \geq \Phi(x), \forall n]$.

Here, the gauge properties above mean

(b03) $(\Phi(x_n))$=descending iff $[(\varphi_i(x_n))$=descending, $\forall i \geq 0]$
(b04) $(x,y \in X)$: $\Phi(x) \geq \Phi(y)$ iff $\varphi_i(x) \geq \varphi_i(y), \forall i$.

Our main result is

Theorem 12. *Suppose that the strongly proper gauge function* $\Phi = (\varphi_i; i \geq 0)$ *belongs to the subclass [Pj] for some* $j \in \{1,2,3,4,5,6\}$*, and* (X, D, Φ) *has the property (gsdc). Then, for each* $u \in Dom(\Phi)$*, there exists* $v \in Dom(\Phi)$*, with*

(a) $d_i(u,v) \leq \varphi_i(u) - \varphi_i(v)$, $\forall i$

(aa) $x \in X$, $[d_i(v,x) \leq \varphi_i(v) - \varphi(x)$, $\forall i] \Longrightarrow v = x$.

Denote the obtained statements as (BCK;Pj;gsdc), where $j \in \{1,2,3,4,5,6\}$; these will be referred to as: "composed" Bae–Cho–Kim gauge variational principles. The relationships between them are given by

Lemma 14. *We have [in (ZF-AC)],*

(i) *(BCK;P(j);gsdc)* $\Longleftrightarrow$ *(BCK;P(j+1);gsdc),* $\forall j \in \{1,3,5\}$

(ii) *the propositional map* $j \mapsto (BCK; Pj; gsdc)$ *is decreasing: if* $j \leq j'$*, then (BCK;Pj;gsdc)* $\Longrightarrow$ *(BCK;Pj';gsdc).*

Proof. (i): Let (X, D, Φ) be as in the premise of (BCK;P(j);gsdc). Then, the triple $(X, D, \Psi := (\psi_i; i \geq 0))$, where $(\psi_i(.) := \varphi_i(.) - \inf \varphi_i(X); i \geq 0)$ fulfills conditions of (BCK;P(j+1);gsdc); and, from this, we are done.

(ii): Evident, by the introduced definitions.

Now, the property (gsdc) is obtainable as an intersection of two components:

(II-a) Completeness properties involving the Fang structure:

[B1] (X, D, Φ) is descending complete:
each D-Cauchy sequence (x_n) in X with $(\Phi(x_n))$=descending is D-convergent

[B2] (X, D, Φ) is complete: each D-Cauchy sequence in X is D-convergent

(II-b) Lower semi-continuity conditions upon the gauge function:

[V1] Φ is descending (X, D)-lsc:
for each sequence (x_n) in X and each element $x \in X$ with
$x_n \stackrel{D}{\longrightarrow} x$ and $(\Phi(x_n))$=descending, we have $\lim_n \Phi(x_n) \geq \Phi(x)$

[V2] Φ is (X, D)-lsc: for each sequence (x_n) in X and each element $x \in X$ with
$x_n \stackrel{D}{\longrightarrow} x$, we have $\liminf_n \Phi(x_n) \geq \Phi(x)$.

Here, by definition,

(b05) $\lim_n \Phi(x_n) \geq \Phi(x)$, iff $[\lim_n \varphi_i(x_n) \geq \varphi_i(x)$, $\forall i \geq 0]$

(b06) $\liminf_n \Phi(x_n) \geq \Phi(x)$, iff $[\liminf_n \varphi_i(x_n) \geq \varphi_i(x)$, $\forall i \geq 0]$.

This yields a family of "factor" type Bae–Cho–Kim gauge variational principles (BCK;Pj;Bh,Vq), where $j \in \{1,2,3,4,5,6\}$, $h, q \in \{1,2\}$; with the properties

(am-1) (BCK;Pj;gdc) $\Longrightarrow$ (BCK;Pj;Bh,Vq), for all admissible (j, h, q)

(am-2) (BCK;P(j);Bh,Vq) $\Longleftrightarrow$ (BCK;P(j+1);Bh,Vq),
for all $j \in \{1,3,5\}$, $h, q \in \{1,2\}$

(am-3) the propositional map $(j, h, q) \mapsto (BCK; Pj; Bh, Vq)$ is decreasing:
if $(j, h, q) \leq (j', h', q')$, then (BCK;Pj;Bh,Vq) $\Longrightarrow$ (BCK;Pj';Bh',Vq').

Now, to get all these, it will suffice proving that the "weakest" variational principle (BCK;P1;gsdc) is deductible in (ZF-AC+DC). This follows from

Proposition 16. *We have [in (ZF-AC)] ((DC) $\Longrightarrow$) (BBg) $\Longrightarrow$ (BCK;P1;gsdc); hence, all these gauge variational principles are deductible in (ZF-AC+DC).*

Proof. Let (X,D,Φ) be as in the premises of (BCK;P1;gsdc). Denote for simplicity $(\preceq) := (\preceq_{(D,\Phi)})$; hence

$$x \preceq y \text{ iff } [d_i(x,y) + \varphi_i(y) \leq \varphi_i(x), \forall i \geq 0].$$

Remember that $(\preceq)$ is antisymmetric [hence, an order] on $\mathrm{Dom}(\Phi)$; in addition, $\mathrm{Dom}(\Phi)$ is $(\preceq)$-invariant. Moreover, both these properties remain valid on the subset $X_u := X(u,\preceq) \subseteq \mathrm{Dom}(\Phi)$. Denote again by (D,Φ) the restriction of the initial couple (D,Φ) to X_u. We claim that conditions of (BBg):=(CUg;P3) (i.e.: the gauge ordering principles in Turinici [44]) are fulfilled over $(X_u,\preceq,\Phi)$. Clearly, Φ is $(\preceq)$-decreasing and belongs to the subclass [P3] (relative to X_u). So, it remains to show that $(X_u,\preceq)$ is sequentially inductive. Let (x_n) be a $(\preceq)$-ascending sequence in X_u:

(b07) $(\forall i \geq 0)$: $d_i(x_n,x_m) \leq \varphi_i(x_n) - \varphi_i(x_m)$, if $n \leq m$.

By [P3], it follows that, for each $i \geq 0$, the sequence $(\varphi_i(x_n))$ is descending and bounded from below; hence, a Cauchy one. This, along with the working hypothesis, tells us that (x_n) is a D-Cauchy sequence in X_u with $(\Phi(x_n))$=descending. By [gsdc], there must be some $y \in X$ with $x_n \xrightarrow{D} y$; and, in addition, $[\Phi(x_n) \geq \Phi(x), \forall n]$. For each pair (i,n) of natural numbers, we have

$$d_i(x_n,y) \leq d_i(x_n,x_m) + d_i(x_m,y) \leq \varphi_i(x_n) - \varphi_i(x_m) + d_i(x_m,y) \leq \varphi_i(x_n) - \varphi_i(y) + d_i(x_m,y), \quad \forall m \geq n.$$

Passing to limit as $m \to \infty$ one derives

$$(\forall n)\text{: } [d_i(x_n,y) \leq \varphi_i(x_n) - \varphi_i(y),\ \forall i]\text{; hence, } x_n \preceq y.$$

This firstly shows that $y \in X_u$; and secondly, that y is an upper bound (modulo $(\preceq)$) of (x_n). Summing up, $(X_u,\preceq)$ is sequentially inductive; as claimed. From (BBg) it then follows that, for the starting $u \in X_u$, there exists $v \in X_u$ with

(j) $u \preceq v$; **(jj)** $v \preceq w \in X_u \Longrightarrow [\varphi_i(v) = \varphi_i(w), \forall i]$.

The former of these is just the first conclusion in the statement. And the latter one gives at once the second conclusion of the same. In fact, let $x \in X$ be such that $[d_i(v,x) \leq \varphi_i(v) - \varphi_i(x), \forall i]$. As a consequence, $z \preceq x$ (hence, $x \in X_u$); so that (by the assertion **(jj)** above) $\varphi_i(v) = \varphi_i(x)$, $\forall i$. This (by the working hypothesis about x), yields $[d_i(z,x) = 0, \forall i]$; so that (as D is sufficient) $z = x$. This ends the argument.

We stress that, by the very proof of the "composed" result above, one has, in (ZF-AC),

(am-4) (BCK;P5;gsdc) $\Longrightarrow$ (BCK;P1;gsdc);

hence, (BCK;P5;gsdc) $\Longleftrightarrow$ (BCK;P1;gsdc).

In fact, if (X,D,Φ) fulfills conditions of (BCK;P1;gsdc), then (X_u,D,Φ) (where $u \in \mathrm{Dom}(\Phi)$) fulfills conditions of (BCK;P5;gsdc); and, from the conclusion of

this gauge variational principle, we are done. However, as the statement above shows, the deduction of these gauge principles requires the system (ZF-AC+DC). Similar properties are valid for the families of "factor" gauge variational principles ((BCK;Pj;B1,V1); $j \in \{1,2,3,4,5,6\}$) and ((BCK;Pj;B2,V2); $j \in \{1,2,3,4,5,6\}$); we do not give details.

(B) In particular, assume that the gauge metric $D = (d_i; i \geq 0)$ and the gauge function $\Phi = (\varphi_i; i \geq 0)$ are constant:

(b08) $d_i = d$, $\varphi_i = \varphi$, $\forall i \geq 0$;

here, d is a metric over X and $\varphi : X \to R \cup \{\infty\}$ is a function. Then, the Bae–Cho–Kim gauge variational principles we just encountered are identical with their corresponding Ekeland (non-gauge) variational principles; so that

$$\begin{array}{l} \text{(BCK;Pj;Bh,Vq)} \Longrightarrow \text{(EVP;Pj;Bh,Vq)}, \\ \forall j \in \{1,2,3,4,5,6\},\ \forall h,q \in \{1,2\}. \end{array} \tag{40}$$

This may be also expressed as: the gauge variational principles in question are gauge versions of their corresponding non-gauge variational principles. For example, (BCK;P1;B2,V2)—denoted as (BCK)—is the gauge version of the principle (EVP;P1,B2,V2):=(EVP). On the other hand, the "strongest" gauge principle in this series, (BCK;P5;B2,V2), is the gauge version of the Ekeland principle (EVP;P5;B2,V2):=(EVP-bf) (=the bounded finitary variant of (EVP)); we denote it as (BCK-bf).

Finally, we note that the gauge variational principle (BCK;P3;B2,V2) is just the 1982 one in Turinici [44]; likewise, the gauge variational principle (BCK;P4,B2,V2) is nothing else than the 2011 statement in Bae et al. [3]; this, among others, explains the conventions.

Concerning the relationships between these gauge variational statements, a direct application of the Dodu–Morillon statement gives the following combined answer:

Proposition 17. *We have [in (ZF-AC)]:*

$$\begin{array}{l} \textit{(DC)} \Longrightarrow \textit{(BCK;Pj;gsdc)} \Longrightarrow \textit{(BCK;Pj;Bh,Vq)} \Longrightarrow \textit{(BCK-bf)} \\ \Longrightarrow \textit{(EVP-bf)} \Longrightarrow \textit{(DC)},\ \forall j \in \{1,2,3,4,5,6\},\ \forall h,q \in \{1,2\}. \end{array} \tag{41}$$

As a consequence of this,

- ***(i)*** *all gauge variational principles above are equivalent with (DC); hence, with (BB) and/or (EVP) as well*
- ***(ii)*** *any gauge variational principle in this series is equivalent with its non-gauge version.*

Despite this equivalence, these gauge variational principles are useful tools in practice. Further aspects in this direction may be found in the above quoted papers.

References

1. Altman, M.: A generalization of the Brezis-Browder principle on ordered sets. Nonlinear Anal. **6**, 157–165 (1982)
2. Artzner, P., Delbean, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
3. Bae, J.-S., Cho, S.-H., Kim, J.-J.: An Ekeland type variational principle on gauge spaces with applications to fixed point theory, drop theory and coercivity. Bull. Korean Math. Soc. **48**, 1023–1032 (2011)
4. Bao, T.Q., Khanh, P.Q.: Are several recent generalizations of Ekeland's variational principle more general than the original principle? Acta Math. Vietnam. **28**, 345–350 (2003)
5. Benbrik, A., Mbarki, A., Lahrech, S., Ouahab, A.: Ekeland's principle for vector-valued maps based on the characterization of uniform spaces via families of generalized quasi-metrics. Lobachevskii J. Math. **21**, 33–44 (2006)
6. Bernays, P.: A system of axiomatic set theory: Part III. Infinity and enumerability analysis. J. Symb. Log. **7**, 65–89 (1942)
7. Bourbaki, N.: Sur le théorème de Zorn. Archiv Math. **2**, 434–437 (1949/1950)
8. Bourbaki, N.: General Topology (Chaps. 5–10). Springer, Berlin (1989)
9. Brezis, H., Browder, F.E.: A general principle on ordered sets in nonlinear functional analysis. Adv. Math. **21**, 355–364 (1976)
10. Brøndsted. A.: On a lemma of Bishop and Phelps. Pac. J. Math. **55**, 335–341 (1974)
11. Brøndsted, A.: Fixed points and partial orders. Proc. Am. Math. Soc. **60**, 365–366 (1976)
12. Brunner, N.: Topologische Maximalprinzipien. Zeitschr. Math. Logik Grundl. Math. **33**, 135–139 (1987)
13. Cârjă, O., Ursescu, C.: The characteristics method for a first order partial differential equation. An. Şt. Univ. "A. I. Cuza" Iaşi (Sect I-a, Mat.) **39**, 367–396 (1993)
14. Cârjă, O., Necula, M., Vrabie, I.I.: Viability, Invariance and Applications. North Holland Mathematics Studies, vol. 207. Elsevier B.V., Amsterdam (2007)
15. Chang, S.S., Cho, Y.J., Lee, B.S., Jung, J.S., Kang, S.M.: Coincidence point theorems and minimization theorems in fuzzy metric spaces. Fuzzy Sets Syst. **88**, 119–127 (1997)
16. Cohen, P.J.: Set Theory and the Continuum Hypothesis. Benjamin, New York (1966)
17. Cristescu, R.: Topological Vector Spaces. Noordhoff Intl. Publishers, Leyden (1977)
18. Dodu, J., Morillon, M.: The Hahn-Banach property and the Axiom of Choice. Math. Log. Q. **45**, 299–314 (1999)
19. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. **47**, 324–353 (1974)
20. Ekeland, I.: Nonconvex minimization problems. Bull. Am. Math. Soc. **1**, 443–474 (1979)
21. Fang, J.X.: The variational principle and fixed point theorems in certain topological spaces. J. Math. Anal. Appl. **202**, 398–412 (1996)
22. Goepfert, A., Tammer, C., Zălinescu, C.: On the vectorial Ekeland's variational principle and minimal points in product spaces. Nonlinear Anal. **39**, 909–922 (2000)
23. Goepfert, A., Riahi, H., Tammer, C., Zălinescu, C.: Variational Methods in Partially Ordered Spaces. Canadian Mathematical Society Books in Mathematics, vol. 17. Springer, New York (2003)
24. Hadžić, O., Ovcin, Z.: Fixed point theorem in fuzzy metric spaces and probabilistic metric spaces. Rev. Res. Fac. Sci. Novi Sad Univ. (Math. Ser.). **24**, 197–209 (1994)
25. Hadžić, O., Žikić, T.: On Caristi's fixed point theorem in F-type topological spaces. Novi Sad J. Math. **28**, 91–98 (1998)
26. Hamel, A.: Variational Principles on Metric and Uniform Spaces. Habilitation Thesis, Martin-Luther University, Halle-Wittenberg, Germany (2005)
27. Hamel, A.: Equivalents to Ekeland's variational principle in uniform spaces. Nonlinear Anal. **62**, 913–924 (2005)

28. Hamel, A., Loehne, A.: A minimal point theorem in uniform spaces. In: Agarwal, R.P., O'Regan, D. (eds.) Nonlinear Analysis and Applications: To V. Lakshmikantham on his 80th birthday, vol. 1, pp. 577–593. Kluwer, Dordrecht (2003)
29. Hicks, T.L.: Some fixed point theorems. Radovi Math. **5**, 115–119 (1989)
30. Hyers, D.H., Isac, G., Rassias, T.M.: Topics in Nonlinear Analysis and Applications. World Scientific Publishing, Singapore (1997)
31. Isac, G.: Sur l'existence de l'optimum de Pareto. Rivista Mat. Univ. Parma (Serie IV) **9**, 303–325 (1983)
32. Isac, G.: The Ekeland's principle and Pareto ϵ-efficiency. In: Tamiz, M. (ed.) Multi-Objective Programming and Goal Programming. Lecture Notes in Economics and Mathematical Systems, vol. 432, pp. 148–163. Springer, Berlin (1996)
33. Jung, C.F.K.: On generalized complete metric spaces. Bull. Am. Math. Soc. **75**, 113–116 (1969)
34. Kang, B.G., Park, S.: On generalized ordering principles in nonlinear analysis. Nonlinear Anal. **14**, 159–165 (1990)
35. Kasahara, S.: On some generalizations of the Banach contraction theorem. Publ. Res. Inst. Math. Sci. Kyoto Univ. **12**, 427–437 (1976)
36. Luxemburg, W.A.J.: On the convergence of successive approximations in the theory of ordinary differential equations (II). Indagationes Math. **20**, 540–546 (1958)
37. Mizoguchi, N.: A generalization of Brøndsted's result and its applications. Proc. Am. Math. Soc. **108**, 707–714 (1990)
38. Moore, G.H.: Zermelo's Axiom of Choice: Its Origin, Development and Influence. Springer, New York (1982)
39. Moskhovakis, Y.: Notes on Set Theory. Springer, New York (2006)
40. Nachbin, L.: Topology and Order. D. van Nostrand Comp. Inc., Princeton (1965)
41. Schechter, E.: Handbook of Analysis and Its Foundation. Academic Press, New York (1997)
42. Tarski, A.: Axiomatic and algebraic aspects of two theorems on sums of cardinals. Fundam. Math. **35**, 79–104 (1948)
43. Turinici, M.: A generalization of Brezis-Browder's ordering principle. An. Şt. Univ. "A. I. Cuza" Iaşi (S I-a: Mat) **28**, 11–16 (1982)
44. Turinici, M.: Mapping theorems via contractor directions in metrizable locally convex spaces. Bull. Acad. Pol. Sci. (Ser. Sci. Math.) **30**, 161–166 (1982)
45. Turinici, M.: Metric variants of the Brezis-Browder ordering principle. Demonstr. Math. **22**, 213–228 (1989)
46. Turinici, M.: A monotone version of the variational Ekeland's principle. An. Şt. Univ. "A. I. Cuza" Iaşi (S. I-a: Mat.) **36**, 329–352 (1990)
47. Turinici, M.: Vector extensions of the variational Ekeland's result. An. Şt. Univ. "A. I. Cuza" Iaşi (S I-a: Mat) **40**, 225–266 (1994)
48. Turinici, M.: Minimal points in product spaces. An. Şt. Univ. "Ovidius" Constanţa (Ser. Math.) **10**, 109–122 (2002)
49. Turinici, M.: Variational statements on KST-metric structures. An. Şt. Univ. "Ovidius" Constanţa (Mat.) **17**, 231–246 (2009)
50. Turinici, M.: Functional variational principles and coercivity over normed spaces. Optimization **59**, 199–222 (2010)
51. Turinici, M.: Brezis-Browder principle and dependent choice. An. Şt. Univ. "Al. I. Cuza" Iaşi (S. N.) Mat. **57**, 263–277 (2011)
52. Wolk, E.S.: On the principle of dependent choices and some forms of Zorn's lemma. Canad. Math. Bull. **26**, 365–367 (1983)
53. Zhu, J., Li, S.J.: Generalization of ordering principles and applications. J. Optim. Theory Appl. **132**, 493–507 (2007)

Brain Network Characteristics in Status Epilepticus

Ioannis Vlachos, Aaron Faith, Steven Marsh, Jamie White-James, Kostantinos Tsakalis, David M. Treiman, and Leon D. Iasemidis

1 Introduction

1.1 Status Epilepticus

Status Epilepticus (SE) is a life-threatening neurological emergency that is commonly treated at tertiary care epilepsy centers. Approximately 250,000 cases of SE occur in the USA annually [6]. Typically defined as greater than 30 min of continuous seizure activity or two or more sequential seizures without full recovery of consciousness between seizures, status epilepticus carries an overall 10–12 % morbidity rate and a further risk of significant morbidity if not arrested promptly [23]. Mortality in children and adults is minimized when SE lasts less than 1 h; however, thereafter, the odds of mortality jump dramatically to 38 % [24]. In addition, it is estimated that SE accounts for more than $4B annual healthcare costs in the USA alone.

Treatment of SE has traditionally involved intravenous administration of antiepileptic drugs (AEDs) that are used to treat chronic epilepsy. The goal of

I. Vlachos (✉) • L.D. Iasemidis
Center Biomedical Engineering and Rehabilitation Science, Louisiana Tech University, 818 Nelson Avenue, P.O. Box 10157/BMEB227, Ruston, LA 71272, USA
e-mail: ivlachos@latech.edu; leonidas@latech.edu

A. Faith
The Harrington Department of Bioengineering, Arizona State University, Tempe, AZ 85287, USA
e-mail: atfaith@asu.edu

S. Marsh • J. White-James • D.M. Treiman
Laboratory for Translational Epilepsy Research, Barrow Neurological Institute, Phoenix, AZ 85013, USA

K. Tsakalis
Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287, USA
e-mail: tsakalis@asu.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_25,

SE treatment is to stop the seizure activity as quickly as possible. Randomized controlled trials have recommended benzodiazepines (in particular, diazepam and lorazepam) as the initial treatment regimen [22]; however, only about 55 % of patients in SE are successfully controlled by this initial AED treatment [27]. A successful clinical response of SE to AED treatment is determined by observance of complete cessation of all seizure (ictal) activity in simultaneously recorded electroencephalogram (EEG). Successful cessation of ictal EEG activity typically occurs within 20 min following AED treatment. On the other hand, in SE patients who do not respond to treatment, patterns of ictal EEG activity persist or reappear within 60 min [27]. It is evident that new drugs and procedures, and new methods to monitor the effectiveness of those drugs and procedures over time, are of immediate need for the treatment of SE.

Treiman et al. [25] have described a sequence of progressive, visually discernible changes in the EEG recorded from generalized convulsive status epilepticus (GCSE) in humans as well as in experimental rat models of SE [26, 30]. Evaluation of AEDs and protocols for SE treatment in terms of the dynamics of concurrently monitored EEG may lead to the design of new, more effective treatment paradigms for successfully controlling SE. Such monitoring techniques may have a profound effect in the treatment of SE in the Emergency Department (ED) and Intensive Care Unit (ICU), where AEDs are given in rapid succession in the hope of patient recovery.

In the past, we analyzed the EEG in SE using measures of nonlinear dynamics and showed that successful administration of AEDs disentrains the pathologically entrained brain network dynamics and correlates well with patient recovery [8, 11]. In the study we herein describe, we employed directional measures of connectivity in terms of information flow between brain sites in the frequency domain and, with the use of graph theoretical indices, we investigated global properties of the network of the brain in SE. The general concepts under consideration were: (a) Brain network connectivity (NC) is denser during than before or after SE, and (b) Balance of information inflow and outflow (NIODc) at brain sites is lower during than before or after SE.

1.2 Frequency-Based Connectivity Measures and Graph Theory

Measures of connectivity estimated in the frequency domain are powerful tools that can provide robust estimates of the frequency interactions between components in multi-component systems [7]. Although different measures of connectivity are inherently related, their properties and capabilities vary. Some measures are able to distinguish between directional (causal) and nondirectional (coupled/correlated) interactions, some can capture only direct, and others both direct and indirect interactions. The ability of measures of connectivity to capture the interactions between a system's components at different frequencies makes them ideal for analysis of biological signals like the EEG that can exhibit different behavior in different frequency bands.

