SPRINGER OPTIMIZATION
AND ITS APPLICATIONS 39

Altannar Chinchuluun · Panos M. Pardalos
Rentsen Enkhbat · Ider Tseveendorj (Eds.)

Optimization and Optimal Control

Theory and Applications

Springer

OPTIMIZATION AND OPTIMAL CONTROL

Springer Optimization and Its Applications

VOLUME 39

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

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

OPTIMIZATION AND OPTIMAL CONTROL

Theory and Applications

Edited By

ALTANNAR CHINCHULUUN
Centre for Process and Systems Engineering,
Imperial College London,
South Kensington Campus,
London SW7 2AZ, UK

PANOS M. PARDALOS
Department of Industrial and Systems Engineering,
University of Florida,
Gainesville, FL 32611, USA

RENTSEN ENKHBAT
School of Economic Studies,
National University of Mongolia,
Ulaanbaatar, Mongolia

IDER TSEVEENDORJ
Computer Science Department,
Université de Versailles-Saint Quentin en Yvelines,
Versailles, France

Springer

Editors
Altannar Chinchuluun
Centre for Process Systems Engineering
Imperial College London
South Kensington Campus
London SW7 2AZ, UK
altannar@imperial.ac.uk

Panos M. Pardalos
Department of Industrial and Systems Engineering
University of Florida
Weil Hall 303
32611-6595 Gainesville Florida
USA
pardalos@ufl.edu

Rentsen Enkhbat
School of Economic Studies
National University of Mongolia
Baga Toiruu 4
Sukhbaatar District
Mongolia
renkhbat46@ses.edu.mn

Ider Tseveendorj
Université Versailles
Labo. PRISM
av. des Etats-Unis 45
78035 Versailles CX
France
ider.tseveendorj@prism.uvsq.fr

ISSN 1931-6828
ISBN 978-0-387-89495-9 e-ISBN 978-0-387-89496-6
DOI 10.1007/978-0-387-89496-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010927368

Conquering the world on horseback is easy; it is dismounting and governing that is hard.
– Chinggis Khan

Translation adapted from *The Gigantic Book of Horse Wisdom* (2007) by Thomas Meagher and Buck Brannaman.

Preface

Optimization and optimal control are the main tools in decision making. In optimization we often deal with problems in finite-dimensional spaces. On the other hand, in optimal control we solve problems in infinite-dimensional spaces. Many problems in engineering, physics, economics and other fields can be formulated as optimization and optimal control problems.

This book brings together recent developments in optimization and optimal control as well as recent applications of these results to a wide range of real-world problems. The book consists of 24 chapters contributed by experts around the world who work with optimization and optimal control either at a theoretical level or at the level of using these tools in practice. Each chapter is not only of expository but also of scholarly nature.

The first 12 chapters focus on optimization theory and equilibrium problems. The chapter by A. Antipin studies optimization problems generated by sensitivity functions for convex programming problems. Methods for these problems are proposed and properties of the sensitivity functions are analyzed. The chapter by M.A. Goberna gives an overview of the state of the art in sensitivity and satiability analysis in linear semi-infinite programming. In the chapter by G. Kassay, scalar equilibrium problems are considered. Applications of these problems in nonlinear analysis are discussed and some new results concerning the existence of exact and approximate solutions are presented. The chapter by G. Isac presents the concept of scalarly compactness in nonlinear analysis. Applications of the concept to the study of variational inequalities and complementarities problems are discussed. The chapter by N.X. Tan and L.J. Lin formulates Blum–Oettli type quasi-equilibrium problems and establishes sufficient conditions for the existence of their solutions. The chapter by R. Enkhbat and Ya. Bazarsad formulated the response surface problems as quadratic programming problems. Solution approaches for these quadratic programming problems based on global optimality conditions are proposed. The chapter by D.Y. Gao et al. proposes a canonical dual approach for solving a fixed cost mixed-integer quadratic programming problem. It is shown that, using so-called canonical duality theory, the problem can be

reduced to canonical convex dual problem with zero gap which can be tackled by many efficient local search methods. The chapter by B. Luderer and B. Wagner considers the problem of finding the intersection of the convex hulls of two sets containing finitely many points each. An algorithm for the problem is proposed based on the equivalent quasi-differentiable optimization problem. The chapter by M.-A. Majig et al. proposes an evolutionary search algorithm for solving the global optimization problem with box constraint. The algorithm finds as many solutions of the problem as possible or all solutions in some cases. The evolutionary search also employs a local search procedure. The chapter by L. Altangerel and G. Wanka deals with the perturbation approach in the conjugate duality for vector optimization on the basis of weak ordering. New gap functions for vector optimization are proposed and their properties are studied. The chapter by D. Li et al. gives an overview of six polynomially solvable classes of binary quadratic programming problems and provides examples and geometric illustrations to give intuitive insights of the problems. The chapter by B. Jadamba et al. deals with an ill-posed multi-valued quasi-variational inequality problem. A parameter identification problem that gives a stable approximation procedure for the ill-posed problem is formulated and generalizations of this approach to other problems are discussed.