Measures of connectivity like Coherence [19], Partial Coherence [15], Partial Directed Coherence [1], and Directed Transfer Function [13] have been widely employed to study the dynamics of the human brain. Applications include epileptogenic focus localization [9, 10, 17], sleep stage analysis [14], cognition [3], and reflex (photosensitive) epilepsy [28].

In general, the brain can be treated as a network of bi-directionally connected nodes, each node corresponding to a recording brain site. Ideas and notions from graph theory have recently found use in the study of the brain network. In [31], node centrality was used to identify brain states during seizure progression, and in [4], clustering index, average path length, and weight dispersion were calculated to characterize brain network organization and function and examine their changes during normal development in children. In a recent review paper by Bullmore and Sporns [5], different graph theoretical approaches to investigate complex brain networks from diverse experimental modalities (e.g., structural and functional MRI, diffusion tensor imaging, magnetoencephalography, and electroencephalography) in humans were discussed. The authors concluded that "the emerging field of complex brain networks raises a number of interesting questions" and that "the power and elegance of graph theoretical analysis suggests that this approach will play an increasingly important part in our efforts to comprehend the physics of the brain's connectome."

2 Materials and Methods

2.1 Data

Intracranial EEG data were recorded from rats that were induced into SE and survived via timely and successful administration of AEDs at the rat epilepsy monitoring unit in the Laboratory for Translational Epilepsy Research at Barrow Neurological Institute, Phoenix, AZ. Three male Sprague-Dawley rats, weighing approximately 375 g each, were implanted with depth wire electrodes for continuous EEG recording. Rats were continuously monitored over days using an electrocorticography/field potential recording machine (XLTEK EEG; Natus Inc.). The analog EEG was band-pass filtered (0.1–100 Hz) and subsequently sampled at 256 Hz and digitally filtered by a 60 Hz notch filter. Following a 4-day resting period and a 3-day baseline EEG recording the animals were induced into status epilepticus by an intraperitoneal (IP) injection of lithium chloride (3 mmol/kg) followed by subcutaneous (SC) injection of pilocarpine (30 mg/kg) 20–24 h later. The EEG of each rat was monitored visually for electrographic signs of SE. At the onset of SE (approximately 2 h after pilocarpine injection), a cocktail of diazepam 10 mg/kg and phenobarbital 25 mg/kg was IP administered to treat SE. The total length of each EEG recording per rat was approximately 150 h.

2.2 Measure of Connectivity Between Nodes: GPDC

The GPDC [2] measure of connectivity is a normalized variant of the traditional Partial Directed Coherence. It is scale invariant and has been applied to the study of biological signals [2, 7]. We have recently utilized GPDC in epilepsy to successfully localize the epileptogenic focus from interictal EEG and MEG recordings [16, 29].

Let $\mathbf{X}(t) = \big(X_1(t), \ldots, X_n(t)\big)'$ be an n-dimensional time series vector representation of recorded EEG signals at n brain sites, with each vector component $X_i(t)$ denoting the EEG signal recorded at the ith recording site. A vector autoregressive model VAR(p) of order p [18] for $\mathbf{X}$ can be constructed as:

$$\mathbf{X}(t) = \sum_{\tau=1}^{p} \mathbf{A}(\tau)\mathbf{X}(t-\tau) + \mathbf{e}(t),$$

where $\mathbf{A}(\tau)$ are the $n \times n$ coefficient matrices of the model, and $\mathbf{e}(t)$ are the residuals that ideally follow a multivariate Gaussian white noise process. The coefficient matrices can be estimated by OLS (Ordinary Least Squares) or another related approach. The order of the model p can be estimated by traditional order selection procedures (e.g., Akaike Information Criterion).

The GPDC measures the direct effect of the component process j on i at frequency f. It is defined as:

$$G_{j\to i}(f) = \frac{|B_{ij}(f)|/\sigma_{ii}}{\sqrt{\sum_{k=1}^{n} |B_{kj}(f)|^2/\sigma_{kk}^2}},$$

where σ_{ii} are the diagonal elements of the covariance matrix $S = [\sigma_{ij}]_{i,j=1,\ldots,n}$ of the noise process $\mathbf{e}(t)$, $B_{ij}(f)$ is the (i,j)th element of the matrix $\mathbf{B}(f) = \mathbf{I} - \sum_{\tau=1}^{p} \mathbf{A}(\tau)e^{-i2\pi f\tau}$, and $\mathbf{I}$ is the $n \times n$ identity matrix.

GPDC provides a measure for the direct linear influence of process X_j on X_i at frequency f, relative to the total influence X_j has on all the other processes of the system. The average GPDC over a given frequency range (f_1, f_2) Hz is estimated and denoted by $\overline{G}_{j\to i}(f_1, f_2)$. This quantity is the "directional connectivity index" from node "j" to "i."

Finally, taking an epoch of T seconds, we obtain an $n \times n$ connectivity matrix $C[i,j]$ with elements the individual indices $\overline{G}_{j\to i}$ that represent the connectivity structure of the brain in this epoch.

2.3 Network Information Measures: Density and Node Degree Correlation

A network represented by the connectivity matrix C (see section above) has some unique properties. It is a fully connected network (complete graph), with weighted, bi-directional connections, so certain traditional graph theoretical measures are not directly applicable. We define two measures that are appropriate for these kinds of network representation: density and node degree correlation.

Network information Density (ND). In classical graph theory, the density of a network is defined as the fraction of existing connections to possible connections between nodes. In our case we define the network information density as the normalized sum of all weights in the connectivity matrix C, i.e.:

$$d(C) = \frac{\sqrt{n-1}}{n(n-1)} \sum_{j=1,i=1,j\neq i}^{n} \overline{G}_{j\to i}(f_1, f_2),$$

where the term $\frac{\sqrt{n-1}}{n(n-1)}$ is the normalization that makes the density $d(C)$ take values in the interval $[0,1]$, due to the property of GPDC that $\sum_{i=1}^{n} G^2_{j\to i}(f) = 1$.

Network Information Node In-Out Degree correlation (NIODc) [20]. In graph theory, the NIODc is the Pearsons correlation coefficient between in-degree and out-degree of a node in a directed network. Nodes with large in-degree are hubs, nodes with large out-degree are authorities. Calculation of the correlation between in-degree and out-degree is a way to check whether hubs are also authorities or not. In our case, we replace the in-degrees and out-degrees with the total flow into a node from all other nodes (Inflow) and total flow from the node to all other nodes (Outflow), respectively. The inflow of a node/site i is estimated by summation over all partial flows towards i from the rest of the nodes j as

$$\mathrm{In}F_i = \sum_{j=1,j\neq i}^{n} \overline{G}_{j\to i}(f_1, f_2).$$

The outflow $\mathrm{Out}F_i$ is defined similarly as the sum of all outflows originating from i. Finally, the Node In-Out Degree correlation (NIODc) index for the connectivity matrix C is defined as

$$r(C) = \frac{\mathrm{Cov}(\mathrm{In}F, \mathrm{Out}F)}{\sqrt{\mathrm{Var}(\mathrm{In}F)\mathrm{Var}(\mathrm{Out}F)}},$$

where Cov denotes the covariance and Var the variance of inflows ($\mathrm{In}F$) and outflows ($\mathrm{Out}F$).

The estimated brain networks during interictal (seizure-free) and ictal (seizure) periods by analysis of a 7-channel intracranial EEG recording from an epileptic rat is shown in Fig. 1. The color plots show the average GPDC values over the 0.1–30 Hz

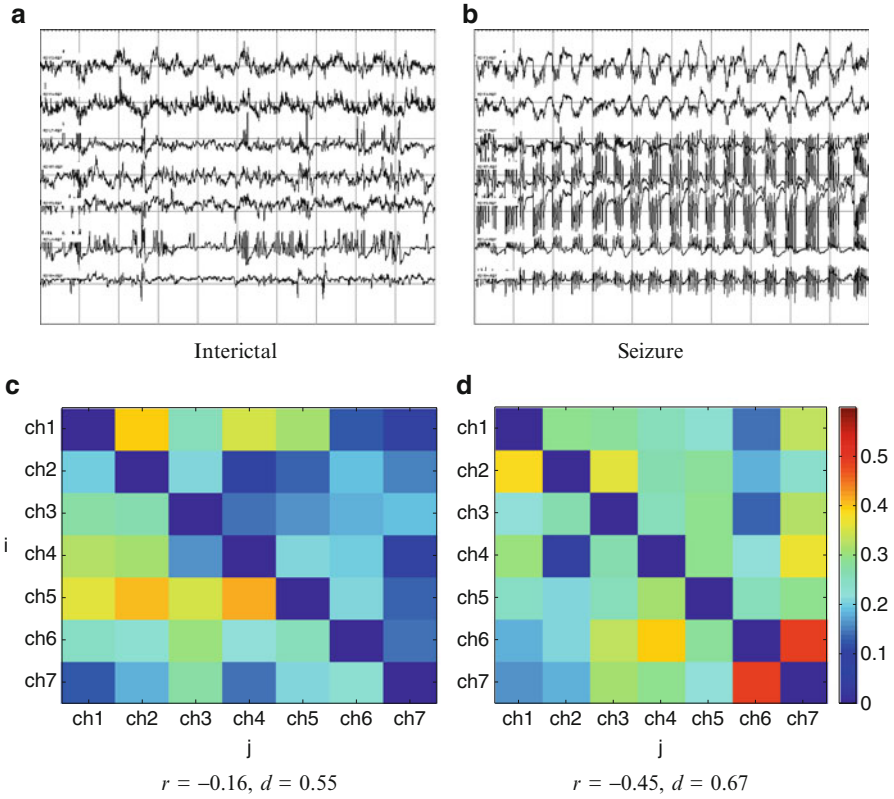


Fig. 1 (**a**) Interictal EEG. (**b**) Ictal (during a seizure) EEG. (**c**) and (**d**) The average GPDC (connectivity matrix C) over 0.1–30 Hz estimated interictally and ictally, respectively. The values for the ND and NIODc measures of the networks are shown below the corresponding sub-figures in (**c**) and (**d**)

band ($\overline{G}_{j\to i}(0.1,30)$). We note that there are stronger (less blue color) connections between brain sites during the seizure, the network density (ND) values $d(C)$ are a bit larger, and the NIODc values $r(C)$ are a lot smaller than interictal ones.

3 Results

The EEG per rat was divided into successive non-overlapping epochs of $T = 10$ seconds in duration and the GPDC values were estimated within each epoch with $p = 7$ and $n = 4$. The value of $p = 7$ is heuristically selected based on our nonlinear dynamical analysis of EEG in the past [12]. $\overline{G}_{j\to i}(f_1, f_2)$ was computed over the traditional rat EEG bands, delta $(0,4)$ Hz, theta $(4,12)$ Hz, beta $(12,30)$ Hz, and

gamma $(30, 100)$ Hz [21]. The ND and NIODc measures of the brain's network were then estimated from the connectivity matrix for each 10 s epoch over the full period of recording (150 h per rat).

Figure 2 shows the network density (ND) for all three rats and all four frequency bands. The first vertical black line corresponds to the time of pilocarpine administration to induce SE, while the second one to the time of AEDs' administration. The profiles of the network density (ND) over time are relatively consistent across rats and frequency bands: almost immediately after the pilocarpine injection, there is a noticeable increase of the network density across all frequency bands that persists for a long period, even after the administration of AEDs. Over time, ND decreases and stabilizes at values a bit higher than its pre-SE induction (baseline) values. This is maximally observed at the higher (beta and gamma) frequency bands. Two additional interesting observations are that (a) the network density appears to be generally inversely related to frequency (higher density at lower frequencies and lower density at higher frequencies) and (b) stabilization of ND values after onset of SE at different rates for different frequency bands.

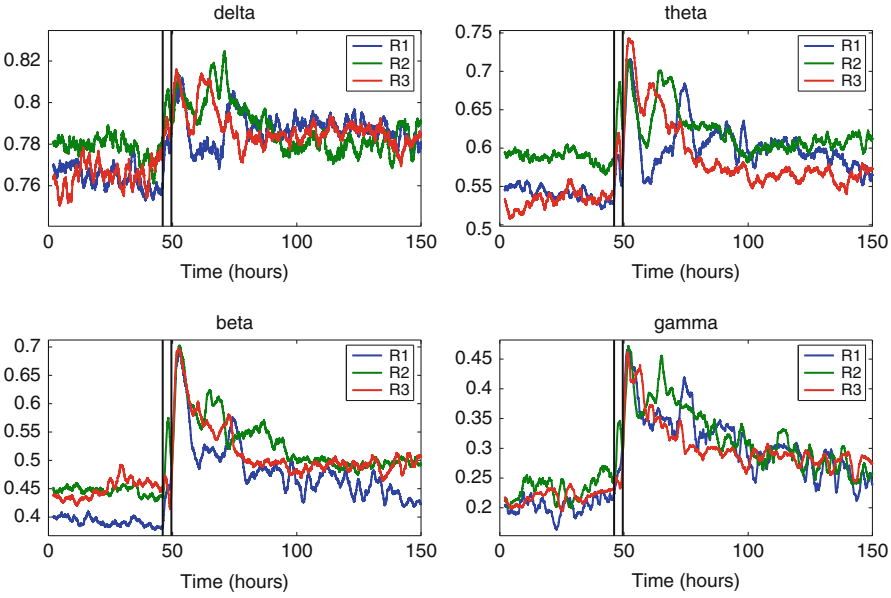


Fig. 2 Network density (ND) over time per frequency band for each of the three SE induced rats (R1, R2, and R3). In each sub-figure, ND is estimated from a different frequency band denoted on the figure's title. The *two black vertical lines* correspond to the times of administration of pilocarpine (to induce SE) and AEDs (to recover from SE), respectively

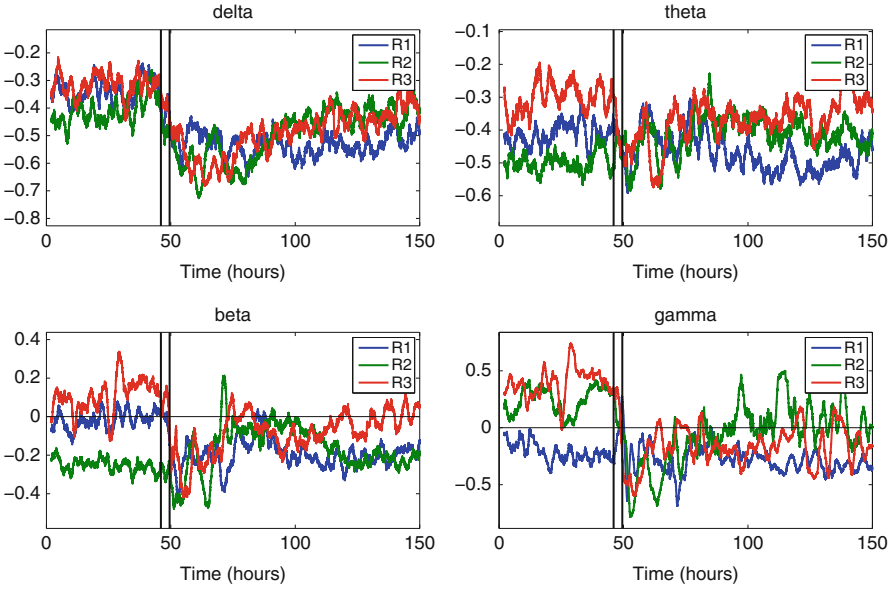


Fig. 3 Network Node In-Out Degree correlation (NIODc) over time per frequency band for all three rats (R1, R2, and R3). *Vertical lines* as in Fig. 2

The NIODc profile of the brain network is shown in Fig. 3. There is a clear overall decrease in NIODc after SE induction with a slow trend of recovery towards baseline (pre-SE) values thereafter, which is more clearly visible in the lowest frequency (delta) band. The results from this network measure (correlation of information inflow and outflow at the recording sites) are not as consistent across rats and frequency bands as the ones of network density. This may be due to NIODc being more sensitive to inflicted and possibly remaining damage after SE in rats' brain network.

4 Conclusions

In an attempt to better characterize the brain's effective network during the transition into and out of SE, we first quantified the directional connectivity between brain sites by multivariate autoregressive analysis of long-term recorded EEGs in an animal model of SE. We then applied measures from graph theory to quantify the global characteristics of the resulted brain network over time. We found that SE was consistently associated with increased information flow between brain regions, quantified by network density (ND), and with decreased balance between inflow and outflow at brain sites, quantified by NIODc. These changes persist long after treatment of SE by AEDs and are more consistent across rats in specific

frequency bands: the highest band (gamma) for ND and the lowest band (delta) for NIODc. The above results suggest that the proposed methodology and measures that combine quantification of information flow with graph theory may assist in the study and monitoring of SE progression, as well as in the evaluation of AEDs' effectiveness in the life-threatening condition of status epilepticus. Studies of a larger scale in both SE animal models and humans with SE episodes to further validate these preliminary results are in progress.

Acknowledgment This work was supported by the National Science Foundation grant ECCS-1102390.

References

1. Baccala, L.A., Sameshima, K.: Partial directed coherence: a new conception in neural structure determination. Biol. Cybern. **84**(6), 463–474 (2001)
2. Baccala, L.A., Takahashi, D.Y., Sameshima, K.: Generalized partial directed coherence. In: Proceedings of the 15th International Conference on Digital Signal Processing, pp. 162–166 (2007)
3. Blinowska, K.J., Kus, R., Kaminski, M., Janiszewska, J.: Transmission of brain activity during cognitive task. Brain Topogr. **23**(2), 205–213 (2010)
4. Boersma, M., Smit, D.J., de Bie, H.M., Van Baal, G.C., Boomsma, D.I., de Geus, E.J., Delemarre-van de Waal, H.A., Stam, C.J.: Network analysis of resting state EEG in the developing young brain: structure comes with maturation. Hum. Brain Mapp. **32**(3), 413–425 (2011)
5. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. **10**(3), 186–198 (2009)
6. Epilepsy Foundation of America: Treatment of convulsive status epilepticus. J. Am. Med. Assoc. **270**, 854–859 (1993)
7. Faes, L., Erla, S., Nollo, G.: Measuring Connectivity in Linear Multivariate Processes: Definitions, Interpretation, and Practical Analysis. Comp. Math. Methods Med. Special issue on "Methodological Advances in Brain Connectivity", 140513 (2012)
8. Faith, A., Sabesan, S., Wang, N., Treiman, D.M., Sirven, J.I., Tsakalis, K., Iasemidis, L.D.: Dynamical analysis of the EEG and treatment of human status epilepticus by antiepileptic drugs. In: Chaovalitwongse, W., Pardalos, P.M., Xanthopoulos, P. (eds.) Computational Neuroscience, Springer Optimization and Its Applications, vol. 38, pp. 305–315. Springer, New York (2010)
9. Franaszczuk, P.J., Bergey, G.K., Kamiski M.: Analysis of mesial temporal seizure onset and propagation using the directed transfer function method. Electroencephalogr. Clin. Neurophysiol. **91**(6), 413–427 (1994)
10. Gersch, W.,Goddard, G.V.: Epileptic focus location: spectral analysis method. Science **169**(3946), 701–702 (1970)
11. Good, L.B., Sabesan, S., Iasemidis, L.D., Tsakalis, K., Treiman, D.M.: Brain dynamical disentrainment by anti-epileptic drugs in rat and human status epilepticus. In: Engineering in Medicine and Biology Society, IEMBS 2004. 26th Annual International Conference of the IEEE, pp. 176–179 (2004)
12. Iasemidis, L.D., Principe, J.C., Sackellares, J.C.: Measurement and quantification of spatiotemporal dynamics of human epileptic seizures. In: Akay, M. (ed.) Nonlinear Biomedical Signal Processing, vol. 2, pp. 294–318. Wiley-IEEE Press, New York (2000)

13. Kaminski, M., Blinowska, K.J.: A new method of the description of the information flow in the brain structures. Biol.Cybern. **65**(3), 203–210 (1991)
14. Kaminski, M., Blinowska, K.J., Szelenberger, W.: Topographic analysis of coherence and propagation of EEG activity during sleep and wakefulness. Electroencephalogr. Clin. Neurophysiol. **102**(3), 216–227 (1997)
15. Kay, S.: Modern Spectral Estimation. Theory & Application. Prentice-Hall, Englewood Cliffs (1988)
16. Krishnan, B., Vlachos, I., Wang, Z.I., Mosher, J., Iasemidis, L., Burgess, R., Alexopoulos, A.V.: Advanced MEG source analysis for epileptogenic focus localization in patients with non-lesional MRI. In: Proceedings of the 29th Southern Biomedical Engineering Conference, Miami (2013)
17. Lu, Y., Yang, L., Worrell, G.A., He, B.: Seizure source imaging by means of FINE spatio-temporal dipole localization and directed transfer function in partial epilepsy patients. Clin. Neurophysiol. **123**(7), 1275–1283 (2012)
18. Lutkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer, New York (2005)
19. Marple, S.L.: Digital Spectral Analysis with Applications. Prentice-Hall, San Diego (1987)
20. Moslonka–Lefebvre, M., Pautasso, M., Jeger, M.J.: Disease spread in small-size directed networks: epidemic threshold, correlation between links to and from nodes, and clustering. J. Theor. Biol. **260**(3), 402–411 (2009)
21. Nerad, L., Bilkey, D.K.: Ten- to 12–Hz EEG oscillation in the rat hippocampus and rhinal cortex that is modulated by environmental familiarity. J. Neurophysiol. **93**(3), 1246–1254 (2005)
22. Sirven, J.I., Waterhouse, E.: Management of status epilepticus. Am. Fam. Physician. **68**(3), 469–476 (2003)
23. Treiman, D.M., Delgado-Escueta, A.V.: Status epilepticus. In: Thompson, R.A., Green, J.R. (eds.) Critical Care of Neurological Emergencies, pp. 55–99. Raven Press, New York (1980)
24. Treiman, D., Walker, M.: Treatment of seizure emergencies: convulsive and non-convulsive status epilepticus. Epilepsy Res. **68**(S1), 77–82 (2006)
25. Treiman, D.M., Walton, N.Y., Kendrick, C.: A progressive sequence of electroencephalographic changes during generalized convulsive status epilepticus. Epilepsy Res. **5**(1), 49–60 (1990)
26. Treiman, D.M., Meyers, P.D., DVA Status Epilepticus Cooperative Study Group: Utility of the EEG pattern as a predictor of success in the treatment of generalized convulsive status epilepticus. Epilepsia **32**(S3), 93 (1991)
27. Treiman, D.M., Meyers, P.D., Walton, N.Y., Collins, J.F., Colling C., Rowan, A.J., Handforth, A., Faught, E., Calabrese, V.P., Uthman, B.M., Ramsay, R.E., Mamdani M.B.: A comparison of four treatments for generalized convulsive status epilepticus. New England J. Med. **339**(12), 792–798 (1998)
28. Varotto, G., Visani, E., Canafoglia, L., Franceschetti, S., Avanzini, G., Panzica, F.: Enhanced frontocentral EEG connectivity in photosensitive generalized epilepsies: a partial directed coherence study. Epilepsia **53**, 359–367 (2012)
29. Vlachos, I., Krishnan, B., Sirven, J., Noe, K., Drazkowski, J., Iasemidis, L.: Frequency-based connectivity analysis of interictal iEEG to localize the epileptogenic focus. In: Proceedings of the 29th Southern Biomedical Engineering Conference, Miami (2013)
30. Walton, N.Y., Treiman, D.M.: Response of status epilepticus induced by lithium and pilocarpine to treatment with diazepam. Exp. Neurol. **101**, 267–275 (1988)
31. Yaffe, R., Burns, S., Gale, J., Park, H.J., Bulacio, J., Gonzalez-Martinez, J., Sarma, S.V.: Brain state evolution during seizure and under anesthesia: a network-based analysis of stereotaxic EEG activity in drug-resistant epilepsy patients. In: Proceedings of the 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), San Diego, 5158–5161 (2012)

A Review on Consensus Clustering Methods

Petros Xanthopoulos

1 Introduction

Unsupervised learning, or clustering, is one of the very fundamental exploratory data analysis methodologies with application in virtually any research area that involves categorization or grouping of data [23]. Modern applications call for ability to analyze massive amounts of data in an efficient time frame [1, 38]. In this sense *unsupervised learning* plays a significant role in data summarization and preliminary structure identification of complex, and often heterogeneous, datasets. Loosely speaking clustering is defined as the process of grouping similar objects (data samples) together based on some basic common similarity properties. This is usually achieved through optimization (maximization or minimization) of a *similarity* related function of interest. This general definition allows many interpretations and sub-definitions as to what constitutes "a good clustering" [11]. This ambiguity is responsible for a richness of algorithms and practical challenges as well. Some notable algorithmic contributions so far include, but are not limited to, the k-means algorithm [30], hierarchical clustering [31, 49], distribution-based model such as expectation maximization [37], spectral clustering [45] and density-based clustering [25]. For a comprehensive literature review of clustering methodology and its applications we refer the reader to [7, 23, 33, 50].

Clustering results on a single problem can vary due to a number of factors. The most important factors that are responsible for this variability are: (1) variability due to local optimality, (2) variability due to algorithm, and (3) variability due to data.

1. *Variability due to local optimality:* In most cases, optimization of clustering objective functions require solution of an NP-complete problem, making

P. Xanthopoulos (✉)
Industrial Engineering and Management Systems Department, University of Central Florida, 4000 Central Florida Blvd., P.O. BOX 16093, Orlando, FL 32816–2993, USA
e-mail: petrosx@ucf.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_26,

heuristic approaches are very common in the literature. Such algorithms usually terminate after finding different locally optimal solutions which can differ for multiple runs of the same algorithm.
2. *Variability due to algorithm:* Since the objectives of each clustering algorithm are different, it is expected to have different clustering results for different algorithms.
3. *Variability due to data:* Sometimes it is possible to have different datasets describing exactly the same objects. Such examples include two different images of the same object under different illumination conditions and/or angles or same users subscribed to different digital content services (e.g., Amazon, Netflix). In such situations it is possible that clustering results will be different even for the same objects.

These inconsistencies motivate the need for *ensemble* algorithms. These algorithms try to combine various clusterings into a single robust clustering of superior quality. The task of combining different clusterings is as computationally challenging as clustering itself; however, research in this area has provided data mining community with more robust algorithms. It can be shown that *consensus clustering* is equivalent to the *median partitioning* problem which is known to be NP-complete [26]. Although in robust data mining traditionally one has to deal with uncertainty induced by measurement or implementation constraints [60, 61], in consensus clustering ambiguity emanates from the choice of clustering algorithm as well. In general there are some exact and a lot of heuristic approaches for consensus clustering. Some interesting theoretical results provide a connection between consensus clustering and the nonnegative matrix factorization (NNMF) problem [29] whereas recently the problem was casted as a network clustering problem [27].

At this moment, we need to note that in the literature the terms *unsupervised ensemble learning*, *consensus clustering*, and *aggregation of clusterings* are used to denote the same process. Here we will use the term *consensus clustering*. In the present chapter we provide a literature overview of the consensus clustering grouped based on their basic principles and background theory.

In this paper we use uppercase letters to denote matrices (e.g., A) and lowercase to denote matrix elements. For example with a_{ij} we denote the element of matrix A that belongs to the ith row and jth column. The vectors are column vectors unless denoted otherwise. With $tr(\cdot)$ we denote the trace function of a square matrix defined by $tr(A) = \sum_{i=1}^{n} a_{ii}$. The input to any clustering algorithm can be described by a set of ordered samples $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and a similarity function $d(.,.)$ that maps a pair of samples to a real number. The output of a clustering algorithm is a set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$, $C_i \subseteq \mathcal{S}$, with the property $\cup_{i=1}^{m} C_i = \mathcal{S}$ where m is a parameter denoting the number of clusters that exist in the dataset and it is either tuned by the user or internally during the clustering process. Moreover if $C_i \cap C_j = \emptyset, i \neq j$ the clustering is termed *hard clustering* and *soft* or *fuzzy clustering* otherwise. This intuitively means that each sample can potentially belong to more than one categories with partial membership. In this article we focus on hard clustering although these approaches can be trivially generalized in soft clustering framework as well.

The rest of the chapter is structured as follows: In Sect. 2 we formulate the consensus clustering problem. In Sect. 3 we describe the exact algorithms that have been proposed to the literature. In Sect. 4 we describe the approximation algorithms whereas in Sect. 5 we provide the relation to other problems such as NNMF and network partitioning problem. In Sect. 6 we provide an overview of the most emerging applications of consensus clustering and in Sect. 7 we describe some of the existing software packages available for this problem. In the last section we provide discussion and some challenges that could be potential direction for future research.

2 The Consensus Clustering Problem

Given a set of clustering $SC = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ and a symmetric distance measure $d(.,.)$ between two clusterings we want to find a clustering $\mathcal{C}^*$ such that:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \sum_{p=1}^{k} d(\mathcal{C}, \mathcal{C}_p) \tag{1}$$

We call $\mathcal{C}^*$ the consensus clustering. This problem is known as the *median partition problem*. The commutation nature of the problem heavily depends on the distance measure $d(.,.)$. Next we discuss the most common function choices in the literature.

2.1 Clustering Distance Functions

One important concept in consensus clustering is the choice of the distance measures between clusterings. These distance measures are usually related to the clustering agreements between two different clustering. One popular distance voice is the *symmetric difference distance* (sdd). Let us define α as the pair of samples that have been clustered in the same cluster in both clusterings, and β as the samples clustered in different clusters in both clusterings. Then sdd is given by:

$$d(\mathcal{C}_i, \mathcal{C}_j) = \binom{n}{2} - \alpha - \beta. \tag{2}$$

Clearly, when two clusterings are identical $\binom{n}{2} = \alpha + \beta$ and $d(\mathcal{C}_i, \mathcal{C}_j) = 0$. In any case distance measures need to be label independent. For example for the clusterings

$$\mathcal{C}_1 = \{1,1,1,0,0,0\}, \quad \mathcal{C}_2 = \{0,0,0,1,1,1\}$$

the distance $d(\mathcal{C}_1, \mathcal{C}_2)$ is zero. It can be shown that this distance can be computed in linear time [12]. As noted in [56] the choice of the similarity function determines

the complexity of the consensus clustering problem. For example, one can construct similarity measures for which the consensus clustering problem is polynomially solvable, however usually such functions are not interesting from a practical perspective. For example, if we consider the following distance function:

$$d(\mathcal{C}_i, \mathcal{C}_j) = \begin{cases} 1, \text{if } \mathcal{C}_i = \mathcal{C}_j \\ 0, \text{otherwise} \end{cases} \tag{3}$$

we can determine the median clustering in $O(1)$ time. On the other hand there can be distance measures for which heuristic optimization techniques do not run in acceptable computational time. Some other categories of distance measures that have been proposed over time in the literature are the following [56]: (1) pair counting measures [5, 14, 34, 39], (2) set matching measures [10, 53, 65], (3) information theory measures [3, 32, 35, 40, 51], and (4) kernel-based measures [48, 54, 55].

3 Exact Approaches

The consensus clustering problem can be formulated as a 0–1 linear program. The resulting polyhedron is exactly the same with the one obtained for the well-known clique partitioning problem. Based on this observation a cutting plane algorithm was proposed for solving this problem [21]. This exact method improved over the previous exact formulation and is able to handle instance of order several hundred data points.

For every clustering $\mathcal{C}_p$, $p = 1, \ldots, k$ we define:

$$r_{ij}^{(p)} = \begin{cases} 1, & (i,j) \in \mathcal{C}_p \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $(i,j) \in \mathcal{C}_p$ denotes that samples s_i and s_j belong to the same cluster in clustering $\mathcal{C}_p$. Also define the decision variables r_{ij}

$$r_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

Then the objective function of Eq. (1) can be written as:

$$\sum_{p=1}^{k} d(\mathcal{C}, \mathcal{C}_p) = \sum_{p=1}^{k} \sum_{i,j} (r_{ij}^{(p)} - r_{ij})^2 \tag{6}$$

Since $r_{ij}^{(p)}, r_{ij} \in \{0,1\}$ Eq. (6) can be linearized

$$\sum_{p=1}^{k} \sum_{i,j} (r_{ij}^{(p)} - r_{ij})^2 = \sum_{p} \sum_{i,j} \left(r_{ij}^{(p)} - 2 r_{ij}^{(p)} \cdot r_{ij} + r_{ij} \right) \tag{7a}$$

$$= \sum_{p} \sum_{i,j} r_{ij}^{(p)} + \sum_{p} \sum_{i,j} \left(1 - 2 \cdot r_{ij}^{(p)} \right) r_{ij} \tag{7b}$$

From the last equation we can observe that the objective is a 0–1 linear. In fact it can be written as

$$c + \sum_{i,j} c_{ij} r_{ij} \tag{8}$$

where

$$c = \sum_{p} \sum_{i,j} r_{ij}^{(p)}, \quad c_{ij} = \sum_{p} \left(1 - 2 \cdot r_{ij}^{(p)}\right) \tag{9}$$

Since variables r_{ij} refer to the clustering task they need to have the properties of reflexiveness, symmetricity, and transitivity. Moreover the term c can be dropped from the objective function since it is constant. Overall the optimization problem can be formulated as

$$\begin{aligned}
\min \quad & \sum_{ij} c_{ij} r_{ij} && && \text{(10a)}\\
\text{s.t.} \quad & r_{ii} = 1, && i = 1,\ldots,n && \text{(10b)}\\
& r_{ij} = r_{ji}, && i,j = 1,\ldots,n && \text{(10c)}\\
& r_{ij} + r_{jk} - r_{ik} \leq 1, && i,j,k = 1,\ldots,n && \text{(10d)}\\
& r_{ij} \in \{0,1\}, && i,j = 1,\ldots,n && \text{(10e)}
\end{aligned}$$

where the reflexive property is implied by constraints (10b), the symmetric property by constraint (10c), and the transitive property by constraint (10d). The size of the problem can be reduced by considering its symmetric nature. Since $r_{ij} = r_{ji}$ these variables can be replaced by a new variable x_{ij}. Also since variables r_{ii} are fixed they can be dropped. Accordingly we can define weights as $w_{ij} = c_{ij} + c_{ji}$. Then the problem can be rewritten in terms of the transformed variables:

$$\begin{aligned}
\min \quad & \sum_{1 \leq i < j \leq n} w_{ij} \cdot x_{ij} && && \text{(11a)}\\
\text{s.t.} \quad & x_{ij} + x_{jk} - x_{ik} \leq 1, && 1 \leq i < j < k \leq n && \text{(11b)}\\
& x_{ij} - x_{jk} + x_{ik} \leq 1, && 1 \leq i < j < k \leq n && \text{(11c)}\\
& -x_{ij} + x_{jk} + x_{ik} \leq 1, && 1 \leq i < j < k \leq n && \text{(11d)}\\
& x_{ij} \in \{0,1\}, && 1 \leq i < j \leq n && \text{(11e)}
\end{aligned}$$

The polyhedron of the last problem is the same with this of clique partitioning problem. Although the last problem is also NP-complete one can use the theoretical results and exact approaches derived for clique partitioning to establish a more efficient algorithm for consensus clustering. Such an algorithm is described in [21]. Despite the theoretical interests of this approach its practical limitations have made it a less favorable option in consensus clustering literature. Recent developments in this field are still limited to solving instances no larger than 300 data samples [52].

4 Approximation Algorithms

In the literature there are several approximation algorithm approaches for consensus clustering. These algorithms run in polynomial time and obtain a solution which is equal to the optimal multiplied with a constant, known as the approximation factor. It can be shown that even a simple naive algorithm can obtain a solution which is guaranteed to be close to the optimal up to a multiplication factor. Some basic algorithms for which there exist approximation results are the following:

- **Pick-a-Cluster**: Simple algorithm that just chooses a cluster randomly and returns as a solution.
- **Best-Clustering**: Choose a clustering with the highest objective function value as the solution.

Most of the approximation algorithms are directly derived by adjusting the existing ones for the correlation clustering problem [4]. These approaches include the following two algorithms:

- **CC-Pivot:** This algorithm resembles the quicksort number ordering routine. The main idea is to choose a pivot element and partitions the rest of the elements according to their relation with the pivot element.
- **CCLP-Pivot:** This is the linear programming version of CC-Pivot and it is an adaptation of the linear programming approach first introduced in [9] for the correlation clustering problem.

Other approximation algorithms for consensus clustering were proposed in [18]:

- *Average linkage:* This is based on the standard agglomerative clustering principle. At the beginning each element belongs to its own separate cluster. At each step the algorithm merges two clusters with the closest distance into one cluster. The process is repeated until the average sdd between clusters is 1/2.
- *Furthest:* This is another greedy approach which is based on an approximation algorithm originally proposed for the k-center problem in [19]. This algorithm's running time depends on the number of output clusters.

In Table 1 we summarize the known approximation results for these algorithms with their known complexity times.

As noted in [6], approximation algorithms for consensus clustering, although they come with a theoretical guarantee, most of the time are not practically useful due to their computational time (as a $O(n^2)$ algorithm becomes impractical for large datasets), and their practical performance is often comparable with heuristics without any theoretical performance guarantee.

5 Relation to Other Problems

So far we have already pointed out the relation between consensus clustering and the clique partitioning problem. In addition, the relation of the consensus clustering problem to the correlation clustering problem has provided most of its existing

Table 1 The summary of existing approximation results and the corresponding computational complexities for consensus clustering, where m stands for the number of clusterings, n for the number of data samples, and k for the resulting number of clusters

Name	Complexity	Approximation factor	Reference
Best clustering	$O(m^2n)$	2	[2]
Pick a cluster	$O(1)$	2	[2]
CC-PIVOT	$O(kmn)$	11/7	[2]
CCLP-PIVOT	$O(n^8)$	4/3	[2, 9]
Average linkage	$O(n^2(\log n+m))$	2	[18]
Furthest	$O(kmn)$	2	[18]

approximation algorithms. In this section we will discuss other equivalency results that have been reported in the literature. These results include the formulation of consensus clustering as a NNMF problem as well as some recent algorithms that solve consensus clustering through spectral graph clustering approaches. We conclude this section with the semidefinite programming relaxation of the problem.

5.1 Nonnegative Matrix Factorization

It has been shown that the consensus clustering problem can be solved through a NNMF problem. The corresponding formulation was originally proposed by [29] and then extended for the weighted consensus clustering problem in [28]. The NNMF problem consists of finding two matrices $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{k\times m}$ whose product is approximately equal to a given matrix $M \in \mathbb{R}^{m\times n}$ with the additional constraint that both A and B are positive element-wise. This can be written as an optimization problem as follows:

$$\min_{A,B} \ \|M - A\cdot B\|_F^2 \tag{12a}$$

$$s.t. \ \ a_{ij} > 0, \qquad i = 1,\ldots,n, j = 1,\ldots,k \tag{12b}$$

$$b_{ij} > 0, \qquad i = 1,\ldots,k, j = 1,\ldots,m \tag{12c}$$

where $\|A\|_F = \sqrt{tr(A^TA)} = \sqrt{\sum_{i,j}|a_{ij}|^2}$ the *Frobenius* norm of matrix A. Using the same variables defined in Eqs. (4) and (5) the problem minimization problem can be written in matrix notation:

$$\min \ \sum_{p=1}^{k} \|R - R^{(p)}\|_F^2 \tag{13}$$

If we define the matrix with the average distance from all clusterings as follows:

$$\tilde{R} = \frac{1}{k}\sum_{i=1}^{k} R^{(p)} \tag{14}$$

we can also write:

$$\frac{1}{k}\sum_{p=1}^{k} \|R - R^{(p)}\|_F^2 = \frac{1}{k}\sum_{p=1}^{k} \|R - \tilde{R} + \tilde{R} - R^{(p)}\|_F^2 \tag{15a}$$

$$= \Delta R^2 + \|R - \tilde{R}\|^2 \tag{15b}$$

where ΔR is the average square difference defined by

$$\Delta R^2 = \frac{1}{k}\sum_{p=1}^{k}\sum_{i,j}\left(R^{(p)} - \tilde{R}\right)^2 \tag{16}$$

and it is constant. Constraints described by (10d) can now be replaced by the NNMF constraints. Then the problem can be described by

$$\min_{H \geq 0} \|\tilde{R} - H^T H\|^2 \tag{17}$$

Solution to problem (17) can be obtained with any from the NNMF algorithms [24].

5.1.1 Semidefinite Programming Relaxation

Note that the problem (17) can be relaxed to a semidefinite program as follows. The quadratic form can be expanded as:

$$\|\tilde{R} - H^T H\|^2 = \|\tilde{R}\|^2 - 2Tr(H^T \tilde{R} H) + \|H^T H\|. \tag{18}$$

Since two of the terms are constants ($\|\tilde{R}\|^2$ and $\|H^T H\|$) the problem can be reduced to:

$$\min_{H \geq 0} Tr(H^T \tilde{R} H) = \min_{H \geq 0} Tr(R H H^T) = \min_{Z \succeq 0} Tr(\tilde{R} Z) \tag{19}$$

where $\succeq$ denotes positive definiteness property. Problem (19) is a linear semidefinite program. However recovering H from Z requires the solution of another norm minimization problem.

5.2 Graph Clustering

In a recent work [27], consensus clustering was described as a spectral graph clustering problem. This approach involves the construction of an intermediate structure

termed the consensus graph. This is a weighted graph with each weighted edge (i, j) being the normalized number of clusterings that points s_i and s_j were assigned to the same cluster. Final consensus clustering is obtained by running a network clustering algorithm (usually spectral clustering) on the consensus graph. This approach was found to be robust compared to other clustering approaches [27]. In addition, it is computationally efficient since it requires only a matrix eigendecomposition. However theoretically it is still a heuristic approach with no performance guarantee.

6 Applications

Although consensus clustering can be employed for any exploratory data analysis problem that involves data grouping it has been used for applications that naturally produce massive amount of data and clustering uncertainty. Primary areas include microarray gene expression analysis and computational chemistry, whereas it has been applied in various other datasets, image segmentation data (including medical datasets), documents clustering, and co-authorship network analysis. In Table 2 we summarize some of the applications found in literature.

As noted in [18] consensus clustering naturally arises in problems where we need to cluster categorical data. In this case each categorical variable defines a clustering, and a clustering over the categorical variables can be seen as a consensus clustering. Another natural application arises when we need to cluster heterogeneous data, i.e., data from different sources and maybe formats, about the same objects (entities).

Table 2 Representative literature of consensus clustering applications

Application	Reference
Chemical structure clustering	[41–43]
Categorical data clustering	[16]
Gene expression microarray data	[36, 47, 63]
Co-authorship network clustering	[27]
Image segmentation	[13, 15, 58, 59]
Magnetic resonance imaging (MRI) clustering	[57]
Image quantization	[8]
Synthetic aperture radar (SAR) image segmentation	[64]
Document clustering	[20, 46, 62]

Due to the robustness of consensus clustering it is useful in particular for identifying the number of clusters when this is now known in advance. In addition consensus clustering is tightly related to outlier detection as well as privacy preserving clustering [18]. In the last case a common clustering has to be generated but one might not be able to directly access to all databases. In such cases individual vendors can exchange clusterings that do not include sensitive (private) information without exchanging the information itself.

7 Software

Several implementations of consensus clustering are readily available through several software packages. Table 3 summarizes some of the most notable implementations.

Table 3 Software packages with consensus clustering capabilities

Software	Environment	Reference
ConsensusCluster	Command line	[44]
clusterCons	R	[47]
Clustering ensembles (CLUE)	R	[22]
Cluster pack toolbox	Matlab	[17]

The command line tool Consensus cluster [44] supports GPU processing for more efficient computations and it is implemented in Python while it supports k-means, hierarchical clustering, self-organizing maps, and partition around medoids algorithm. The consensus clustering is performed by another clustering algorithm on the consensus graph which can be constructed by Euclidean or correlation distance metrics. The software package clusterCons has been implemented for the R platform[1] providing implementation of resampling algorithms that have appeared in the literature [36, 47]. It supports a number of clustering algorithms such as k-means, hierarchical, as well as a number of visual evaluation tools such as Receiver Operating Characteristic (ROC) curve plots. Finally the cluster pack toolbox is a clustering toolbox that includes the consensus framework as an embedded part.

[1] http://www.r-project.org.

8 Conclusion and Future Research

Consensus clustering is an important problem in exploratory data analysis with particular value for applications that involve data heterogeneity. So far despite the interesting theoretical results about the problem that include exact formulations and approximation algorithms the most practically notable approaches are heuristic based. Modern needs for massive data analysis make exact approaches impractical since they are only able to handle problems of limited size.

The interesting relationship between consensus clustering and other well-studied problems such as clique partition, correlation clustering, NNMF, and graph clustering has provided theoretical and practical tools enabling robust solutions within reasonable amount of time.

Acknowledgment The author would like to thank Dr. Sibel B. Sonuç for proofreading the manuscript and providing useful comments.