The next five chapters are concerned with optimal control theory and algorithms. The chapter by Z.G. Feng and K.L. Teo considers a class of optimal feedback control problems where its dynamical system is described by stochastic linear systems subject to Poisson processes and with state jumps. They show that the problem is equivalent to a deterministic impulsive optimal parameter selection problem with fixed jump times and provide an efficient computational method for the later problem. In the chapter by V. Maksimov, controlled differential inclusions involving subdifferentials of convex functions are considered. In particular, the three problems, the problem of prescribed motion realization, the problem of robust control, and the problem of input dynamical reconstruction, are suited. Stable feedback control-based algorithms for solving the problems are presented. The chapter by B.D.O. Anderson et al. proposes a new algorithm for solving Riccati equations and certain Hamilton-Jacobi-Bellman-Isaacs equations arising in H_∞ control. In the chapter by D. Vrabie and F. Lewis, a new online direct adaptive scheme is constructed in order to find an approximate solution to the state feedback, infinite-horizon, optimal control problem. In the chapter by A.S. Buldaev, iterative perturbation methods for nonlinear optimal control problems which are polynomial with respect to the state are proposed.

The remaining seven chapters are largely devoted to applications of optimization and optimal control. The chapter by H.P. Geering et al. explains how stochastic optimal control theory can be applied to optimal asset allocation problems under consideration of risk aversion. Two types of problems are studied and corresponding solution techniques are presented. The chapter by F.D. Fagundez et al. considers scheduling problems in the process industry.

A nonlinear dynamic programming model for the process scheduling is proposed and the results are compared with those of different mixed integer nonlinear programming models. The chapter by D. Fortin is concerned with quantum computing and Grothendieck's constant. A non-cooperative quantum game is presented and it is also shown that for many instances of rank-deficient correlation matrices Grothendieck's constants go beyond $\sqrt{2}$ for sufficiently large size. The chapter by H. Damba et al. considers a problem of identifying a pasture region where the grass mass in the region is maximized. The chapter by W.-J. Hwang et al. considers the rate control problem in wired-cum-wireless networks. It is shown that there is a unique solution for end-to-end session rates and infinitely many corresponding optimal values for wireless link transmission rates of the optimization problems, where the optimization variables are both end-to-end session rates and wireless link transmission rates. The chapter by N. Fan et al. explores the relationship between biclustering and graph partitioning. Several integer programming formulations for the different cuts including ratio cut and normalized cut are presented. In the chapter by M. Tamaki and Q. Wang, a best choice problem in queue theory is considered. The problem is to find a procedure to select the best applicant by selecting or rejecting the applicants. They give the explicit rule for the best choice problem where the number of applicants is uniformly distributed.

We would like to take this opportunity to thank the authors of the chapters, the anonymous referees, and Springer for making the publication of this book possible.

London, UK — A. Chinchuluun
Gainesville, FL, USA — P.M. Pardalos
Ulaanbaatar, Mongolia — R. Enkhbat
Versailles, France — I. Tseveendorj

Contents

Sensibility Function as Convolution of System of Optimization Problems

Anatoly Antipin

Computing Center of Russian Academy of Sciences, Vavilov str., 40, 119333 Moscow, Russia
antipin@ccas.ru

Summary. The sensibility function generated by a convex programming problem is viewed as an element of a complex system of optimization problems. Its role in this system is clarified. The optimization problems generated by the sensibility function are considered. Methods for their solution are proposed.

Key words: sensibility function, system of optimization problems, extraproximal method

1 Introduction

The sensibility function has been intensively studied since the first publications on this subject [1, 2]. A complete bibliography can be found in [3], where the directional differentiability of the sensibility function was defined and its properties were examined. Issues concerning perturbation theory and the associated properties of the sensibility function in convex programming problems were discussed in [4]. The convexity of the sensibility function generated by a convex programming problem was proved in [5, 6]. In [7] the relationship between the sensibility function and the set of Pareto solutions of a multicriteria optimization problem was established in the case when the problem's vector criterion is formed of the objective function and of the functional constraints in the nonlinear programming problem. For convex programming problems, the sensibility function can be treated as a parametrization of the subset of Pareto solutions that are in the positive orthant, since the graph of the sensibility function coincides with this subset. Methods for computing multicriteria solutions for a nonconvex Pareto manifold were proposed in [8]. In [9] the sensibility function was treated as a usual element of the space of differentiable functions.

In this chapter the sensibility function is viewed as an element of a system of optimization problems, i.e., in fact, of game problems with a Nash equilibrium. In the framework of this system, the sensibility function itself forms

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_1,

an optimization problem whose solution solves the original system. Moreover, the optimization problem generated by the sensibility function can be treated as a convolution or scalarization of the original system. Accordingly, methods for solving systems of optimization problems are those for optimizing the sensibility function on different sets for different systems.