References

1. Abello, J., Pardalos, P.M., Resende, M.G.: Handbook of Massive Data Sets, vol. 4. Kluwer Academic, London (2002)
2. Ailon, N., Charikar, M., Newman, A.: Aggregating inconsistent information: ranking and clustering. J. ACM (JACM) **55**(5), 23 (2008)
3. Bakus, J., Hussin, M., Kamel, M.: A som-based document clustering using phrases. In: Proceedings of the 9th International Conference on Neural Information Processing, 2002 (ICONIP'02), vol. 5, pp. 2212–2216. IEEE, Piscataway (2002)
4. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Mach. Learn. **56**(1–3), 89–113 (2004)
5. Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pacific Symposium on Biocomputing, vol. 7, pp. 6–17 (2001)
6. Bertolacci, M., Wirth, A.: Are approximation algorithms for consensus clustering worthwhile? In: Proceedings of the 2007 SIAM International Conference on Data Mining (2007)
7. Butenko, S., Chaovalitwongse, W.A., Pardalos, P.P.M.: Clustering challenges in biological networks. World Scientific, New Jersey (2009)
8. Chang, Y., Lee, D.J., Hong, Y., Archibald, J., Liang, D.: A robust color image quantization algorithm based on knowledge reuse of k-means clustering ensemble. J. Multimedia **3**(2), 20–27 (2008)
9. Charikar, M., Guruswami, V., Wirth, A.: Clustering with qualitative information. J. Comput. Syst. Sci. **71**(3), 360–383 (2005)
10. Dongen, S.: Performance criteria for graph clustering and markov cluster experiments. CWI (Centre for Mathematics and Computer Science) Amsterdam, The Netherlands (2000)
11. Estivill-Castro, V.: Why so many clustering algorithms: a position paper. ACM SIGKDD Explorations Newsl. **4**(1), 65–75 (2002)
12. Filkov, V., Skiena, S.: Integrating microarray data by consensus clustering. Int. J. Artif. Intell. Tools **13**(04), 863–880 (2004)
13. Forestier, G., Wemmert, C., Gançarski, P.: Collaborative multi-strategical clustering for object-oriented image analysis. In: Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 71–88. Springer, Berlin (2008)

14. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983)
15. Fred, A.: Finding consistent clusters in data partitions. In: Multiple Classifier Systems, pp. 309–318. Springer, Berlin (2001)
16. Gao, C., Pedrycz, W., Miao, D.: Rough subspace-based clustering ensemble for categorical data. Soft. Comput. **17**, 1–16 (2013)
17. Ghosh, J., Strehl, A., Merugu, S.: A consensus framework for integrating distributed clusterings under limited knowledge sharing. In: Proceedings of the NSF Workshop on Next Generation Data Mining, pp. 99–108 (2002). URL http://strehl.com/download/ghosh-ngdm02.pdf
18. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Trans. Knowl. Discov. Data (TKDD) **1**(1), 4 (2007)
19. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theor. Comput. Sci. **38**, 293–306 (1985)
20. Gonzàlez, E., Turmo, J.: Comparing non-parametric ensemble methods for document clustering. In: Natural Language and Information Systems, pp. 245–256. Springer, Berlin (2008)
21. Grötschel, M., Wakabayashi, Y.: A cutting plane algorithm for a clustering problem. Math. Program. **45**(1–3), 59–96 (1989)
22. Hornik, K.: A clue for cluster ensembles. J. Stat. Software **14**(12), 1–25 (2005). URL http://www.jstatsoft.org/v14/i12
23. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
24. Kim, J., He, Y., Park, H.: Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. J. Global Optim. **58**(2), 285–319 (2014)
25. Kriegel, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. Wiley Interdiscip. Rev. Data Mining Knowl. Discov. **1**(3), 231–240 (2011)
26. Křivánek, M., Morávek, J.: NP-hard problems in hierarchical-tree clustering. Acta Informatica **23**(3), 311–323 (1986)
27. Lancichinetti, A., Fortunato, S.: Consensus clustering in complex networks. Sci. Rep. **2**, 336 (2012). URL http://www.nature.com/srep/2012/120327/srep00336/full/srep00336.html
28. Li, T., Ding, C.: Weighted consensus clustering. In: Proceedings of the 2008 SIAM International Conference on Data Mining (2008)
29. Li, T., Ding, C., Jordan, M.I.: Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Seventh IEEE International Conference on Data Mining, 2007 (ICDM 2007), pp. 577–582. IEEE, Los Alamitos (2007)
30. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, p. 14. California (1967)
31. McQuitty, L.L.: Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. Educ. Psychol. Meas. **17**(2), 207–229 (1957)
32. Meilă, M.: Comparing clusterings – an information based distance. J. Multivariate Anal. **98**(5), 873–895 (2007)
33. Milligan, G.W., Cooper, M.C.: Methodology review: clustering methods. Appl. Psychol. Meas. **11**(4), 329–354 (1987)
34. Mirkin, B.: Mathematical Classification and Clustering: From How to What and Why. Springer, Dordrecht (1998)
35. Mirkin, B.: Reinterpreting the category utility function. Mach. Learn. **45**(2), 219–228 (2001)
36. Monti, S., Tamayo, P., Mesirov, J., Golub, T.: Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach. Learn. **52**(1–2), 91–118 (2003)
37. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal Process. Mag. **13**(6), 47–60 (1996)
38. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, Cambridge (2011)

39. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
40. Rosenberg, A., Hirschberg, J.: V-measure: a conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL, vol. 7, pp. 410–420 (2007)
41. Saeed, F., Salim, N., Abdo, A.: Voting-based consensus clustering for combining multiple clusterings of chemical structures. J. Cheminformatics **4**(1), 1–8 (2012)
42. Saeed, F., Salim, N., Abdo, A., Hentabli, H.: Combining multiple individual clusterings of chemical structures using cluster-based similarity partitioning algorithm. In: Advanced Machine Learning Technologies and Applications, pp. 276–284. Springer, New York (2012)
43. Saeed, F., Salim, N., Abdo, A.: Information theory and voting based consensus clustering for combining multiple clusterings of chemical structures. Mol. Inform. **32**(7), 591–598 (2013)
44. Seiler, M., Huang, C.C., Szalma, S., Bhanot, G.: Consensuscluster: a software tool for unsupervised cluster discovery in numerical data. OMICS J. Integr. Biol. **14**(1), 109–113 (2010)
45. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 731–737. IEEE, Los Alamitos (1997)
46. Shinnou, H., Sasaki, M.: Ensemble document clustering using weighted hypergraph generated by nmf. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 77–80. Association for Computational Linguistics, Prague (2007)
47. Simpson, T.I., Armstrong, J.D., Jarman, A.: Merged consensus clustering to assess and improve class discovery with microarray data. BMC Bioinform. **11**(1), 590 (2010)
48. Smola, A.J., et al.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge (2002)
49. Sneath, P.H.: The application of computers to taxonomy. J. Gen. Microbiol. **17**(1), 201–226 (1957)
50. Steinbach, M., Karypis, G., Kumar, V., et al.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining, vol. 400, pp. 525–526. Boston (2000)
51. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
52. Sukegawa, N., Yamamoto, Y., Zhang, L.: Lagrangian relaxation and pegging test for the clique partitioning problem. Adv. Data Anal. Classif. **7**(4), 363–391 (2013)
53. van Rijsbergen, C.J.: Foundation of evaluation. J. Doc. **30**(4), 365–373 (1974)
54. Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J.: Weighted cluster ensemble using a kernel consensus function. In: Progress in Pattern Recognition, Image Analysis and Applications, pp. 195–202. Springer, Berlin (2008)
55. Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J.: Weighted partition consensus via kernels. Pattern Recognit. **43**(8), 2712–2724 (2010)
56. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. Int. J. Pattern Recognit. Artif. Intell. **25**(03), 337–372 (2011)
57. Viswanath, S., Bloch, B.N., Genega, E., Rofsky, N., Lenkinski, R., Chappelow, J., Toth, R., Madabhushi, A.: A comprehensive segmentation, registration, and cancer detection scheme on 3 tesla in vivo prostate dce-mri. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008, pp. 662–669. Springer, Berlin (2008)
58. Wattuya, P., Jiang, X.: Ensemble combination for solving the parameter selection problem in image segmentation. In: Structural, Syntactic, and Statistical Pattern Recognition, pp. 392–401. Springer, Berlin (2008)
59. Wattuya, P., Rothaus, K., Praßni, J.S., Jiang, X.: A random walker based approach to combining multiple segmentations. In: 19th International Conference on Pattern Recognition, 2008 (ICPR 2008), pp. 1–4. IEEE, Piscataway (2008)
60. Xanthopoulos, P., Guarracino, M.R., Pardalos, P.M.: Robust generalized eigenvalue classifier with ellipsoidal uncertainty. Ann. Oper. Res. **216**(1), 327–342 (2014)
61. Xanthopoulos, P., Pardalos, P.M., Trafalis, T.B.: Robust Data Mining. Springer, New York (2013)

62. Xu, S., Lu, Z., Gu, G.: An efficient spectral method for document cluster ensemble. In: The 9th International Conference for Young Computer Scientists, 2008 (ICYCS 2008), pp. 808–813. IEEE, Los Alamitos (2008)
63. Yu, Z., Wong, H.S., Wang, H.: Graph-based consensus clustering for class discovery from gene expression data. Bioinformatics **23**(21), 2888–2896 (2007)
64. Zhang, X., Jiao, L., Liu, F., Bo, L., Gong, M.: Spectral clustering ensemble applied to sar image segmentation. IEEE Trans. Geoscience Remote Sensing **46**(7), 2126–2136 (2008)
65. Zhao, Y., Karypis, G.: Criterion Functions for Document Clustering: Experiments and Analysis. UMN CS 01-040 (2001)

Influence Diffusion in Social Networks

Wen Xu, Weili Wu, Lidan Fan, Zaixin Lu, and Ding-Zhu Du

1 Introduction

With hundreds of millions of users worldwide, social networks provide great opportunities for social connection, learning, political and social change, as well as individual entertainment and enhancement in a wide variety of forms. Although social interaction is possible in the workplace, universities, communities, it is most popular online. Online social networks (OSNs) allow individuals to present themselves, articulate their social networks, and establish or maintain connections with others. Now, massive amounts of information about social networks and social interactions are recorded, which can allow social scientists to study social interactions on a scale and at a level of detail that has never before been possible. With the rapid development of online communities and devices connecting to the Internet such as smart phones, new possibilities of basic human activities have emerged. For instance, the process by which people locate, organize, and coordinate groups of individuals with shared interests, the number and nature of information and news sources available, and the ability to solicit and share opinions and ideas across various topics have all undergone dramatic change with the rise of social networks.

Social networks have already become a significant medium for the widespread distribution of news and instructions in mass convergence events such as presidential elections [1, 2], and emergencies like the landfall of Hurricanes Ike and Gustav in the fall of 2008 [1]. OSNs such as Facebook and Twitter have also been well known for providing great ease during the recent demonstrations in Middle East [3]. In light of these notable events, understanding information diffusion in OSNs has become a critical research goal. This great understanding can be achieved through effective data analysis, the development of reliable models that can predict outcomes of social

W. Xu (✉) • W. Wu • L. Fan • Z. Lu • D.-Z. Du
Department of Computer Science, University of Texas at Dallas,
800 W. Campbell Road, MS EC31, Richardson, TX 75080, USA
e-mail: wen.xu@utdallas.edu; weiliwu@utdallas.edu; lidan.fan@utdallas.edu;
zaixinlu@utdallas.edu; dzdu@utdallas.edu

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_27,

processes, and ultimately the creation of applications that can shape the outcome of these processes. In this paper, we provide an overview of the recent research in this area based on a wide variety of techniques such as optimization algorithms, data mining, data streams covering a large number of problems such as influence spread maximization, misinformation limitation, and study of trends in OSNs.

2 Characteristics of Social Networks

In order to study influence diffusion in OSNs, it is necessary to understand the significance and characteristics of social networks first. In this section, we give an overview about social networks identifying its significance and characteristics.

As a complex network, social networks have some well-known theoretical properties like Power-law distribution, Small-world, Scale-free et al. Power-law distribution means that the probability that a node has degree k is proportional to $k^{-\gamma}$, for large k and $\gamma > 1$. The parameter γ is called the power-law coefficient. Researchers have shown that many real-world networks are power-law networks, including Internet topologies [4], the Web [5, 6], social networks [7], neural networks [8], and power grids [9]. Scale-free networks are a class of power-law networks where the high-degree nodes tend to be connected to other high-degree nodes. Scale-free graphs are discussed in detail by Li et al. [10], and they propose a metric to measure the scale-freeness of graphs. Small-world networks have a small diameter and exhibit high clustering. Studies have shown that the Web [11, 12], scientific collaboration on research papers [13], film actors [14], and general social networks [7] have small-world properties. Kleinberg [15, 16] proposes a model to explain the small-world phenomenon in off-line social networks and also examines navigability in these networks.

There are also many other interesting properties about social networks that have been studied by various sociologists. Milgram [17] shows that the average path length between two Americans is six hops, and Pool and Kochen [18] provide an analysis of the small-world effect. The influential paper by Granovetter [19] argues that a social network can be partitioned into strong and weak ties, and that the strong ties are tightly clustered. For an overview of social network analysis techniques, we refer the reader to the book by Wasserman and Faust [20]. A prominent study of the Web link structure [12] shows that the Web has a "bow-tie" shape, consisting of a single large strongly connected component (SCC), and other groups of nodes that can either reach the SCC or can be reached from the SCC. OSNs have a similar large component, but that its relative size is much larger than that of the Web's SCC. Faloutsos et al. [4] show that the degree distribution of the Internet follows a power-law. Siganos et al. demonstrate that the high-level structure of the Internet resembles a "jellyfish" [21]. Kleinberg [22] demonstrates that high-degree pages in the Web can be identified by their function as either hubs (containing useful references on a subject) or authorities (containing relevant information on a subject). Kleinberg also

presents an algorithm [23] for inferring which pages function as hubs and which as authorities. The well-known PageRank algorithm [24] uses the Web structure to determine pages that contain authoritative information.

As OSNs are gaining popularity, sociologists and computer scientists are beginning to investigate their properties. Adamic et al. [7] study an early OSN at Stanford University and find that the network exhibits small-world behavior, as well as significant local clustering. Liben-Nowell et al. [25] find a strong correlation between friendship and geographic location in social networks by using data from LiveJournal. Kumar et al. [26] analyze two OSNs and discover that both possess a large SCC. Girvan and Newman observe that users in OSNs tend to form tightly knit groups, which is also called communities [27]. Backstrom et al. [28] examine snapshots of group membership in LiveJournal and present models for the growth of user groups over time.

3 Diffusion of Influence

In this section, we outline the techniques used in optimizing or facilitating information diffusion in social networks. There are a large number of problems that are related to the diffusion of information in social networks. We study two example problem definitions through which a broad survey of techniques in recent research is provided, namely, (a) maximizing the spread of influence and (b) minimizing the spread of misinformation in social networks. Next we delve into details about influence diffusion from the following subsections: (1) overview of influence diffusion, (2) formalization and optimization, (3) large-scale data analysis.

3.1 Overview of Influence Diffusion

Diffusion of influence refers to circumstances where a point of view or behavior is widely spread in specific structures of propagation channels [29]. A diffusion can be associated with topological properties, such as scale, range, and temporal properties. This concept has been widely researched in the field of epidemiology, sociology, and marketing. In early time, biology and epidemiology have conducted in-depth study on diffusion of virus within the group [30], and two classical models: SIS and SIR are proposed. In sociology and marketing area, research on diffusion focuses on the problems of innovation diffusion. In the early twentieth century, Schumpeter et al. [31] created innovative theory. Then the BASS model [32] opened up new research directions for this research area and derived a series of related models. Westerman et al. [33] studied the effect of system generated reports of connectedness on credibility, and have shown that there are curvilinear effects for the number of followers exist, such that having too many or too few connections results in lower judgments of expertise and trustworthiness. Lopez-Pintado et al. [34]

studied the product diffusion in complex social networks. He considered the mutual influence among individuals on the micro-level into the propagation equation based on mean-field theory, and found out that innovation diffusion in complex networks has a threshold which is closely related to the degree distribution and propagation functions of the network.

Information propagation considering two repellent relationship is studied in the field of competitive marketing and disease control. Virus propagation and immunization can be considered as two kinds of mutually exclusive information diffusing in networks in the field of disease control. Meier et al. [35] studied inoculation game problem in OSNs, that is, whether each node can select to protect itself when virus diffusing in networks. Salathé et al. [36] developed an algorithm that acts only on locally available network information and is able to quickly identify targets for successful immunization intervention. They also demonstrated that community structure strongly affects disease dynamics.

Understanding, capturing, and being able to predict influence diffusion can be helpful for several areas such as marketing, security, and Web search. For instance, if we consider the case of marketing, it may be useful to know which are the features that control the process of diffusing information when it is created to, e.g. better advertise a product or to better protect it against attacks on the network. The marketing may also benefit from information such as how many initial users to start with in a marketing campaign (budget optimization), how much time to wait between actions etc. In the case of security, criminal investigators generally need to understand the information flow between, e.g. members of a given community, to extract hints regarding possible guilt or innocence of a person or a group of persons. This is clearly an observation phase where the user wants to understand the route that information took and possible links. Finally, as Web search evolves, if we consider the case of subscriptions to feeds, a propagation prediction model may be useful for the user to, e.g. subscribe to the most interesting topic according to its expected growth (in addition to his interests). This reflects a more active usage of the diffusion prediction.

3.2 Formalization and Optimization

To better understand the underlying ideas behind diffusion and social networks, we study the formulations and optimizations for two important problems in social networks, (1) maximizing the spread of influence, (2) limiting the spread of misinformation, which is also called rumor blocking in some related work.

To begin with, we will cover two basic diffusion models that have been researched intensely and will illustrate the differences between them. Since diffusion models and the process of diffusion are the foundation of related research, we provide as much background as possible.

3.2.1 Two Basic Diffusion Models

Diffusion is the process by which information passes from neighbor to neighbor. Real-world examples include viral marketing, innovation of technologies, and infection propagation. Diffusion models are the framework on which diffusion occurs.

Definition 1. A diffusion model is a graph $G = V, E$ along with a collection of activation functions $F = (f_v)_{v \in V}$, where f_v is a $\{\emptyset, \{v\}\}$ valued function on $2^{|V|}$.

The output of a function f_v is a random variable based on the activation function.

Vertices on this graph are usually individuals and the activation function models the influence individuals exert on others. The activation function usually depends only on the neighbors of v, denoted $N(v)$. This means that $f_v(S) = f_v(N(v) \bigcap S)$.

Definition 2. Diffusion is the process on a diffusion model M, $S = (S_t)_{t=0}^{n-1}$ started at $S \subseteq V$:

1. set $S_0 = S$
2. for $t > 1$ set $S_t = f(S_{t-1}) = \text{def} \bigcup_{v \in V} f_v(S_{t-1})$

The set of nodes activated at the end of diffusion is denoted as $\sigma(S) = \bigcup_{t=0}^{n}(S_t)$.

Diffusion occurs in time steps t. At each time step, all previously activated nodes remain activated and individuals are either activated or deactivated based on the activation functions. Diffusion can run on a fixed number of time steps or indefinitely. Diffusion is said to have stopped when the set of activated nodes in time step t_k is the same as the set in time step t_{k+n} for all $n \geq 1$.

One class of diffusion models, namely threshold model, adds an influence threshold to each individual, which, when overcome, triggers the individual to be activated. There is a cumulative effect of these models, as it takes a critical number of influential neighbors to activate an individual. The linear threshold model is a specialized form of general threshold models. The linear threshold model, LT model in short, is more often used in marketing research.

Definition 3. The **linear threshold model** is a diffusion model with all of the following properties:

1. A set of threshold values $(\theta_v)_{v \in V}$, where θ_v is in the range $[0, 1]$.
2. Node v being activated if $f_v(S) \geq \theta_v$, where S is the set of neighbors of v.
3. A set of weights $(p(u,v))_{(u,v) \in E}$ with the property $\Sigma_{u \in N(v)} p(u,v) \leq 1$.
4. Activation function of the form $f_v(S) = \sum_{u \in N(v)} p(v,u)$ with $f(\emptyset) = 0$.

Cascade models of diffusion give each individual the ability to influence their neighbors as soon as they are activated. This is opposed to the threshold models that rely on a cumulative effect. This model has the property that the more nodes that have attempted to influence a node, the less likely the node is to be activated. Here we give a definition of a specialized cascade model, namely the independent cascade model, IC model in short.

Definition 4. The **independent cascade model** is a diffusion model with the following properties:

1. Each arc (u,v) has associated the probability $p(u,v)$ of u influencing v.
2. Time unfolds in discrete steps.
3. At time t, nodes that became active at $t-1$ try to active their inactive neighbors, and succeed according to $p(u,v)$.

Note that the probability of a node u influencing a node v is independent of the set of nodes S that has attempted to influence v.

There is an assumption of monotonicity on this model made to reflect that adding active neighbors to a node increases likelihood of the node being activated.

3.2.2 Influence Maximization

An intensively studied problem in viral marketing is that, by picking a small group of influential individuals in a social network—say, convincing them to adopt a product—it will trigger the largest cascade of influence by which many users will try the product ultimately. Domingos and Richardson [37] are the first to pose it as an algorithmic problem and solve it as a probabilistic model of interaction. In [38], Kempe et al. formalize it as the problem of influence maximization.

A social network is modeled as a directed graph $G=(V,E)$ with vertices in V modeling the individuals and edges in E modeling the relationship between individuals. For example, in co-authorship graphs, vertices are authors of academic papers and two vertices have an edge if the two corresponding authors have coauthored a paper. Let p denote the influence probabilities between two vertices. The influence is propagated in the network according to a diffusion model m. Let S be the subset of vertices selected to initiate the influence propagation, which is also called seed set. Let $\sigma_m(S)$ be the expected number of influenced nodes at the end of propagation process. The formal definition of influence maximization problem is given as follows:

Problem 1 (Influence Maximization). Given a directed and edge-weighted social graph $G=(V,E,p)$, a propagation model m, and an integer $k \leq |V|$, find a seed set $S \subset V$, $|S|=k$, such that the expected influence $\sigma_m(S)$ is maximum.

This problem is also referred to as the identification of influential users or opinion leaders in a social network. This problem under both independent cascade (IC) and linear threshold (LT) propagation models is shown to be NP-hard, and so attempts have been made at approximating the value of $\sigma_m(S)$ [39].

For a diffusion model with a nonnegative, monotone submodula activation function, a greedy hill-climbing algorithm approximates the optimum within a factor of $(1-1/e)-\varepsilon$ for any real number ε, as shown by Kempe et al. [38]. The complexity of influence maximization problem has been further discussed in [40–42]. By greedy hill-climbing algorithm we mean an algorithm which, at every step, adds to the output set the element that currently has the highest value.

Submodularity is a property on a diffusion model that states that the influence gained from adding nodes to the infected set decreases or stays the same as the set becomes larger. This condition can be read as a principle of diminishing returns, where the value of adding a node to the infected set decreases based on the size of the infected set.

The challenge of the greedy algorithm rises when selecting a new vertex v that provides the largest marginal gain $\sigma_m(S+v) - \sigma_m(S)$ compared to the influence spread of current seed set S. Computing the expected spread given a seed set turns out to be a difficult task under both the IC model and the LT model. Instead of finding an exact algorithm, Kempe et al. run Monte-Carlo simulations of the propagation model for sufficiently many times (10,000 trials) to obtain an accurate estimate of the influence spread, leading to a very long computation time. In [43], Mathioudakis et al. simplified the network to accelerate the speed of finding seeds. Facing serious efficiency and scalability limits, several heuristics [44–46] are proposed to overcome it. In [45], Chen et al. propose a scalable heuristic called DAGs (local directed acyclic graphs) for the LT model. They construct local DAGs for each node and computing the expected spread over DAGs can be done in linear time while over general graphs it is #P-hard. In [44], Chen et al. also propose a PMIA heuristic to estimate the influence spread under the IC model. However, these heuristics lack of theoretical guarantees.

Another issue for Kempe's method is that it assumes a weighted social graph as input and does not address the problem of learning influence probabilities. In [47], Saito et al. study how to learn the probabilities of the IC model from a set of past propagations by formalizing this as a likelihood maximization problem and then applying the expectation maximization (EM) algorithm to solve it; Goyal et al. [48, 49] propose a credit model for learning influence probability from pure historical action logs which takes the temporal nature of influence into account.

Some variations are also proposed to handle different real-world requirements. Leskovec et al. [50] optimized placements for a set of social sensors such that the propagation of information or virus can be effectively detected in a social network. Lappas et al. [51] discover a set of key mediators which determine the bottlenecks of influence propagation if seed nodes try to activate some target nodes.

A characteristic common to the studies discussed so far is the assumption that information cascades of campaigns happen in isolation. Next we introduce a group of problem formulations that capture the notion of competing campaigns in a social network [52–57]. This scenario will frequently arise in the real world: multiple companies with comparable products will vie for sales with competing word-of-mouth cascades; similarly, many innovations face active opposition also spreading by word of mouth.

Dubey et al. [58] study competitive information game problem in networks based on quasilinear model. They find the Nash equilibrium by considering the adoption of the costs, benefits, and external functions of the different information conditions. Carnes et al. [52] study the strategies of a company that wishes to invade an existing market and persuade people to buy their product. This turns the problem into a Stackelberg game where in the first player (leader) chooses a strategy in the first

stage, which takes into account the likely reaction of the second players (followers). In the second stage, the followers choose their own strategies having observed the Stackelberg leader decision i.e., they react to the leader's strategy. Carnes et al. use models similar to the ones proposed in [38] and show that the second player faces an NP-hard problem if aiming at selecting an optimal strategy. Furthermore, the authors prove that a greedy hill-climbing algorithm lead to a $(1-1/e-\varepsilon)$-approximation.

Around the same time, Bharathi et al. [53] introduce roughly the same model for competing rumors and they also show that there exists an efficient approximation algorithm for the second player. Moreover they present an FPTAS for the single player problem on trees. Kostka et al. [54] considered the rumors diffusion as a game theoretical problem under a much more restricted model compared with IC and LT. They showed that the first player did not always obtain benefit although he/she started earlier. Trpevski et al. [55] propose a competitive rumors spreading model based on SIS model in epidemic domain, but they did not address the issue of influence maximization or rumor blocking. Borodin et al. in [56] study competitive influence diffusion in several different models extended from LT. Chen et al. [57] address positive influence maximization under an extension of the IC model with negative opinions about the product or service quality.

3.2.3 Misinformation Minimization

While the ease of information propagation in social networks can be very beneficial, it can also have disruptive effects. A number of examples of this sort are the spread of misinformation on swine flu in Twitter [59], exaggerated reports on a bomb attack in Grand Central and celebrities who are falsely claimed as being dead [60]. We specifically focus on the study that addresses the problem of influence limitation [61, 63] where a "bad" campaign starts propagating from a certain node in the network and use the notion of limiting campaigns to counteract the effect of misinformation. The problem of misinformation minimization can also be called as rumor blocking problem, or influence limitation problem. Its definition is defined as follows:

Problem 2 (Misinformation Minimization). Given a graph $G = (V; E; p)$, where p represents its positive and negative edge weights, a negative seed set N_0, and a positive integer k, the goal is to find a positive seed set S of size at most k such that the expected number of negatively activated nodes is minimized, or equivalently, the reduction in the number of negatively activated nodes is maximized.

Kimura et al. in [62] deal with influence limitation problem through blocking a certain number of links in a network. The most recent works regarded with this problem include [63–65]. In [63], Budak et al. study the controlling of negative information in social networks, that is, when negative information is diffused in networks, how to select some nodes to implant positive information in order to correct the information attitude in the whole network to a maximizing extent. They prove that under an extension of the IC model, the eventual influence limitation (EIL) problem

is NP-hard. They also examine a more realistic problem of influence limitation in the presence of missing information and introduced an algorithm called predictive hill-climbing approach which has good performance.

In [64], He et al. propose a competitive linear threshold (CLT) model to address the influence blocking maximization (IBM) problem, which is an extension to the classic linear threshold model. They prove that this problem under CLT model was submodular and theoretically obtained a $(1-1/e)$-approximation ratio by a greedy strategy. To improve the efficiency, they further propose the CLDAG algorithm that is similar to the LDAG algorithm in [45]. In [65], a β_T^I-Node Protector problem is proposed by Nguyen et al., which is actually the extensions of the Misinformation Minimization problem under LT and IC models. The goal is to find the smallest set of highly influential nodes that can limit the viral spread of misinformation originated from set I to a desired rate $(1-\beta)$ $(\beta \in [0,1])$ in T time steps. They present a greedy viral stopper (GVS) algorithm that greedily adds nodes with the best influence gain for β Node Protectors to the current solution. They also apply GVS to the network restricted to T-hop neighbors of the initial set I and reached a slightly better bound for β_T^I-Node Protector problems. Besides, a community-based algorithm which returns a good selection of nodes to decontaminate in a timely manner is proposed.

3.3 Large-Scale Data Analysis

No matter which technique is used in studying information diffusion, large-scale data analysis is a significant aspect of study as well as being a significant challenge. In this part, we will introduce several representative data analysis techniques used in the social influence analysis. With the increase of studies in social networks, there are a number of data sets available to researchers [66–69].

As data grows, data mining and machine learning applications start to embrace the Map-Reduce paradigm, e.g., news personalization with Map-Reduce EM algorithm [70], Map-Reduce of several machine learning algorithms on multicore architecture [71]. For the networking data, graphical probabilistic models are often employed to describe the dependencies between observation data. Markov random field [72], factor graph [73], restricted Boltzmann machine (RBM) [74], and many others are widely used graphical models. In [75], Tang et al. proposed a topical factor graph (TFG) model, for quantitatively analyzing the topic-based social influences. Compared with the existing work, the TFG can incorporate the correlation between topics. They also proposed a very efficient algorithm for learning the TFG model. In particular, a distributed learning algorithm has been implemented under the Map-reduce programming model.

The techniques used in Web community discovery can also be applied in social influence analysis. The problem of detecting such communities within networks has been well studied. Early approaches such as spectral partitioning, the Kernighan–Lin algorithm, hierarchical clustering, and G-N algorithm work well for spe-

cific types of problems (particularly graph bisection), but perform poorly in real networks. Recently, most works focus on graph partitioning approaches. The most popular partition technique in the literature is k-means clustering, which aims to separate the nodes in k clusters such to maximize/minimize a given cost function based on distances between nodes and/or from nodes to centroids. In [76], Q. Yan et al. proposed a two-phase method that combines community detection with naive greedy algorithm to improve time efficiency of influence maximizing problem with multiple spread model. In the first phase, they use efficient clustering algorithm such as kernel k-means to partition graph nodes into k clusters, with the parameter k related to the number of influential nodes. In the second phase, in each community, they apply techniques in social influence maximization to find influential nodes in each cluster. Similar work has [77].

4 Research Trends

Social networks provide large-scale information infrastructures for people to discuss and exchange ideas about different topics. The general problem of network influence analysis represents a new and interesting research direction in social network mining. There are many potential future directions of this work. Even though the influence diffusion in social networks has been intensively studied, we note that there are three essential dimensions emerging from the analysis we performed, which could be of great benefits for future researchers.

4.1 Learn Influence Probabilities for IC and LT Models

In social network analysis, two information diffusion models: the independent cascade (IC) and the linear threshold (LT) are widely used to solve such problems as the influence maximization problem and the misinformation minimization problem. These two models focus on different information diffusion aspects. The IC model is sender-centered (push type) and each active node independently influences its inactive neighbors with given diffusion probabilities. The LT model is receiver-centered (pull type) and a node is influenced by its active neighbors if their total weight exceeds the threshold for the node. What is important to note is that both models have parameters that need to be specified in advance: diffusion probabilities for the IC model, and weights for the LT model. However, their true values are not known in practice. This poses yet another problem of estimating them from a set of information diffusion results that are observed as timesequences of influenced (activated) nodes. This falls in a well-defined parameter estimation problem in machine learning framework.

In [78], K. Saito et al. extended both IC and LT models to be able to simulate asynchronous time delay. They learned the dependency of the diffusion probability

and the time-delay parameter on the node attributes by solving a formulated problem named as the maximum likelihood estimation problem, and an efficient parameter update algorithm that guarantees the convergence is derived. Other efforts of learning parameters of the influence graph from history data include the work [49, 79]. In [49], A. Goyal et al. proposed both static and time-dependent models for capturing influence. Moreover, they presented optimized algorithms for learning the parameters of the various models based on social networks and historical action logs.

4.2 Learn the Speed of Influence Spread in Networks

It has been observed that information spreads extremely fast in social networks. There has been some but not enough theoretical results about the analysis of influence spread speed. In [80], B. Doerr et al. have shown that for preferential attachment graphs the classic push–pull strategy needs $\Theta(\log n)$ rounds to inform all vertices. The slightly improved version which avoids that a vertex contacts the same neighbor twice in a row only needs $\Theta(\log n/\log\log n)$ rounds, which is best possible since the diameter is of the same order of magnitude. In [81], N. Fountoulakis et al. establishe for a class of random graphs ultrafast time bounds on the running time of the synchronous push–pull protocol that is needed until the majority of the vertices are informed. We present the first theoretical analysis of this protocol on random graphs that have a power law degree distribution with an arbitrary exponent $\beta > 2$. Their main findings reveal a striking dichotomy in the performance of the protocol that depends on the exponent of the power law. More specifically, it is shown that if $2 < \beta < 3$, then the rumor spreads to almost all nodes in $\Theta(\log\log n)$ rounds with high probability. On the other hand, if $\beta > 3$, then $\Theta(\log n)$ rounds are necessary.

4.3 Study the Time Constrained Influence Diffusion Problem

Traditional diffusion models including IC and LT do not fully incorporate important temporal aspects that have been well observed in the dynamics of influence propagation. Firstly, the propagation of influence from one person to another may incur a certain amount of time delay, which is obvious from recent studies by statistical physicists on empirical social networks. Secondly, the spread of influence may be time-critical in practice. In a certain viral marketing campaign, a company might wish to trigger a large cascade of product adoption in a fairly short time frame, e.g., a 3-day sale. Therefore it is very meaningful to extend the influence maximization problem to have a time constraint.

Chen et al. [82] proposed the time-critical influence maximization problem, in which one wants to maximize influence spread within a given deadline. In their

model influence delays are constrained to follow the geometric distribution. In [83], B. Liu et al. proposed a new problem of the time constrained influence maximization in social networks based on a Latency Aware Independent Cascade model. They also proposed to use Influence Spreading Paths to quickly and effectively approximate the time constrained influence spread for a given seed set.

References

1. Hughes, A.L., Palen, L.: Twitter adoption and use in mass convergence and emergency events. In: Proceedings of the 6th International Information Systems for Crisis Response and Management Conference (2009)
2. Grossman, L.: Iran protests: Twitter, the medium of the movement. Time (online) (June 2009). http://www.time.com/time/world/article/0,8599,1905125,00.html
3. Smith, C.: Egypt's facebook revolution: Wael ghonim thanks the social network. The Huffington Post, February 2011. http://www.huffingtonpost.com/2011/02/11/egypt-facebook-revolution-wael-ghonim_n_822078.html
4. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM), Cambridge, August 1999
5. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
6. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the web for emerging cyber-communities. Comput. Netw. **31**, 1481–1493 (1999)
7. Adamic, L.A., Buyukkokten, O., Adar, E.: A social network caught in the Web. First Monday **8**(6), 35–42 (2003)
8. Braitenberg, V., Schüz, A.: Anatomy of a Cortex: Statistics and Geometry. Springer, Berlin (1991)
9. Phadke, A.G., Thorp, J.S.: Computer relaying for power systems. Wiley, New York (1988)
10. Li, L., Alderson, D., Doyle, J.C., Willinger, W.: Towards a theory of scale-free graphs: definitions, properties, and implications. Internet Math. **2**(4), 431–523 (2006)
11. Albert, R., Jeong, H., Barabasi, A.L.: The diameter of the world wide web. Nature **401**, 130 (1999)
12. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the web: Experiments and models. In: Proceedings of the 9th International World Wide Web Conference (WWW), Amsterdam, May 2000
13. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. (PNAS) **98**, 409–415 (2001)
14. Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of small-world networks. Proc. Natl. Acad. Sci. (PNAS) **97**, 11149–11152 (2000)
15. Kleinberg, J.: The small-world phenomenon: An algorithmic perspective. In: Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC), Portland, May 2000
16. Kleinberg, J.: Navigation in a small world. Nature **406**, 845–845 (2000)
17. Milgram, S.: The small world problem. Psychol. Today, **2**(60), 60–67 (1967)
18. Pool, I., Kochen, M.: Contacts and influence. Soc. Netw. **1**, 1–48 (1978)
19. Granovetter, M.: The strength of weak ties. Am. J. Sociol. **78**(6), 1360–1380 (1973)
20. Wasserman, S., Faust, K.: Social Networks Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
21. Siganos, G., Tauro, S.L., Faloutsos, M.: Jellyfish: A conceptual model for the AS internet topology. J. Commun. Netw. **8**(3), 339–350 (2006)
22. Kleinberg, J., Lawrence, S.: The structure of the web. Science **294**, 1849–1850 (2001)

23. Kleinberg, J.: Authoritative sources in a hyperlinked environment. J. ACM **46**, 604–632 (1999)
24. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University (1998)
25. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proc. Natl. Acad. Sci. (PNAS) **102**(33), 11623–11628 (2005)
26. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, August 2006
27. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. (PNAS) **99**, 7821–7826 (2002)
28. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: Membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Philadelphia, August 2006
29. Centola, D.: The spread of behavior in an online social network experiment. Science **329**(5996), 1194–1197 (2010)
30. Anderson, R.M., May, R.M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford (1992)
31. Schumpeter, J., Bakhays, U.: The Theory of Economics Development. Springer, New York (2003)
32. Albert, R., Barabasi, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)
33. Westermana, D., Spenceb, P.R., Heide, B.V.D.: A social network as information: The effect of system generated reports of connectedness on credibility on twitter. Comput. Hum. Behav. **28**(1), 199–206 (2012)
34. Lopez-Pintado, D.: Diffusion in complex social networks. Games Econ. Behav. **62**(2), 573–590 (2008)
35. Meier, D., Oswald, Y.A., Schmid, S., Wattenhofer, R.: On the windfall of friendship: Inoculation strategies on social networks. In: ICEC, pp. 294–301 (2008)
36. Salathé, M., Jones, J.H.: Dynamics and control of diseases in networks with community structure. PLoS Comput. Biol. **6**(8), e1000736 (2010). doi:10.1371/journal.pcbi.1000736
37. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 57–66 (2001)
38. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedingsof the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)
39. Kempe, D., Kleinberg, J., Tardos, E.: Influential nodes in a diffusion model for social networks. In: ICALP, pp. 1127–1138. Springer, New York (2005)
40. Mossel, E., Roch, S.: On the submodularity of influence in social networks. In: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing (STOC), p. 128 (2007)
41. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions-i. Math. Program. **14**(1), 265–294 (1978)
42. Lu, Z., Zhang, W., Wu, W., Kim, J., Fu, B.: The complexity of influence maximization problem in the deterministic linear threshold model. J. Comb. Optim. **24**(3), 374–378 (2012)
43. Mathioudakis, M., Bonchi, F., Castillo, C., Gionis, A., Ukkonen, A.: Sparsification of influence networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'11), pp. 529–537, New York, USA (2011)
44. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'10), pp. 1029-1038, New York, USA (2010)
45. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: Proceedings of the 10th IEEE International Conference on Data Mining (ICDM'10), pp. 88–97 (2010)

46. Kimura, M., Saito, K.: Tractable models for information diffusion in social networks. In: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp. 259–271 (2006)
47. Saito, K., et al.: Prediction of information diffusion probabilities for independent cascade model. In: Knowledge-Based Intelligent Information and Engineering Systems (KES'08), Lecture Notes in Computer Science, **5179**, 67–75 (2008)
48. Goyal, A., Lu, W., Lakshmanan, L.V.S.: A data-based approach to social influence maximization. In: PVLDB **5**(1), 73–84 (2011)
49. Goyal, A., Bonchi, F., Lakshmanan, L.V.S.: Learning influence probabilities in social networks. In: Proceedings of the third ACM international conference on Web search and data mining (WSDM'10), pp. 241–250, New York, USA (2010)
50. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., Van-Briesen, J., Glance, N.S.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 420–429 (2007)
51. Lappas, T., Terzi, E., Gunopulos, D., Mannila, H.: Finding effectors in social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'10), pp. 1059–1068, New York, USA (2010)
52. Carnes, T., Nagarajan, C., Wild, S.M., van Zuylen, A.: Maximizing influence in a competitive social network: A follower's perspective. In: Proceedings of the 9th International Conference on Electronic Commerce (ICEC) (2007)
53. Bharathi, S., Kempe, D., Salek, M.: Competitive influence maximization in social networks. In: Proceedings of the 3rd international conference on Internet and network economics (WINE'07), **4858**, 306–311 (2007)
54. Kostka, J., Oswald, Y.A., Wattenhofer, R.: Word of mouth: Rumor dissemination in social networks. In: Structural Information and Communication Complexity (SIROCCO'08), Lecture Notes in Computer Science, **5058**, 185–196 (2008)
55. Trpevski, D., Tang, W.K.S., Kocarev, L.: Model for rumor spreading over networks. Phys. Rev. E, **81**(5), 056102 (2010)
56. Borodin, A., Filmus, Y., Oren, J.: Threshold models for competitive influence in social networks. In: Proceedings of the 6th international conference on Internet and network economics (WINE'10), pp. 539–550 (2010)
57. Chen, W., Collins, A., Cummings, R., Ke, T., Liu, Z., Rincn, D., Sun, X., Wang, Y., Wei, W., Yuan, Y.: Influence maximization in social networks when negative opinions may emerge and propagate. In: SIAM Data Mining (SDM), pp. 379–390 (2011)
58. Dubey, P., Garg, R., Meyer, B.D.: Competing for customers in a social network: The quasi-linear case. Internet Netw. Econ. **4286**, 162–173 (2006)
59. Morozov, E.: Swine flu. Twitter's power to misinform. Foreign Policy, April 2009. http://neteffect.foreignpolicy.com/posts/2009/04/25/swine_flu_twitters_power_to_misinform
60. Heussner, K.M.: Enough already! 7 twitter hoaxes and half-truths. ABC News (January 2010)
61. Fan, L., Lu Z., Wu W., Thuraisingham B., Ma H., and Bi Y.: Least Cost Rumor Blocking in Social Networks. In: Proceedings of the 33rd Internaltional Conference on Distributed Computing Systems (ICDCS), pp. 540–549 (2013)
62. Kimura, M., Saito, K., Motoda, H.: Minimizing the spread of contamination by blocking links in a network. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (2008)
63. Budak, C., Agrawal, D., Abbadi, A.E.: Limiting the spread of misinformation in social networks. In: International World Wide Web Conference (WWW'11), March 28–April 1, Hyderabad, India, pp. 665–674 (2011)
64. He, X., Song, G., Chen, W., Jiang, Q.: Influence blocking maximization in social networks under the competitive linear threshold model. In: SIAM Data Mining (SDM), pp. 463–474 (2012)
65. Nguyen, N.P., Yan, G., Thai, M.T., Eidenbenz, S.: Containment of Misinformation Spread in Online Social Networks. In: Proceedings of the 3rd Annual ACM Web Science Conference (WebSci'12), pp. 213–222, ACM New york, USA (2012)

66. C. for Computational Analysis of Social and O. S. (CASOS). Casos networks. http://www.casos.cs.cmu.edu/computational_tools/data2.php (2005)
67. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? http://an.kaist.ac.kr/traces/WWW2010.html (2010)
68. Leskovec, J.: Stanford large network dataset collection. http://snap.stanford.edu/data/index.html (2009)
69. Newman, M.: Network data. http://www-personal.umich.edu/~mejn/netdata/ (2013)
70. Das, A., Datar, M., Garg, A., Rajaram, S.: Google news personalization: Scalable online collaborative filtering. In: Proceeding of the 16th International Conference on World Wide Web (WWW) (2007)
71. Chu, C.-T., Kim, S.K., Lin, Y.-A., Yu, Y., Bradski, G.R., Ng, A.Y., Olukotun, K.: Map-reduce for machine learning on multicore. In: Proceedings of the 18th Neural Information Processing Systems (NIPS) (2006)
72. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, **1**, 194–281. MIT Press/Bradford Books, Cambridge (1986)
73. Kschischang, F.R., Member, S., Frey, B.J., Andrea Loeliger, H.: Factor graphs and the sum-product algorithm. IEEE Trans. Inf. Theory **47**, 498–519 (2001)
74. Welling, M., Hinton, G.E.: A new learning algorithm for mean field boltzmann machines. In: Proceedings of International Conference on Artificial Neural Network (ICANN), pp. 351–357 (2001)
75. Tang, J., Sun, J., Wang, C., Yang, Z.: Social influence analysis in large-scale networks. In: KDD (2009)
76. Yan, Q., Guo, S., Yang, D.: Influence maximizing and local influenced community detection based on multiple spread model. In: Advanced Data Mining and Applications (ADMA'11), Part II, LNAI 7121, pp. 82–95, (2011).
77. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 160–168 (2008)
78. Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., Motoda, H.: Learning diffusion probability based on node attributes in social networks. In: ISMIS, pp. 153–162 (2011)
79. Rodriguez, M.G., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: ICML, pp. 561–568 (2011)
80. Doerr, B., Fouz, M., Friedrich, T.: Social networks spread rumors in sublogarithmic time. In: Proceedings of the 43rd Annual ACM Symposium on Theory of Computing, pp. 21–30 (2011)
81. Fountoulakis, N., Panagiotouy, K., Sauerwaldz, T.: Ultra-fast rumor spreading in social networks. In: Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1642–1660 (2012)
82. Chen, W., Lu, W., Zhang, N.: Time-critical influence maximization in social networks with time-delayed diffusion process. In: AAAI, pp. 1–5 (2012)
83. Liu, B., Cong, G., Xu, D., Zeng, Y.: Time constrained influence maximization in social networks. In: IEEE International Conference on Data Mining (ICDM), December 2012

A New Exact Penalty Function Approach to Semi-infinite Programming Problem

Changjun Yu, Kok Lay Teo, and Liansheng Zhang

1 Introduction

Many real-world optimization problems in engineering design, such as the design of earthquake-resistant structures, multi-input multi-output control systems, wide-band amplifiers, and robot trajectory planning [6, 13–15], can be formulated as semi-infinite programming problems (SIPs). Some interesting applications in statistics can be found in [2, 10]. They include optimal experimental design in regression, constrained multinomial maximum likelihood estimation, robustness in Bayesian statistics, and actuarial risk theory.

A general SIPs can be stated in the form given below:

$$\min \ f(\mathbf{x}) \tag{1a}$$

$$\text{subject to} \qquad g_j(\mathbf{x}, \omega) \leq 0, \ \forall \ \omega \in \Omega, \ j = 1, \ldots, m, \tag{1b}$$

where $\mathbf{x} \in \mathbb{R}^n$ is a decision vector, Ω is a compact interval in $\mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable in x, and for each $j = 1, \ldots, m$, $g_j : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function in $\mathbf{x}$ and ω. Let this problem be referred to as Problem $(\mathbf{P})$.

C. Yu
Business School, Central South University, Kent Street, Bentley, WA 6102, Australia

Department of Mathematics and Statistics, Curtin University, South Lushan Road, Changsha, China
e-mail: yuchangjun@126.com

K.L. Teo (✉)
Department of Mathematics and Statistics, Curtin University,
Kent Street, Bentley, WA 6102, Australia
e-mail: k.l.teo@curtin.edu.au

L. Zhang
Shanghai University, 99 Shangda Road, Baoshan district Shanghai 200444, China

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_28,

Since there are infinite many inequality constraints in (1b), it is difficult to solve Problem (**P**) directly. Hence, SIP has become an active research area in optimization both in theory and numerical algorithms since 1970s. Many important publications have appeared in the literature. Examples include [1, 3–5, 9, 12, 17–20], and the relevant references cited therein. There are also several excellent review papers (see, for example, [6, 11]) devoted to SIP. The methods developed are mainly based on exchange methods, discretization methods, dual parametrization methods, method based on constraint transcription techniques, or methods based on local reduction.

In [21, 22], an exact penalty function approach is proposed for solving continuous inequality constraint optimization problems where the summation of the integrals of some smooth approximation functions is appended to the objective function forming an exact penalty objective function

$$f_\sigma(\mathbf{x},\varepsilon)=\begin{cases} f(\mathbf{x}), & \text{if } \varepsilon=0, g_j(\mathbf{x},\omega)\le 0\ (\omega\in\Omega),\\ f(\mathbf{x})+\varepsilon^{-\alpha}\Delta(\mathbf{x},\varepsilon)+\sigma\varepsilon^{\beta}, & \text{if } \varepsilon>0,\\ +\infty, & \text{otherwise}\end{cases} \tag{2}$$

where

$$\Delta(\mathbf{x},\varepsilon)=\sum_{j=1}^{m}\int_{\Omega}\left[\max\left\{0,g_j(\mathbf{x},\omega)-\varepsilon^{\gamma}\right\}\right]^2 d\omega.$$

Convergence analysis and numerical results show that the proposed method is effective.

In this paper, a new exact penalty function approach is proposed for solving the semi-infinite optimization problem (**P**). The purpose is to develop an alternative effective computational method for solving the semi-infinite optimization problem. In this approach, a logarithmic form function of the constraint violation is appended to the objective function forming a new exact penalty objective function $f_\sigma(\mathbf{x},\varepsilon)$. This gives rise to a sequence of optimization problems subject to $\varepsilon>0$. We shall show that any local minimizer of these optimization problems is a local minimizer of the original problem when the penalty parameter is sufficiently large.

The rest of the paper is organized as follows. In Sect. 2, we give a new exact penalty function and analyze its convergent properties. In Sect. 3, we devise an algorithm for solving Problem (**P**) via solving a sequence of optimization problems. Several examples are solved by using the algorithm proposed. Section 4 concludes the paper.

2 New Exact Penalty Function

Consider Problem (**P**). Define

$$S_\varepsilon=\{(\mathbf{x},\varepsilon)\in\mathbb{R}^n\times\mathbb{R}_+ : g_j(\mathbf{x},\omega)\le\varepsilon^{\gamma},\ \forall\ \omega\in\Omega,\ j=1,\ldots,m\}, \tag{3}$$

where $\mathbb{R}_+=\{\alpha\in\mathbb{R}\ :\ \alpha\ge 0\}$, $j=1,\ldots,m$, are fixed constants and γ is a positive real number. Clearly, Problem (**P**) is equivalent to the following problem, which is denoted as Problem ($\hat{\mathbf{P}}$).

$$\min\ f(\mathbf{x}) \tag{4a}$$

subject to

$$(\mathbf{x}, \varepsilon) \in S_0, \tag{4b}$$

where $S_0 = S_\varepsilon$ with $\varepsilon = 0$.

We assume that the following conditions are satisfied:

1. There exists a global minimizer of Problem (**P**), implying that $f(\mathbf{x})$ is bounded from below on S_0.
2. The number of distinct local minimum values of the objective function of Problem (**P**) is finite.
3. The objective function $f(\mathbf{x}) \to \infty$, as $\|\mathbf{x}\| \to \infty$, where $\|\mathbf{x}\|$ denotes the usual Euclidean norm of the vector $\mathbf{x}$.

Motivated by the exact penalty function introduced in [7] and the constraint transcription method for converting continuous inequality constraints into a sequence of inequality constraints in integral form (see [8]), we introduce a new exact penalty function $f_\sigma(\mathbf{x}, \varepsilon)$ defined below:

$$f_\sigma(\mathbf{x}, \varepsilon) = \begin{cases} f(\mathbf{x}), & \text{if } \varepsilon = 0, g_j(\mathbf{x}, \omega) \le 0 \ (\omega \in \Omega), \\ f(\mathbf{x}) - \varepsilon^{-\alpha} \log(1 - \Delta(\mathbf{x}, \varepsilon)) + \sigma \varepsilon^{\beta}, & \text{if } \varepsilon > 0, \Delta(\mathbf{x}, \varepsilon) < 1, \\ +\infty, & \text{otherwise} \end{cases} \tag{5}$$

where $\Delta(\mathbf{x}, \varepsilon)$, which is referred to as the constraint violation, is defined by

$$\Delta(\mathbf{x}, \varepsilon) = \sum_{j=1}^{m} \int_\Omega \left[\max\left\{0, g_j(\mathbf{x}, \omega) - \varepsilon^\gamma\right\} \right]^2 d\omega, \tag{6}$$

γ is a positive real number, $\beta > 2$, and $\sigma > 0$ is a penalty parameter. We now introduce a surrogate optimization problem, which is referred to as Problem ($\mathbf{P}_\sigma$), as follows:

$$\min \ f_\sigma(\mathbf{x}, \varepsilon) \tag{7a}$$

subject to

$$(\mathbf{x}, \varepsilon) \in \mathbb{R}^n \times [0, +\infty). \tag{7b}$$

Compared with the exact penalty function proposed in [21], which approximates the optimal solution from outside of the feasible region, the new exact penalty function (5) is more like a traditional barrier function. However, our method does not require to choose an initial guess to be within the feasible region of the problem. We only require to choose an initial guess close to the feasible region of the problem. Due to the structure of the logarithmic function, i.e., $\log(1 - \Delta(\mathbf{x}, \varepsilon))$, it is clear that the constraint violation of the new proposed exact penalty function has an upper bound of 1. It forces the iterates to stay within a small neighborhood of the feasible region. When the penalty parameter σ is large, the constraint violation will be forced to reduce. This means that the value of

$$\left[\max\left\{0, g_j(\mathbf{x}, \omega) - \varepsilon^\gamma\right\} \right]^2$$

must go down, and eventually, leading to the satisfaction of the continuous inequality constraints, i.e.,

$$g_j(\mathbf{x},\omega) \le 0,\ \forall\ \omega \in \Omega,\ j = 1,\ldots,m.$$

In the next section, we shall present our main theoretical results.

2.1 Convergence Analysis

Taking the gradients of $f_\sigma(\mathbf{x},\varepsilon)$ with respect to x and ε gives

$$\frac{\partial f_\sigma(\mathbf{x},\varepsilon)}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} + \frac{2\varepsilon^{-\alpha}}{1-\Delta(\mathbf{x},\varepsilon)} \sum_{j=1}^{m} \int_\Omega \max\{0, g_j(\mathbf{x},\omega) - \varepsilon^\gamma\} \frac{\partial g_j(\mathbf{x},\omega)}{\partial \mathbf{x}} d\omega \tag{8}$$

$$\frac{\partial f_\sigma(\mathbf{x},\varepsilon)}{\partial \varepsilon} = \alpha\varepsilon^{-\alpha-1}\log(1-\Delta(\mathbf{x},\varepsilon)) - \varepsilon^{-\alpha}\frac{2\gamma}{1-\Delta(\mathbf{x},\varepsilon)} \sum_{j=1}^{m} \int_\Omega \max\{0, g_j(\mathbf{x},\omega) - \varepsilon^\gamma\}\varepsilon^{\gamma-1} d\omega + \sigma\beta\varepsilon^{\beta-1} \tag{9}$$

For every positive integer k, let $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ be a local minimizer of Problem $(\mathbf{P}_{\sigma_k})$. To obtain our main result, we need

Lemma 1. *Let $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ be a local minimizer of Problem $(\mathbf{P}_{\sigma_k})$. Suppose that $f_{\sigma_k}(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ is finite and that $\varepsilon^{(k),\star} > 0$. Then*

$$(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star}) \notin S_{\varepsilon^{(k),\star}}$$

where $S_{\varepsilon^{(k),\star}}$ is defined by (3) with $\varepsilon = \varepsilon^{(k),\star}$.

Proof. The proof is similar to that given for Lemma 2.1 in [21] and hence is omitted.

To continue, we introduce

Definition 1. It is said that the constraint qualification is satisfied for the continuous inequality constraints (1b) at $\mathbf{x} = \bar{\mathbf{x}}$, if the following implication is valid. Suppose that

$$\int_\Omega \sum_j \varphi_j(\omega) \frac{\partial g_j(\bar{x},\omega)}{\partial x} d\omega = 0.$$

Then, $\varphi_j(\omega) = 0, \forall \omega \in \Omega,\ j = 1,\ldots,m.$

By Assumption (A3), the existence of an accumulating point of the sequence $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ is assured. Let the conditions of Lemma 1 be satisfied. Then, we have

Theorem 1. *Suppose that* $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ *is a local minimizer of Problem* $(\mathbf{P}_{\sigma_k})$ *such that* $f_{\sigma_k}(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ *is finite and* $\varepsilon^{(k),\star} > 0$. *If* $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star}) \to (\mathbf{x}^{\star}, \varepsilon^{\star})$ *as* $k \to +\infty$, *and the constraint qualification is satisfied for the continuous inequality constraints (1b) at* $\mathbf{x} = \mathbf{x}^{\star}$, *then* $\varepsilon^{\star} = 0$ *and* $\mathbf{x}^{\star} \in S_0$.

Proof. It follows from the conditions of the theorem that

$$\begin{aligned}
&\frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})}{\partial \mathbf{x}} \\
&= \frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} + \frac{2(\varepsilon^{(k),\star})^{-\alpha}}{1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})} \\
&\quad \sum_{j=1}^{m} \int_{\Omega} \max\{0, g_j(\mathbf{x}^{(k),\star}, \omega) - (\varepsilon^{(k),\star})^{\gamma}\} \frac{\partial g_j(\mathbf{x}^{(k),\star}, \omega)}{\partial \mathbf{x}} d\omega \\
&= 0, \qquad (10)
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial f_{\sigma_k}(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})}{\partial \varepsilon} \\
&= \alpha(\varepsilon^{(k),\star})^{-\alpha-1} \log(1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})) - (\varepsilon^{(k),\star})^{-\alpha} \frac{2\gamma}{1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})} \\
&\quad \sum_{j=1}^{m} \int_{\Omega} \max\{0, g_j(\mathbf{x}^{(k),\star}, \omega) - (\varepsilon^{(k),\star})^{\gamma}\} (\varepsilon^{(k),\star})^{\gamma-1} d\omega + \sigma_k \beta (\varepsilon^{(k),\star})^{\beta-1} \\
&= (\varepsilon^{(k),\star})^{-\alpha-1} \Big\{ \alpha \log(1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})) - \frac{2\gamma}{1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})} \\
&\quad \sum_{j=1}^{m} \int_{\Omega} \max\{0, g_j(\mathbf{x}^{(k),\star}, \omega) - (\varepsilon^{(k),\star})^{\gamma}\} (\varepsilon^{(k),\star})^{\gamma} d\omega \Big\} + \sigma_k \beta (\varepsilon^{(k),\star})^{\beta-1} \\
&= 0 \qquad (11)
\end{aligned}$$

Suppose that $\varepsilon^{(k),\star} \to \varepsilon^{\star} \neq 0$. Then, by (11), we observe that

$$\begin{aligned}
&(\varepsilon^{(k),\star})^{-\alpha-1} \Big\{ \alpha \log(1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})) - \frac{2\gamma}{1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})} \\
&\quad \sum_{j=1}^{m} \int_{\Omega} \max\{0, g_j(\mathbf{x}^{(k),\star}, \omega) - (\varepsilon^{(k),\star})^{\gamma}\} (\varepsilon^{(k),\star})^{\gamma} d\omega \Big\}
\end{aligned}$$

tends to a finite value, while $\sigma_k \beta (\varepsilon^{(k),\star})^{\beta-1}$ tends to positive infinity as $\sigma_k \to +\infty$, when $k \to +\infty$. This is impossible for the validity of (11). Thus, $\varepsilon^{\star} = 0$.

Now, by rearranging (10), we obtain

$$\begin{aligned}
&(\varepsilon^{(k),\star})^{\alpha} \big(1 - \Delta(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})\big) \frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}} \qquad (12) \\
&\quad + 2 \sum_{j=1}^{m} \int_{\Omega} \max\{0, g_j(\mathbf{x}^{(k),\star}, \omega) - (\varepsilon^{(k),\star})^{\gamma}\} \frac{\partial g_j(\mathbf{x}^{(k),\star}, \omega)}{\partial \mathbf{x}} d\omega \\
&\quad = 0. \qquad (13)
\end{aligned}$$

Thus,

$$\lim_{k\to+\infty}\left\{(\varepsilon^{(k),\star})^{\alpha}\big(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\big)\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}\right.$$
$$\left.+2\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}d\omega\right\}$$
$$=2\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{\star},\omega)\right\}\frac{\partial g_j(\mathbf{x}^{\star},\omega)}{\partial \mathbf{x}}d\omega=0. \tag{14}$$

Since the constraint qualification is satisfied for the continuous inequality constraints (1b) at $\mathbf{x}=\mathbf{x}^{\star}$, it follows that, for each $j=1,\ldots,m$,

$$\max\left\{0,g_j(\mathbf{x}^{\star},\omega)\right\}=0,$$

for each $\omega\in\Omega$. This, in turn, implies that, for each $j=1,\ldots,m$, $g_j(\mathbf{x}^{\star},\omega)\le 0$, $\forall\ \omega\in\Omega$. The proof is completed.

Corollary 1. *If* $\mathbf{x}^{(k),\star}\to\mathbf{x}^{\star}\in S_0$ *and* $\varepsilon^{(k),\star}\to\varepsilon^{\star}=0$, *then* $\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\to\Delta(\mathbf{x}^{\star},\varepsilon^{\star})=0$.

Proof. The conclusion follows readily from the definition of $\Delta(\mathbf{x},\varepsilon)$ and the continuity of $g_j(\mathbf{x},\omega)$.

The next theorem shows that, under some mild conditions, $f_{\sigma}(\mathbf{x},\omega)$ is continuously differentiable with continuous limits.

Theorem 2. *Assume that* $g_j(\mathbf{x}^{(k),\star},\omega)=o((\varepsilon^{(k),\star})^{\delta})$, $\delta>0$, $j=1,\ldots,m$. *Suppose that* $\gamma>\alpha$, $\delta>\alpha$, $-\alpha-1+2\delta>0$, $2\gamma-\alpha-1>0$. *Then*

$$f_{\sigma_k}(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\xrightarrow[\mathbf{x}^{(k),\star}\to\mathbf{x}^{\star}\in S_0]{\varepsilon^{(k),\star}\to\varepsilon^{\star}=0} f_{\sigma_k}(\mathbf{x}^{\star},0)=f(\mathbf{x}^{\star}), \tag{15}$$

$$\nabla_{(\mathbf{x},\varepsilon)}f_{\sigma_k}(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\xrightarrow[\mathbf{x}^{(k),\star}\to\mathbf{x}^{\star}\in S_0]{\varepsilon^{(k),\star}\to\varepsilon^{\star}=0}\nabla_{(\mathbf{x},\varepsilon)}f_{\sigma_k}(\mathbf{x}^{\star},0)=(\nabla f(\mathbf{x}^{\star}),0). \tag{16}$$

Proof. The proof is similar to that given for Theorem 2.4 in [21] and hence is omitted.

Theorem 3. *There exists a* $k_0>0$, *such that for any* $k\ge k_0$, *every local minimizer* $(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})$ *of the penalty problem with finite* $f_{\sigma_k}(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})$ *has the form* $(\mathbf{x}^{\star},0)$ *with* $\mathbf{x}^{\star}$ *being a local minimizer of Problem* (**P**).

Proof. On the contrary, we assume that the conclusion is false. Then, there exists a subsequence of $\{(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\}$, which is denoted by the original sequence, such that for any $k_0>0$, there exists a $k'>k_0$ satisfying $\varepsilon^{(k'),*}\neq 0$. By Theorem 1, we have

$$\varepsilon^{(k),\star} \to \varepsilon^\star = 0,\ \mathbf{x}^{(k),\star} \to \mathbf{x}^\star \in S_0,\ \text{as } k \to +\infty.$$

Since $\varepsilon^{(k),\star} \neq 0$ for all k, it follows from dividing (11) by $(\varepsilon^{(k),\star})^{\beta-1}$ that

$$\begin{aligned}(\varepsilon^{(k),\star})^{-\alpha-\beta}\Big\{\alpha\log(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))-\frac{2\gamma}{1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})}\\ \sum_{j=1}^{m}\int_\Omega \max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}(\varepsilon^{(k),\star})^\gamma d\omega\Big\}+\sigma_k\beta=0.\end{aligned} \tag{17}$$

This is equivalent to

$$\begin{aligned}&(\varepsilon^{(k),\star})^{-\alpha-\beta}\Bigg\{\alpha\log(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))\\ &+\frac{2\gamma}{1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})}\sum_{j=1}^{m}\int_\Omega\Big[\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}\big(-(\varepsilon^{(k),\star})^\gamma\big)\\ &+\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}g_j(\mathbf{x}^{(k),\star},\omega)\\ &-\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}g_j(\mathbf{x}^{(k),\star},\omega)\Big]d\omega\Bigg\}+\sigma_k\beta=0.\end{aligned} \tag{18}$$

Note that

$$\begin{aligned}&\sum_{j=1}^{m}\int_\Omega\Big[\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}\big(-(\varepsilon^{(k),\star})^\gamma\big)\\ &+\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}g_j(\mathbf{x}^{(k),\star},\omega)\Big]d\omega\\ &=\sum_{j=1}^{m}\int_\Omega\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}\big(g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\big)d\omega\\ &=\sum_{j=1}^{m}\int_\Omega\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}^2 d\omega\\ &=\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})\end{aligned} \tag{19}$$

Substitute (19) to (18) and rearranging (18) yields

$$\begin{aligned}&(\varepsilon^{(k),\star})^{-\alpha-\beta}\left\{\alpha\log(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))+\frac{2\gamma\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})}{1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})}\right\}+\sigma_k\beta\\ &=\frac{2\gamma(\varepsilon^{(k),\star})^{-\alpha-\beta}}{1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})}\sum_{j=1}^{m}\int_\Omega\max\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^\gamma\}g_j(\mathbf{x}^{(k),\star},\omega)d\omega.\end{aligned} \tag{20}$$

Letting $k \to +\infty$, it follows that the left-hand side of (20) tends to infinity, which means the right-hand side of (20) should also goes to infinity. By Theorem 3, we have that $\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}) \to 0$, as $k \to +\infty$. Thus, in view of the right-hand side of (20), we have

$$2\gamma(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}g_j(\mathbf{x}^{(k),\star},\omega)d\omega\to+\infty. \tag{21}$$

Again, by Theorem 3 and the continuity of g_j, it follows that for a sufficiently large k,

$$\begin{aligned}
&2\gamma(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}g_j(\mathbf{x}^{(k),\star},\omega)d\omega\\
&\quad\le 2\gamma(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}|g_j(\mathbf{x}^{(k),\star},\omega)|d\omega\\
&\quad\le 2\gamma(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}G_{\max}d\omega
\end{aligned} \tag{22}$$

where $G_{\max}=\max\{g_j(\mathbf{x}^{\star}),j=1,\ldots,m\}$. Define

$$y^k=(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}d\omega. \tag{23}$$

Then, by (21) and (22), we obtain

$$2\gamma G_{\max}y^k\to+\infty\ ,\ \text{as}\ k\to+\infty. \tag{24}$$

Thus, $y^k\to+\infty$, as $k\to+\infty$.

Let

$$z^k=y^k/|y^k|. \tag{25}$$

Clearly

$$\lim_{k\to+\infty}|z^k|=|z^*|=1. \tag{26}$$

Dividing (12) by $|y^k|$ yields

$$\begin{aligned}
&\frac{\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial\mathbf{x}}}{|y^k|}+\frac{2(\varepsilon^{(k),\star})^{-\alpha}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)\right.\\
&\quad\left.-(\varepsilon^{(k),\star})^{\gamma}\right\}\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial\mathbf{x}}d\omega=0.
\end{aligned} \tag{27}$$

Note that $\mathbf{x}^{(k),\star}\to\mathbf{x}^{\star}$ as $k\to+\infty$ and that $\dfrac{\partial f(\mathbf{x})}{\partial\mathbf{x}}$ and, for each $j=1,\ldots,m$, g_j and $\dfrac{\partial g_j(\cdot\,,\omega)}{\partial\mathbf{x}}$ are continuous in $\mathbb{R}^n$ for each $\omega\in\Omega$, where Ω is a compact set. Then, it can be shown that there exist constants $\hat{K}$ and $\overline{K}$, independent of k, such that, for all $k=1,2,\ldots,$

$$\left|\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}\right| \leq \hat{K}, \tag{28}$$

$$\left|\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}\right| \leq \overline{K}, \text{ for } j=1,\ldots,m. \tag{29}$$

Dividing (27) by $(\varepsilon^{(k),\star})^{\beta}$, we obtain

$$\frac{\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}}{|y^k|(\varepsilon^{(k),\star})^{\beta}} + \frac{2(\varepsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, g_j(\mathbf{x}^{(k),\star},\omega) - (\varepsilon^{(k),\star})^{\gamma}\right\}\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}d\omega = 0. \tag{30}$$

By (23), we have

$$\begin{aligned}\frac{1}{|y^k|(\varepsilon^{(k),\star})^{\beta}} &= \frac{1}{\left|(\varepsilon^{(k),\star})^{-\alpha-\beta}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, g_j(\mathbf{x}^{(k),\star},\omega) - (\varepsilon^{(k),\star})^{\gamma}\right\}d\omega\right|(\varepsilon^{(k),\star})^{\beta}} \\ &= \frac{1}{\left|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, g_j(\mathbf{x}^{(k),\star},\omega) - (\varepsilon^{(k),\star})^{\gamma}\right\}d\omega\right|(\varepsilon^{(k),\star})^{-\alpha}}.\end{aligned} \tag{31}$$

From the conditions of Theorem 2, we recall that $g_j(\mathbf{x}^{(k),\star},\omega) = o((\varepsilon^{(k),\star})^{\delta})$ and $\gamma > \alpha,\ \delta > \alpha$. Thus

$$\begin{aligned}&\lim_{k\to+\infty}\left|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, g_j(\mathbf{x}^{(k),\star},\omega) - (\varepsilon^{(k),\star})^{\gamma}\right\}d\omega\right|(\varepsilon^{(k),\star})^{-\alpha} \\ &= \lim_{k\to+\infty}\left|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, \frac{g_j(\mathbf{x}^{(k),\star},\omega)}{(\varepsilon^{(k),\star})^{\delta}}(\varepsilon^{(k),\star})^{\delta-\alpha} - (\varepsilon^{(k),\star})^{\gamma-\alpha}\right\}d\omega\right| \\ &= \lim_{k\to+\infty}\left|\sum_{j=1}^{m}\int_{\Omega}\max\left\{0, \frac{o((\varepsilon^{(k),\star})^{\delta})}{(\varepsilon^{(k),\star})^{\delta}}(\varepsilon^{(k),\star})^{\delta-\alpha} - (\varepsilon^{(k),\star})^{\gamma-\alpha}\right\}d\omega\right| \\ &= 0,\end{aligned} \tag{32}$$

and hence,

$$\lim_{k\to\infty}\frac{1}{|y^k|(\varepsilon^{(k),\star})^{\beta}} \to +\infty. \tag{33}$$

From (28) and (33), it is clear that

$$\frac{\left|\frac{\partial f(\mathbf{x}^{(k),\star})}{\partial \mathbf{x}}\right|}{|y^k|(\varepsilon^{(k),\star})^{\beta}} \to +\infty,\ \ k\to+\infty. \tag{34}$$

On the other hand,

$$\begin{aligned}
&\left|\frac{2(\varepsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}d\omega\right| \\
&\quad\leq \frac{2(\varepsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\left|\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}\right|d\omega \\
&\quad= \frac{2(\varepsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}\left|\frac{\partial g_j(\mathbf{x}^{(k),\star},\omega)}{\partial \mathbf{x}}\right|d\omega \\
&\quad\leq \frac{2(\varepsilon^{(k),\star})^{-\alpha-\beta}}{|y^k|(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}\sum_{j=1}^{m}\int_{\Omega}\max\left\{0,g_j(\mathbf{x}^{(k),\star},\omega)-(\varepsilon^{(k),\star})^{\gamma}\right\}\overline{K}d\omega \\
&\quad= \frac{2\overline{K}z^k}{(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))},
\end{aligned} \tag{35}$$

where z^k is defined by (25). Clearly, $|z^k| = 1$. On the other hand, note that $1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}) \to 1$, as $k \to +\infty$. Thus, $\frac{2\overline{K}z^k}{(1-\Delta(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star}))}$ is bounded when $k \to \infty$. This together with (34) is a contradiction to (30). This completes the first part of the proof.

For sufficiently large k, every local minimizer $(\mathbf{x}^{(k),\star},\varepsilon^{(k),\star})$ has the form $(\mathbf{x}^{\star},0)$. It is obvious from Theorem 1 that $\mathbf{x}^{\star}$ is a feasible point of Problem $(\mathbf{P})$. This indicates that there is a neighborhood of $\mathbf{x}^{\star}$, such that for any feasible x of Problem $(\mathbf{P})$

$$f(\mathbf{x}) = f_{\sigma_k}(\mathbf{x},0) \geq f_{\sigma_k}(\mathbf{x}^{\star},0) = f(\mathbf{x}^{\star}).$$

Therefore, $\mathbf{x}^{\star}$ is a local minimizer of Problem $(\mathbf{P})$. This completes the proof.

We may now conclude that, under some mild assumptions and the constraint qualification condition, when the parameter σ is sufficiently large, a local minimizer of Problem $(\mathbf{P}_{\sigma})$ is a local minimizer of Problem $(\mathbf{P})$.

Based on the results obtained in Theorem 1, Corollary 1, Theorems 2, and 3 we are in a position to present an effective computational method in the next section.

3 Algorithm and Numerical Results

To show the effectiveness of the proposed method, we consider three examples. The optimization tool box *fmincon* within MATLAB environment is used to solve the optimization Problem $(\mathbf{P}_{\sigma})$, where the integral appeared in $f_{\sigma}(\mathbf{x},\varepsilon)$ is calculated by using the *Simpson's Rule* with a discretization step size h. For *Simpson's Rule*, the global error is of order h^4. Thus, by choosing a sufficiently small h, the required accuracy of the integrations can be achieved.

Let $\sigma^\star$ be the upper bound of the penalty parameter, and $\varepsilon^\star$ be the lower bound of ε. Based on the proposed new exact penalty function, an efficient algorithm for solving Problem $(\mathbf{P}_\sigma)$ is given below:

Algorithm 1

Step 1:

Set $\sigma^{(0)} = 1$, $\varepsilon^{(0)} = 0.1$, $\sigma^\star = 10^5$ $\varepsilon^\star = 10^{-9}$, choose an initial point $(\mathbf{x}_0, \varepsilon_0)$. Set the iteration index $k = 0$. Choose appropriate values of β, γ and α. Note that their choices depend on the specific structure of Problem $(\mathbf{P})$ concerned.

Step 2:

Solve Problem $(\mathbf{P}_{\sigma_k})$, and let $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ be the minimizer obtained.

Step 3:

If $\varepsilon^{(k),\star} > \varepsilon^\star$, $\sigma^{(k)} < \sigma^\star$, set $\sigma^{(k+1)} = 10 \times \sigma^{(k)}$, $k := k+1$. Go to **Step 2** with $(\mathbf{x}^{(k),\star}, \varepsilon^{(k),\star})$ taken as the new initial point for the new optimization process
Else set $\varepsilon^{(k),\star} := \varepsilon^\star$, then go to **Step 4**

Step 4:

Check the feasibility of $\mathbf{x}^{(k),\star}$. **If** $\mathbf{x}^{(k),\star}$ is feasible, then it is a local minimizer of Problem $(\mathbf{P})$. **Else** go to **Step 5**

Step 5:

Adjust the parameters α, β and γ such that conditions of Lemma 1 are satisfied. Set $k := 0$. Go to **Step 2**.

Note that, in **Step 4**, it is impossible to check the feasibility of $g_j(\mathbf{x}, \omega) \le 0$, $j = 1, \ldots, m$, for every $\omega \in \Omega$. In practice, we choose a set $\hat{\Omega}$, which contains a dense enough of points in Ω. Then, the feasibility of $g_j(\mathbf{x}, \omega) \le 0$ is checked over $\hat{\Omega}$ for each $j = 1, \ldots, m$.

Example 1. The following example is taken from [4]. It is also used in [16, 17, 20, 21] to test the effectiveness of their algorithms. In this problem, the objective function:

$$f(\mathbf{x}) = \frac{x_2(122 + 17x_1 + 6x_3 - 5x_2 + x_1x_3) + 180x_3 - 36x_1 + 1224}{x_2(408 + 56x_1 - 50x_2 + 60x_3 + 10x_1x_3 - 2x_1^2)} \tag{36}$$

is to be minimized subject to

$$\phi(\mathbf{x}, \omega) \le 0\ ,\ \forall\ \omega \in \Omega, \tag{37}$$

$$0 \le x_1, x_3 \le 100,\ 0.1 \le x_2 \le 100, \tag{38}$$

where $\Omega = [10^{-6}, 30]$ and $(i = \sqrt{-1})$, while

$$\phi(\mathbf{x}, \omega) = \Im\big(T(\mathbf{x}, \omega)\big) - 3.33[\Re\big(T(\mathbf{x}, \omega)\big)]^2 + 1.0,$$
$$T(\mathbf{x}, \omega) = 1 + \frac{x_1 + \frac{x_2}{i\omega} + i\omega x_3}{(i\omega + 3)(-\omega^2 + 2i\omega + 2)}.$$

Here, $\Im\big(T(\mathbf{x}, \omega)\big)$ and $\Re\big(T(\mathbf{x}, \omega)\big)$ are, respectively, the imaginary and real parts of $T(\mathbf{x}, \omega)$. The initial point is $[50, 50, 50]^\top$, and we choose α, β, and γ to be 1.7, 2.2, and 3, respectively. *Simpson's Rule* with $\Omega = [10^{-6}, 30]$ being divided into 3,000 equal subintervals is used to evaluate the integral. Moreover, the required dense subset $\hat{\Omega}$ of Ω is taken to be the set which contains all these discretized points.

By applying Algorithm 1, the solution obtained is listed below:

$$\mathbf{x}^\star = [16.9328,\ 45.4605,\ 34.6875]^\top$$

The corresponding cost is $f(\mathbf{x}^\star) = 0.174627$ and the maximum value of the continuous inequality constraint on $[10^{-6}, 30]$ is -6.05×10^{-6}, meaning that the solution is feasible in regard to the continuous inequality constraint. Note that this solution is slightly better than the solution reported in [21], where a cost of 0.174778 is obtained.

Example 2. Consider the following semi-infinite optimization problem:

$$\begin{array}{ll} \min & x_1^2 + (x_2 - 3)^2 \\ \text{subject to} & x_2 - 2 + x_1 \sin(\frac{t}{x_2 - \omega}) \le 0,\ \forall\, t \in [0, \pi] \\ & -1 \le x_1 \le 1,\ 0 \le x_2 \le 3. \end{array}$$

where ω is a parameter which is chosen to be 2.032 as in [17]

Simpson's Rule with interval $[0, \pi]$ being divided into 1,000 equal subintervals is used to evaluate the integral. These discretized points also form a dense subset $\hat{\Omega}$ of the interval $[0, \pi]$. The feasibility check is carried over $\hat{\Omega}$. By using Algorithm 1 with the initial point taken as $(0.5, 0.5)$, the solution obtained is $(x_1^\star, x_2^\star) = [0, 2]^\top$ with the cost value $f^\star = 1$. This cost value is the same as the one obtained in [21]. The maximum value of the continuous inequality constraint on $[0, \pi]$ is 0, meaning that the solution is feasible in regard to the continuous inequality constraint.

Example 3. Consider the following semi-infinite optimization problem:

$$\begin{array}{ll} \min & (x_1 + x_2 - 2)^2 + (x_1 - x_2)^2 + 30[\min\{0, x_1 - x_2\}]^2 \\ \text{subject to} & x_1 \cos t + x_2 \sin t - 1 \le 0,\ \forall\, t \in [0, \pi]. \end{array}$$

Again, *Simpson's Rule* with the interval $[0, \pi]$ being partitioned into 1,000 equal subintervals is used to evaluate the corresponding constraint violation in the exact penalty function. These discretized points also form the required dense subset $\hat{\Omega}$ of the interval $[0, \pi]$. The check of the feasibility of the continuous inequality constraint is carried out over $\hat{\Omega}$. Now, by using Algorithm 1 with the initial point taken as

$[0.5, 0.5]^\top$, we obtain the solution $\mathbf{x} = [0.707100, 0.707114]^\top$. The corresponding cost value is 0.343145. This is also slightly better than the result reported in [21], where the cost value obtained is 0.34325. Furthermore, the solution obtained by Algorithm 1 is a feasible point (the maximum value of the continuous inequality constraint on $[0, \pi]$ is -4.80×10^{-7}).

4 Conclusions

In this paper, a barrier-like penalty function is introduced to develop an alternative effective computational method for solving a class of SIPs. This computational algorithm forces the iterates to stay in a small neighborhood of the feasible region. There is no need to find an interior point to start with. Any local minimizer of the penalized optimization problems is also a local minimizer of the original semi-infinite optimization problem when the penalty parameter is sufficiently large. The numerical results indicate that the proposed exact penalty method is effective when compared with other existing methods.

References

1. Brosowski, B.: Parametric Semi-infinite Optimization. Verlag Peter Lang, Frankfurt (1982)
2. Dall'Aglio, M.: On some applications of lsip to probability and statistics. In: Goberna, M.A., L*ó*pez, M.A. (eds.) Semi-infinite Programming, Recent Advances, Nonconvex Optimization and Its Applications, **57**, Kluwer, Dordrecht (2001)
3. Glashoff, K., Gustafson, S.A.: Linear Optimization and Approximation, Springer Singapore Pte. Limited, (1983)
4. Gonzaga, C., Polak, E., Trahan, R.: An improved algorithm for optimization problems with functional inequality constraints. IEEE Trans. Autom. Control **25**(1), 49–54 (1980)
5. Hettich, R. (ed.): Semi-infinite Programming. Lecture Notes in Control and Information Sciences. Springer, Berlin (1979)
6. Hettich, R., Kortanek, K.O.: Semi-infinite programming: Theory, methods, and applications. SIAM Rev. **35**(3), 380–429 (1993)
7. Huyer, W., Neumaier, A.: A new exact penalty function. SIAM J. Optim. **13**(4), 1141–1158 (2003)
8. Jennings, L.S., Teo, K.L.: A computational algorithm for functional inequality constrained optimization problems. Automatica **26**, 371–375 (1990)
9. Kortanek, K.O., Fiacco, A.V. (eds.): Semi-Infinite Programming and Applications. Lecture Notes in Economics and Mathematical Systems, **215**, Springer-Verlag (1983)
10. L*ó*pez, M.A., Goberna, M.A.: Linear Semi-infinite Optimization. Wiley, Chichester (1998)
11. L*ó*pez, M., Still, G.: Semi-infinite programming. Eur. J. Oper. Res. **180**, 491–518 (2006)
12. Liu, Y., Teo, K.L., Wu, S.Y.: A new quadratic semi-infinite programming algorithm based on dual parametrization. J. Glob. Optim. **29**, 401–413 (2004)
13. Polak, E., Mayne, D.Q.: An algorithm for optimization problems with functional inequality constraints. IEEE Trans. Autom. Control **21**, 184–193 (1976)
14. Polak, E., Pister, K.S., Ray, D.: Optimal design framed structures subjected to earthquak. Eng. Optim. **12**, 65–71 (1976)

15. Polak, E., Mayne, D.Q., Stimler, D.M.: Control system design via semi-infinite optimization: A review. Proc. IEEE **72**(12), 1777–1794 (1984)
16. Teo, K.L., Rehbock, V., Jennings, L.S.: A new computational algorithm for functional inequality constrained optimization problems. Automatica **29**, 789–792 (1993)
17. Teo, K.L., Yang, X.Q., Jennings, L.S.: Computational discretization algorithms for functional inequality constrained optimization. Ann. Oper. Res. **98**, 215–234 (2000)
18. Wu, S.Y., Fang, S.C., Lin, C.J.: Relaxed cutting plane method for solving linear semi-infinite programming problems. J. Optim. Theory Appl. **99**(3), 759–779 (1998)
19. Wu, S.Y., Li, D.H., Qi, L.Q., Zhou, G.L.: An iterative method for solving KKT system of the semi-infinite programming. Optim. Methods Softw. **20**(6), 629–643 (2005)
20. Yang, X.Q., Teo, K.L.: Nonlinear Lagrangian functions and applications to semi-infinite programs. Ann. Oper. Res. **103**, 235–250 (2001)
21. Yu, C.J., Teo, K.L., Zhang, L.S., Bai, Y.Q.: A new exact penalty function method for continuous inequality constrained optimization problems. J. Ind. Manag. Optim. **8**(2), 485–491 (2010)
22. Yu, C.J., Teo, K.L., Zhang, L.S., Bai, Y.Q.: On a refinement of the convergence analysis for the new exact penalty function method for continuous inequality constrained optimization problem. J. Ind. Manag. Optim. **6**(4), 895–910 (2012)

On the Statistical Models-Based Multi-objective Optimization

Antanas Žilinskas

1 Introduction

Nonlinear multi-objective optimization is a very active research area. Depending on the properties of a multi-objective optimization problem, different approaches to its solution can be applied. The best direction developed is optimization of convex problems; for the latter problems, the methods that generalize the ideas of classical mathematical programming suit well [2, 14, 32]. For the problems with the objectives not satisfying the assumption of convexity, metaheuristic methods are frequently favorable [1, 3, 7, 16]. However, there remains a class of important problems without sufficient attention of researchers: the problems with black-box, multimodal, and expensive objectives. In the present paper, namely those problems with black-box expensive objective functions are considered. The construction of algorithms for such problems is difficult even at the conceptual level because of scarce black-box information and the expensiveness of objective functions. The latter factor restricts eliciting of the desirable information. The rational decision theory and statistical models of uncertainty seem well appropriate to tackle such problems. A new approach to constructing global optimization algorithms, induced by the rational decision theory, is proposed. We postulate the properties to be satisfied by a rational decision concerning the current optimization step. As shown in [4, 28] such properties are inherent for several well-known single-objective optimization algorithms, e.g. for the P-algorithm [27]. Those properties also facilitate the implementation of the respective algorithms in the arithmetic of infinity [18, 19]. The proposed approach, from a new more general perspective, substantiates the single-objective P-algorithm. For the multi-objective optimization, this approach not only constitutes a new more general substantiation of the known algorithms

A. Žilinskas (✉)
Institute of Mathematics and Informatics, Vilnius University,
Akademijos 4, 08663 Vilnius, Lithuania
e-mail: antanas.zilinskas@mii.vu.lt

T.M. Rassias et al. (eds.), *Optimization in Science and Engineering: In Honor of the 60th Birthday of Panos M. Pardalos*, DOI 10.1007/978-1-4939-0808-0_29,

but also facilitates construction of a family of algorithms, similar in a sense to the multi-objective P-algorithm. The paper is completed with several numerical examples which illustrate the performance of an algorithm constructed according to the proposed ideas.

This paper is devoted to commemorate the 60th anniversary of Professor Panos M. Pardalos.

2 On Statistical Models in Single-Objective Global Optimization

Global optimization (GO) of non-convex functions is a challenging problem. The development of single-objective GO algorithms for some sub-classes of non-convex functions is facilitated by the exploitation of analytical properties of objectives [9]. However, in some applications there occur optimization problems where objectives are available either as a complicated computational model or as unfamiliar software. We focus on the problems where objective functions are expensive because of the complexity of the computational model; expensiveness here means a long-lasting computation of a value of the objective function. The complexity of the computational model normally implies not only the expensiveness of the objective function but also the uncertainty in its properties. The black-box optimization of expensive functions in many respects is quite opposite to the optimization of objective functions defined by analytical formulae. The limitation in collecting general information about the function, and particulary about its minima, strongly requires the rationality in distribution of the points where to compute the objective function values. Therefore the algorithms, founded on the principles of rational decision theory, here are of special interest. To construct such algorithms in the single-objective optimization case, statistical models of multimodal functions have proved very helpful [15, 20, 21, 23].

The global minimization problem $\min_{x\in A} f(x)$, $\mathbf{A} \subset R^d$, is considered, where $f(\cdot)$ is a continuous function, and A is a compact set. The concept of black-box optimization includes the assumption on the uncertainty in properties of $f(x)$, e.g. such an assumption is natural in applied problems where only the values of an objective function are available computed by an unfamiliar software. Besides the continuity, other analytical properties of $f(x)$ cannot be substantiated. By the expensiveness it is supposed that the long-lasting computations are needed to evaluate a single value of $f(x)$. Such unfavorable, from the optimization point of view, properties of $f(x)$ as non-differentiability, non-convexity, and multimodality cannot be excluded. To justify the search strategy in the described situation of uncertainty, a "rational optimizer" should define a model of uncertainty, e.g. to choose a statistical model of uncertainty as it is justified in the expected utility theory [6]. We focus on the statistical models of uncertainty, although other models such as fuzzy logic and rough sets would also be interesting to investigate.

Let us consider the current minimization step, where n function values have been computed at the previous steps: $y_i = f(x_i)$, $i = 1,\ldots,n$. A rational choice of

a point for the next computation of the objective function value cannot be performed without the assessment of the uncertainty in the result of the computation. The only objective information on $f(\cdot)$ is $x_i, y_i, i = 1, \ldots, n$. Besides that objective information, normally some subjective information is available, e.g. the experience of solution of similar problems in the past. As shown in [26], very general assumptions on the rational perception of uncertainty imply a random variable model for the objective function value to be computed, i.e. those assumptions imply a random variable ξ_x as a model of $f(x), x \neq x_i, i = 1, \ldots, n$. We refer to [21, 23] for the bottom-up construction of a computational statistical model of objective functions, where the mentioned above result of the existence of a statistical model in the form of a random variable has been augmented by constructive details. Such a construction of the statistical model is more advantageous than selection of the known stochastic function for a model of objective functions. In the latter case, $f(x)$ is also considered as a random variable, but its distribution is complicated to compute since the formulae of conditional probability should be applied here which are rather complicated from the computational point of view. That is true even for Gaussian stochastic functions with a notable exception of Gaussian–Markovian stochastic processes the conditional distribution of which is defined by simple formulas. Therefore such stochastic processes are attractive models for the construction of one-variable global optimization. The Wiener process was the first stochastic process model successfully used for constructing the one-variable global optimization algorithms [13, 24, 25]. However, similar simple cases are not known for $d \geq 2$. The bottom-up construction gives more flexibility in a definition of the computationally simple statistical model, i.e. to define parameters of the distribution of ξ_x which are more simply computable than in the case of Gaussian stochastic function.

As mentioned above, the first global optimization algorithm, based on a stochastic function model, was proposed in [13]. That one-variable algorithm was constructed using the Wiener process for a model. The current point for computing the objective function value is chosen to maximize the probability that this function value falls below a certain level y^{on}:

$$x_{n+1} = \arg\max_{x \in A} \mathbf{P}\{\xi(x) \leq y^{\text{on}} \,|\, \xi(x_1) = y_1, \ldots, \xi(x_n) = y_n\}, \tag{1}$$

where it is supposed that $y^{\text{on}} < \min_{1 \leq i \leq n} y_i$. That algorithm was substantiated axiomatically in [27], where it was named as the P-algorithm. For the generalization of the P-algorithm to the multidimensional ($d > 1$) case, its theoretical analysis, and applications we refer to [23]. The so-called Bayesian methods, where the idea of minimization of the average error is implemented, are presented in detail in [15]. An other well-established direction in the development of statistical model-based global optimization algorithms is rooted in the information theory [20]. In the present paper, we propose a new idea for substantiation of a statistical model-based global optimization algorithm.

Let us consider the choice of a point for the current computation of the objective function value. Such a choice in the black-box situation is a decision under uncertainty, and the rational decision theory [6] can be applied to make the choice

rationally. The theory suggests to make a decision by maximizing the average utility. To compute the latter a statistical model of uncertainty is needed as well as a utility function. The axioms in [26] substantiate the acceptance of a random variable as a model of uncertainty for an unknown value of the objective function. Accordingly, a family of random variables ξ_x is acceptable as a statistical model of the objective function. A utility function corresponding to the conception of global optimization is proposed in [27]. These results substantiate the P-algorithm, i.e. to choose the point of current computation of the objective function value, where the probability of improvement is maximal. To implement the algorithm, the family of random variables ξ_x should be defined constructively, and generally the Gaussian distribution is used to describe ξ_x. In the present paper, we construct an algorithm bypassing the necessity to define the utility function and the distribution of ξ_x.

Any characterization of a random variable normally includes a location parameter (e.g., mean) and a spread parameter (e.g., standard deviation); we use a minimal description of ξ_x by these two parameters denoted by $m(x)$ and $s(x)$. The dependence of both parameters on the information available at the current optimization step $(x_i, y_i, i = 1,\dots,n)$ will be included into the notation where needed. Assume that the utility $u_{n+1}(x)$ of computation of the current objective function value at the point x depends on x via $m(x)$ and $s(x)$. The value of $f(\cdot)$ desired to achieve y^{on}, $y^{\mathrm{on}} < \min_{1\le i\le n} y_i$, is also assumed as a parameter which defines $u_{n+1}(x)$

$$u_{n+1}(x) = U(m(x), s(x), y^{\mathrm{on}}), \tag{2}$$

and the point of the current computation is defined as the maximizer of $u_{n+1}(x)$. The following assumptions on $U(\cdot)$ express the rationality of invariance of the utility with respect to the scales of the objective function values:

$$\begin{aligned} U(m(x)+c, s(x), y^{\mathrm{on}}+c) &= U(m(x), s(x), y^{\mathrm{on}}), \\ U(m(x)\cdot C, s(x)\cdot C, y^{\mathrm{on}}\cdot C) &= U(m(x), s(x), y^{\mathrm{on}}), C>0. \end{aligned} \tag{3}$$

Since the minimization problem is considered, it is desirable to find a possibly small objective function value at every iteration; therefore, we postulate that

$$m < \mu \text{ implies } U(m,s,y) > U(\mu,s,y). \tag{4}$$

The postulated properties are inherent for several well-known optimization algorithms as shown in [4, 28].

Theorem 1. *The function that satisfies assumptions (3) is of the following structure*

$$U(m(x), s(x), y^{on}) = P\left(\frac{y^{on} - m(x)}{s(x)}\right). \tag{5}$$

Moreover, if assumption (4) is satisfied, then $P(\cdot)$ is an increasing function.

Proof. The substitution of $-y^{\mathrm{on}}$ for c in the equality $U(m(x)+c, s(x), y^{\mathrm{on}}+c) = U(m(x), s(x), y^{\mathrm{on}})$ results in

$$U(m(x),s(x),y^{\text{on}}) = U(m(x)-y^{\text{on}},s(x),0). \tag{6}$$

The substitution of $1/s(x)$ for C in (6) and the second equality in (3) yield the following equality:

$$U(m(x),s(x),y^{\text{on}}) = U\left(\frac{m(x)-y^{\text{on}}}{s(x)},1,0\right). \tag{7}$$

Now, equality (5) is obtained simply by denoting $P(z) = U(-z,1,0)$. Assumption (4) obviously implies that $P(z)$ is an increasing function of z.

The theorem substantiates the construction of the global optimization algorithm the $n+1$ iteration of which is defined by the solution of the following maximization problem

$$x_{n+1} = \max_{x\in A} P\left(\frac{y^{\text{on}}-m(x|x_i,y_i,\,i=1,\ldots,n)}{s(x|x_i,y_i,\,i=1,\ldots,n)}\right). \tag{8}$$

Formula (8) is coincident with that derived in [27], where the assumption on the Gaussian distribution of $\xi(x)$ is made. In that case, $P(\cdot)$ is a cumulative distribution function of Gaussian distribution, and the utility of the current computation is interpreted as the improvement probability. For single-objective optimization the theorem generalizes the known algorithm. For multi-objective optimization such a generalization is more important since it provides more flexibility in constructing relevant algorithms.

3 Multi-objective Optimization Based on Statistical Models

Recently several papers have been published which propose multi-objective optimization algorithms that generalize single-objective optimization algorithms based on statistical models of objective functions [5, 10–12, 16, 22, 30]. The numerical results included there show the relevance of the proposed algorithms to the problems of multi-objective optimization with black-box expensive objectives. We propose here a new idea for constructing relevant algorithms.

A multi-objective minimization problem can be stated almost identically to a single-objective problem,

$$\min_{x\in A} F(x),\ F(x) = (f_1(x),f_2(x),\ldots,f_r(x))^T,\ \mathbf{A}\subset\mathbf{R}^d, \tag{9}$$

however, the concept of solution in this case is more complicated. For the definitions of the solution to a multi-objective optimization problem with nonlinear objectives we refer to [14].

Generally speaking, the solution to a multi-objective optimization problem can be described as a set of objective vectors which well represents either the set of Pareto optimal solutions or its favorable subset; for a rigorous analysis, we refer

to [17]. The following two cases can be pointed as extreme: the approximation (discrete representation) of the whole Pareto optimal set, and an objective vector sufficiently close to a desirable one. In the real-world applications, the notion of solution can change. Frequently, in the starting optimization phase, a rough approximation of the whole Pareto optimal set is of interest; in the intermediate phase, a subset of the Pareto optimal set of interest is intended to be approximated more precisely; finally, a specific Pareto optimal solution is sought. A similar strategy is also justified in single-objective global optimization: starting from a uniform search over the feasible region, concentrating the search in prospective subregions, and finishing with a local algorithm chosen according to the local properties of the objective function.

In the case of multi-objective optimization, a vector objective function $F(x) = (f_1(x), f_2(x), \ldots, f_r(x))^T$ is considered. The same arguments, as in the case of single-objective optimization, corroborate the applicability of statistical models. The assumptions on black-box information and expense of the objective functions together with the standard assumptions of rational decision making imply the acceptability of a family of random vectors $\Xi(x) = (\xi_1(x), \ldots, \xi_r(x))^T$, $x \in \mathbf{A}$, as a statistical model of $F(x)$. Similarly, the location and spread parameters of $\xi_i(x)$, denoted by $m_i(x)$, $s_i(x)$, $i = 1, \ldots, r$, are essential in the characterization of $\xi_i(x)$. For a more specific characterization of $\Xi(x)$, e.g. by a multidimensional distribution of $\Xi(x)$, the available information usually is insufficient. If the information on, e.g. correlation between $\xi_i(x)$ and $\xi_j(x)$ were available, the covariance matrix could be included into the statistical model. However, in the present paper we assume that the objectives are independent, and the spread parameters are represented by a diagonal matrix $\Sigma(x)$ which diagonal elements are equal to $s_1, \ldots, s_r$. Similarly to the case of single-objective optimization we assume that the utility of choice of the point for the current computation of the vector value $F(x)$ has the following structure

$$u_{n+1}(x) = U(m(x), \Sigma(x), y^{\text{on}}), \tag{10}$$

where $m(x) = (m_1(x), \ldots, m_r(x))^T$, and y^{on} denotes a vector desired to improve.

At the current optimization step a point for computing the value of $F(x)$ is sought by an optimization algorithm which maximizes $u_{n+1}(x)$ which should be invariant with respect to the scales of data. Such a rationality assumption can be expressed by the following properties of $U(\cdot)$:

$$\begin{aligned} U(m(x) + c, \Sigma(x), y^{\text{on}} + c) &= U(m(x), \Sigma(x), y^{\text{on}}), c = (c_1, \ldots, c_r)^T, \\ U(C \cdot m(x), C \cdot \Sigma(x), C \cdot y^{\text{on}}) &= U(m(x), \Sigma(x), y^{\text{on}}), C_i > O, \\ C &= \begin{pmatrix} C_1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & C_r \end{pmatrix}. \end{aligned} \tag{11}$$

Since the minimization problem is considered, it is desirable to find a vector objective with possibly small values at every iteration; therefore, we postulate that for $\mu = (\mu_1, \ldots, \mu_r)^T$, where $\mu_i \geq m_i$, $i = 1, \ldots, r$ and at least one inequality is strict, the following inequality is valid

$$U(m,\Sigma,y) > U(\mu,\Sigma,y). \tag{12}$$

Theorem 2. *The function that satisfies assumptions (11) is of the following structure:*

$$U(m(x),\Sigma(x),y^{on}) = \pi\left(\frac{y_1^{on}-m_1(x)}{s_1(x)},\ldots,\frac{y_r^{on}-m_r(x)}{s_r(x)}\right). \tag{13}$$

Moreover, if assumption (12) is satisfied, then $P(\cdot)$ is an increasing function of all variables.

Proof. The proof repeats the main steps of the proof of Theorem 1 replacing the operations with scalar variables, where necessary, by the operations with vectors/matrices.

The substitution of $-y^{\text{on}}$ for c in the equality $U(m(x)+c,\Sigma(x),y^{\text{on}}+c) = U(m(x),\Sigma(x),y^{\text{on}})$ results in

$$U(m(x),\Sigma(x),y^{\text{on}}) = U(m(x)-y^{\text{on}},\Sigma(x),0). \tag{14}$$

The substitution of $\Sigma(x)^{-1}$ for C in (14) and the second equality in (3) gives the following equality:

$$U(m(x),\Sigma(x),y^{\text{on}}) = U\left(\Sigma(x)^{-1}\cdot(m(x)-y^{\text{on}}),I,0\right). \tag{15}$$

Now, equality (13) is obtained simply by denoting $\pi(z_1,\ldots,z_r) = U(-z,I,0)$. Assumption (12) obviously implies that $\pi(z)$ is an increasing function of z_i.

Theorem 2 states that a rational choice of a point for the current computation of objectives is the maximization problem of aggregated objectives $(y_i^{\text{on}}-m_i(x))/s_i(x)$. Such a conclusion is not surprising since the implementation of optimal, in some sense, single-objective optimization algorithms normally involves the optimization of an auxiliary function which formalizes the concept of optimality. The previously developed multi-objective P-algorithm [30] uses a special case of scalarization, where $P(\cdot)$ means a probability that the Gaussian random vector $\Xi(x)$ dominates y^{on}:

$$P\left(\frac{y_1^{\text{on}}-m_1(x)}{s_1(x)},\ldots,\frac{y_r^{\text{on}}-m_r(x)}{s_r(x)}\right) = \prod_{i=1}^{r}\Phi\left(\frac{y_i^{\text{on}}-m_i(x)}{s_i(x)}\right),$$

$$\Phi(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)dt. \tag{16}$$

The substantiation of rationality of various scalarizations opens a broad potentiality of the development of multi-objective optimization algorithms based on statistical models of objective functions. However, the investigation of compatibility of a priori information on the properties of objective functions with particular scalarization methods is needed to realize the mentioned potentiality. The newly proposed algorithm is called π-algorithm to show its close relationship with the

earlier developed P-algorithm; a Greek letter is used since this algorithm is proposed in the paper dedicated to commemorate the 60th anniversary of the famous Greek, the leader of global optimization, Professor Panos M. Pardalos.

As an example, the bi-objective π-algorithm has been implemented. The statistical model described in [23, pp. 158–159], has been used. A product of two arctangents was used for $\pi(\cdot)$. Then the $n+1$ step of the π-algorithm is defined as the following optimization problem

$$x_{n+1} = \arg\max_{x\in\mathbf{A}} \arctan\left(\frac{y_1^{\text{on}} - m_1(x)}{s_1(x)} + \frac{\pi}{2}\right) \cdot \arctan\left(\frac{y_2^{\text{on}} - m_2(x)}{s_2(x)} + \frac{\pi}{2}\right), \quad (17)$$

where the information collected at previous steps is taken into account when computing $m_i(x) = m_i(x|x_j, y_j,\ j = 1,\ldots,n)$ and $s_i(x) = s_i(x|x_j, y_j,\ j = 1,\ldots,n)$. The maximization in (17) was performed by a simple version of multistart: from the best of 1,000 points, generated randomly with uniform distribution over the feasible region, a local descent was performed using the codes from the MATLAB Optimization Toolbox. By this implementation we wanted to check whether the function $\arctan(\cdot)\cdot\arctan(\cdot)$ chosen rather arbitrarily could be as good as the Gaussian cumulative distribution function for constructing statistical model-based multi-objective optimization algorithms. The experimentation with this version of the algorithm can be helpful also in selecting the most appropriate statistical model for a further development where two alternatives seem competitive: a Gaussian random field versus a statistical model, based on the assumptions of subjective probability [21].

4 Experimental Results

The algorithm proposed in the present paper was implemented in MATLAB, and some experiments have been done for its comparison with the multi-objective P-algorithm, constructed using a homogeneous isotropic random field for the statistical model [30]. Also the optimization results from [30], obtained by a uniform random search, are included to highlight the properties of the selected test problems. The results obtained by a multi-objective genetic algorithm (the MATLAB implementation in [8]) are also provided for the comparison. Two examples are presented and commented; we think that extensive competitive testing would be premature, as argued in [30].

Two bi-objective test problems of two variables are chosen: the first multi-objective test problem is composed using typical test functions for a single-objective global optimization, and the second one is chosen from the set of functions, frequently used for testing multi-objective algorithms. The first vector function is composed of two Shekel functions:

$$f_1(X) = -\frac{1}{0.1+(x_1-0.1)^2+2(x_2-0.1)^2} - \frac{1}{0.14+20(x_1-0.45)^2+(x_2-0.55)^2},$$

$$f_2(X) = -\frac{1}{0.15+40(x_1-0.55)^2+(x_2-0.45)^2} - \frac{1}{0.1+(x_1-0.3)^2+(x_2-0.95)^2},$$
$$0 \le x_i \le 1,\ i = 1,2. \tag{18}$$

Shekel test functions are frequently used for testing single-objective global optimization algorithms [21]. For the representation of both objective functions by contour lines as well as for the drawing of the feasible objective region we refer to [30]. The second problem used was proposed in [7]; see also [3, pp. 339–340]. We present below its definition for an arbitrary dimension of the decision variables d:

$$f_1(X) = 1-\exp\left(-\sum_{i=1}^{d}(x_i - 1/\sqrt{n})^2\right),$$
$$f_2(X) = 1-\exp\left(-\sum_{i=1}^{d}(x_i + 1/\sqrt{n})^2\right),$$
$$-4 \le x_i \le 4,\ i = 1,\ldots,d. \tag{19}$$

Both problems are constructed using the objective functions which are hard from the single-objective global optimization point of view: their response surface is rather flat over a large part of the feasible decision region, and the minima have the form of sharp spikes. The worst case problems of multi-objective optimization are also of this type [29].

Since the considered approach is oriented to "expensive" problems, we are interested in the quality of the result obtained computing a modest number of the values of objectives. Following the concept of experimentation in [30], a termination condition of all the considered algorithms was defined by the maximum number of computations of the objective function values, equal to 100. The parameters of the statistical model, needed by the π-algorithm, have been estimated using a sample of $F(x)$ values, chosen similarly to the experiments in [30]: the sites for the first 50 computations of $F(x)$ were chosen randomly with a uniform distribution over the feasible region; the obtained data were used not only for estimating parameters but also in planning of the next 50 observations according to (17).

An important parameter of the π-algorithm is y^{on}. The vector y^{on} should be not dominated by the known values $y_1,\ldots,y_n$. A heuristic recommendation is to select y^{on} at a possibly symmetric site with respect to the global minima of objectives. We have selected the values of y^{on} used in [30]: $y^{\text{on}} = (-0.6,-0.6)$ in the case of problem (18), and $y^{\text{on}} = (0.6,0.6)$ in the case of problem (19). Typical results are illustrated in Fig. 1.

Several metrics have been proposed in recent publications for the comparison of multi-objective algorithms; for the comprehensive list and discussion, we refer to [3]. Generally speaking, it is aimed to assess the quality of approximations of the Pareto set and the efficiency of algorithms, used to compute these approximations. In the present paper, we consider only the approximation quality. Besides the results of experimentation with the π-algorithm, the results of the P-algorithm, the uniform random search, and the genetic algorithm are presented for the comparison. Since all

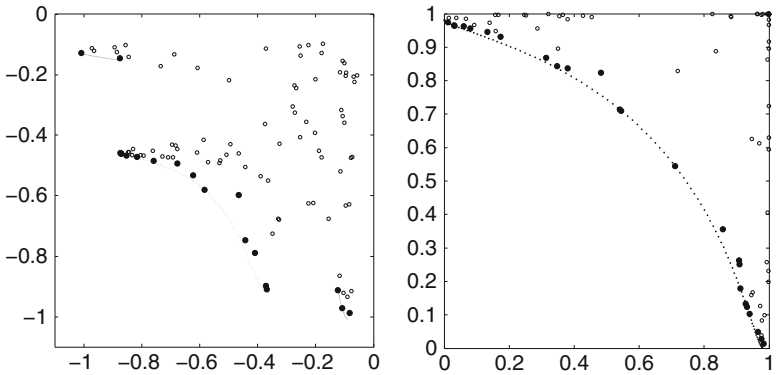


Fig. 1 The points generated by the π-algorithm in the objective feasible region of problem (18) on the *left side*, and of problem (19) on the *right side*. Non-dominated solutions are denoted by *thicker points*. A *line* indicates the Pareto set

the considered algorithms are randomized, the statistical estimates of the considered metrics are presented. The results of the P-algorithm are taken from [30]. The results of the π-algorithm are obtained as the averages from 200 independent runs.

The following metrics, used for the quantitative assessment of the precision of Pareto set approximation, have been evaluated: the number of nondominated solutions found (NN), the generational distance (GD), and the epsilon indicator (EI). GD is used to estimate, how close to the Pareto set are the found non-dominated points; the role of this metric in testing of multi-objective optimization algorithms is discussed, e.g., in [3]. GD is computed as the maximum of distances between the found non-dominated solutions and their closest neighbors from the Pareto set. EI is a metric suggested in [31] which integrates the measures of approximation precision and spread: it is the max–min distance between the Pareto set and the set of the found non-dominated solutions

$$\mathrm{EI} = \max_{1\le i\le K} \min_{1\le j\le N} ||Z_i - F_j||, \tag{20}$$

where F_j are the non-dominated solutions found by the considered algorithm, and $\{Z_i, i = 1,\dots,K\}$ is the set of points well representing the Pareto set, i.e. Z_i are sufficiently densely and uniformly distributed over the Pareto set as described in [30]. The mean values and standard deviations of the considered metrics are presented in Table 1.

Almost all the results of the π-algorithm are somewhat better than those of the P-algorithm. The only exception is GD in the case of problem (19). However, the isolated comparison of two numbers here is insufficient, since NN of the π-algorithm in this case is double of that of the P-algorithm.

The experimental results corroborate the acceptability of the departure from the Gaussian model, and the good adaptivity of the generalized statistical model [21, 26]

Table 1 The performance criteria of the considered algorithms for problems (18), and (19)

Algorithm	π-algorithm				P-algorithm			
Problem	Problem (18)		Problem (19)		Problem (18)		Problem (19)	
NN	20.2	3.10	19.6	2.4	15.7	2.0	9.87	1.4
GD	0.044	0.027	0.055	0.028	0.070	0.051	0.015	0.0061
EI	0.13	0.074	0.14	0.047	0.13	0.053	0.20	0.034
Algorithm	GA				RUS			
Problem	Problem (18)		Problem (19)		Problem (18)		Problem (19)	
NN	9.3	2.5	8.2	2.2	12.1	3.0	6.6	1.6
GD	0.30	0.079	0.18	0.098	0.23	0.081	0.11	0.068
EI	0.38	0.078	0.55	0.21	0.26	0.056	0.33	0.097

to the properties of the objective functions; such a conclusion well consists with the properties of the generalized statistical models in case of single-objective global optimization.

In the second half of the Table 1 similar data of the experiments with the random uniform search (RUS) and the genetic algorithm (GA) is presented for the comparison. As mentioned above, functions (18) and (19) are somewhat similar to the worst case Lipshitz objective functions [29]. Therefore it is interesting to assess the performance of the worst case optimal algorithm for these problems. RUS is a randomized approximation of the optimal worst case algorithm which computes the values of the objective functions at the centers of balls optimally covering the feasible region. The results of experiments with RUS from [30], where the search was stopped after 100 computations of the objectives, are included in Table 1. The performance of RUS is obviously worse than that of the π-algorithm and of the P-algorithm. In [30] it is also reported how many computations are needed to RUS to achieve the average values of the metrics EI and GD comparable with those by the P-algorithm after 100 observations. The RUS, while solving problem (18), needs about 600 observations to reach average values of EI (0.13) and GD (0.096), which are close to those of the P-algorithm presented in Table 1. The averaged values of EI and GD, equal to 0.2331 and 0.0810 correspondingly, are reached by the RUS for (19) after 200 observations; in that case, the RUS needs only twice as many observations to match up the P-algorithm with respect to EI and GD. As it could be expected from the theory in [29], the relative performance of the uniform search was better in the case where the objective functions are more similar to the worst case ones.

The experiments with GA are reported to illustrate the hardness of the considered black-box multi-objective expensive global optimization problems, by demonstrating that the conventional algorithms are inappropriate here. The "Pareto Genetic Algorithm" from the book oriented to practical applications [8] was chosen for the experimentation. The intrinsic parameters of the GA have been selected as recommended by the authors in [8]. The problem relevant parameters, i.e., the population size, and the number of iterations, were chosen equal to 20

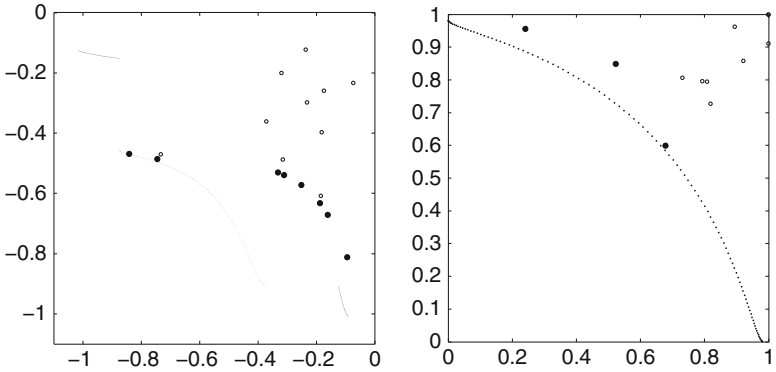


Fig. 2 The solutions (of problem (18) on the *left side*, and of problem (19) on the *right side*) generated by the GA algorithm at the last (fifth) iteration. Non-dominated solutions are denoted by *thicker points*. A *line* indicates the Pareto set

and 5 correspondingly, taking into account the termination condition of the other algorithms considered in our investigation. The values of metrics presented in Table 1 are obtained as averages of the 200 independent runs of the algorithm. The typical results presented in Fig. 2 illustrate large values of EI and GD.

Even with ten times larger number of computation of the objective functions values, where the population size was equal to 100 and the number of iterations was equal to 10, the average values of the considered metrics for GA were worse than those of the π-algorithm in Table 1. For the problem (18) the average values of EI and GD were 0.27 and 0.23 correspondingly; for the problem (19), those values were 0.29 and 0.10 correspondingly.

The presented results of the preliminary numerical experiments corroborate the theoretical conjecture that the proposed approach is competitive in the field of black-box expensive global optimization and deserves further investigation. The development of algorithms for the problems of larger dimensionality and with larger number of objectives is, although challenging, but highly promising.

5 Conclusions

The statistical model-based approach to single-objective global optimization is generalized relaxing some assumptions presented in the previous publications. The generalized approach is extended to the black-box multi-objective optimization of expensive objectives. The proposed approach facilitates construction of a family of algorithms, similar to the multi-objective P-algorithm. The results of experimentation with small size test problems are promising, and for

the development of algorithms, suitable for solving real-world problems, the generalization of the theory of statistical model-based global optimization to the multi-objective case is highly desirable.

Acknowledgment This research is supported by the Research Council of Lithuania under Grant No. MIP-063/2012.

References

1. Branke, J., Deb, K., Miettinen, K., Słowiński, R. (eds.): Multiobjective Optimization: Interactive and Evolutionary Approaches; Dagstuhl Seminar on Practical Approaches to Multi-Objective Optimization, Schlosss Dagstuhl, 10–15 December 2006. Lecture Notes in Computer Science, vol. 5252. Springer, New York (2008)
2. Chinchuluun, A., Pardalos, P.M.: A survey of recent developments in multiobjective optimization. Ann. Oper. Res. **154**(1), 29–50 (2007)
3. Deb, K.: Multi-Objective Optimization Using Evolutionary Algorithms. Wiley, New York (2009)
4. Elsakov, S.M., Shiryaev, V.I.: Homogeneous algorithms for multiextremal optimization. Comput. Math. Math. Phys. **50**(10), 1642–1654 (2010)
5. Emmerich, M., Giannakoglou, K., Naujoks, B.: Single- and multi-objective evolutionary optimization assisted by gaussian random field metamodels. IEEE Trans. Evol. Comput. **10**(4), 421–439 (2006)
6. Fishburn, P.: Utility Theory for Decision Making. Wiley, New York (1970)
7. Fonseca, C., Fleming, P.: An overview of evolutionary algorithms in multi-objective optimization. Evol. Comput. J. **3**(1), 1–16 (1995)
8. Haupt, R., Haupt, S.: Practical Genetic Algorithms. Wiley-Interscience, Hoboken, New Jersey (2004)
9. Horst, R., Pardalos, P., Thoai, N.: Introduction to Global Optimization. Kluwer Academic, Dordecht (2007)
10. Keane, A., Scalan, J.: Design search and optimization in aerospace engineering. Philos. Trans. R. Soc. A **365**, 2501–2529 (2007)
11. Knowles, J.: Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans. Evol. Comput. **10**(1), 50–66 (2006)
12. Knowles, J., Corne, D., Reynolds, A.: Noisy Multiobjective Optimization on a Budget of 250 Evaluations. Lecture Notes in Computer Science, vol. 5467, pp. 36–50, Springer-Verlag, Berlin, Heidelberg (2009)
13. Kushner, H.: A versatile stochastic model of a function of unknown and time-varying form. J. Math. Anal. Appl. **5**, 150–167 (1962)
14. Miettinen, K.: Nonlinear Multiobjective Optimization. Springer, New York (1999)
15. Mockus, J.: Bayesian Approach to Global Optimization. Kluwer Academic, Dordecht (1988)
16. Nakayama, H., Yun, Y., Yoon, M.: Sequential Approximate Multiobjective Optimization Using Computational Intelligence. Springer, New York (2009)
17. Sayin, S.: Measuring the quality of discrete representation of efficient sets in multiple objective mathematical programming. Math. Program. **87 A**, 543–560 (2000)
18. Sergeyev, Ya.D.: Numerical computations and mathematical modelling with infinite and infinitesimal numbers. J. Appl. Math. Comput. **29**, 177–195 (2009)
19. Sergeyev, Ya.D.: Lagrange lecture: Methodology of numerical computations with infinities and infinitesimals. Rendiconti del Seminario Matematico dell'Università e del Politecnico di Torino **68**(2), 95–113 (2010)

20. Strongin, R.G., Sergeyev, Y.D.: Global Optimization with Non-convex Constraints: Sequential and Parallel Algorithms. Kluwer Academic, Dordecht (2000)
21. Törn, A., Žilinskas, A.: Global Optimization. Lecture Notes in Computer Science, vol. 350, pp. 1–255, Springer-Verlag, Berlin, Heidelberg (1989)
22. Wagner, T., Emmerich, M., Deutz, A., Ponweiser, W.: On Expected-Improvement Criteria for Model-Based Multi-objective Optimization. Lecture Notes in Computer Science, vol. 6238, pp. 718–727, Springer-Verlag, Berlin Heidelberg (2010)
23. Zhigljavsy, A., Žilinskas, A.: Stochastic Global Optimization. Springer, Berlin (2008)
24. Žilinskas, A.: One-step bayesian method for the search of the optimum of one-variable functions. Cybernetics and Systems Analysis, **11**(1), 160–166 (1975)
25. Žilinskas, A.: Otimization of one-dimensional multimodal functions, algorithm 133. J. R. Stat. Soc. Ser. C **23**, 367–385 (1978)
26. Žilinskas, A.: Axiomatic approach to statistical models and their use in multimodal optimizatio theory. Math. Program. **22**, 104–116 (1982)
27. Žilinskas, A.: Axiomatic characterization of a global optimization algorithm and investigation of its search strategies. Oper. Res. Lett. **4**, 35–39 (1985)
28. Žilinskas, A.: On strong homogeneity of two global optimization algorithms based on statistical models of multimodal objective functions. Appl. Math. Comput. **218**(16), 8131–8136 (2012)
29. Žilinskas, A.: On the worst-case optimal multi-objective global optimization. Optim. Lett. **7**(8), 1921–1928 (2013)
30. Žilinskas, A.: A statistical model-based algorithm for black-box multi-objective optimization. Int. J. Syst. Sci. **45**(1), 82–93 (2014)
31. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., Fonseca, V.G.D.: Performance assessment of multiobjective optimizers: An analysis and review. IEEE Trans. Evol. Comput. **7**, 117–132 (2003)
32. Zopounidis, C., Pardalos, P. (eds.): Handbook of Multicriteria Analysis. Springer, Berlin (2010)