Let us first review the properties of the sensibility function. In contrast to the traditional approach, we give new definitions of the convexity and subdifferentiability of the function that are based on a saddle point of the Lagrangian for a convex programming problem.

The sensibility function is generated by the following parametric convex programming problem with the right-hand side column vector of functional constraints $y \in R^m_+$ used as a parameter:

$$\varphi(y) = \min\{f(w) \mid g(w) \leq y,\ w \in W_0\},\ y \in R^m_+. \tag{1}$$

Here, the objective function $f(w)$ and each component of the vector function $g(w)$ are convex scalar functions, $W_0 \subset R^n$ is a convex closed set, and $y \in R^m_+$. In the general case, $\varphi(y)$ is defined on the entire space R^m (if the feasible set of the problem is empty for some y, then by definition $\varphi(y) = +\infty$), but in this chapter we restrict ourselves to the case of R^m_+.

Recall some properties of the sensibility function.

Property 1. The sensibility function is monotonically decreasing.

Indeed, if $y_1 \leq y_2$ (in the sense of a partial order), then the feasible set corresponding to y_2 includes that corresponding to y_1. On a larger set, the objective function value can be only smaller than on the original feasible set corresponding to y_1. Therefore, $\varphi(y_1) \geq \varphi(y_2)$.

Recall that, by definition, we have (1). Assume also that this problem is regular (e.g., the Slater condition holds) for any $y \in R^m_+$. This in turn means that the system of inequalities

$$f(w_y) + \langle p, g(w_y) - y\rangle \leq f(w_y) + \langle p_y, g(w_y) - y\rangle \leq f(w) + \langle p_y, g(w) - y\rangle \tag{2}$$

holds for all $w \in W_0, p \geq 0$. Here, $w_y \in W_0, p_y \geq 0$ is a saddle point of $L(p, w, y) = f(w) + \langle p, g(w) - y\rangle$ for a fixed parameter value $y \geq 0$.

Given arbitrary convex $f(w)$, $g(w)$, and arbitrary y, the function $L(p, w, y)$ usually has several saddle points. However, since $L(p, w, y)$ is a continuous and convex function of its variables [10], its saddle points form a convex closed set. If, additionally, problem (2) is regular, then the set of saddle points is bounded with respect to $p \geq 0$. Indeed, setting $w = w_0$ in the right inequality in (2), where w_0 is a Slater point, i.e., a point satisfying $g_i(w_0) < 0, i = 1, 2, \ldots, m$, we obtain

$$0 \leq \langle p_y, g(w_0) - y\rangle \leq f(w_0) - f(w_y).$$

Now, assuming that a certain component of p_y is infinitely large, we obtain a contradiction to the estimate.

It is useful to rewrite system (2) in the equivalent form

$$w_y \in \operatorname{Argmin}\{f(w) + \langle p, g(w) - y\rangle \mid w \in W_0\}, \tag{3}$$

$$\langle p - p_y, g(w_y) - y\rangle \le 0,\ p \ge 0. \tag{4}$$

Since the variational inequality of this system is defined on the positive orthant, it splits into two relations that form a complementarity problem. To show this, it suffices to set $p = 0$ and, then, $p = 2p_y$ in this inequality. Then we obtain

$$\langle p_y, g(w_y) - y\rangle = 0, \quad g(w_y) - y \le 0. \tag{5}$$

In view of (5), we can see that system (3), (4) is equivalent to

$$w_y \in \operatorname{Argmin}\{f(w) \mid g(w) \le y,\ w \in W_0\}, \tag{6}$$

$$\langle p - p_y, g(w_y) - y\rangle \le 0,\ p \ge 0, \tag{7}$$

where (6) coincides with (1). Let us show that $\varphi(y)$ is convex and subdifferentiable.

Property 2. The sensibility function $\varphi(y)$ of a regular convex programming problem is convex and subdifferentiable.

Definition 1. *The function $\varphi(y)$ is said to be convex and subdifferentiable if for any y from its domain there exists a subdifferential $\nabla\varphi(y)$ (a convex closed bounded set) and $\varphi(y)$ satisfies the system of inequalities*

$$\langle \nabla\varphi(y_0), y - y_0\rangle \le \varphi(y) - \varphi(y_0) \le \langle \nabla\varphi(y), y - y_0\rangle \tag{8}$$

for all $y \ge 0$ and $y_0 \ge 0$.

For illustrative purposes, we rewrite system (2) for a fixed parameter value $y = y_0$:

$$f(w_{y_0}) + \langle p, g(w_{y_0}) - y_0\rangle \le f(w_{y_0}) + \langle p_{y_0}, g(w_{y_0}) - y_0\rangle \le f(w) + \langle p_{y_0}, g(w) - y_0\rangle \tag{9}$$

for all $w \in W_0, p \ge 0$. According to (5), the left variational inequality of this system

$$\langle p - p_{y_0}, g(w_{y_0}) - y_0\rangle \le 0, \quad p \ge 0 \tag{10}$$

is equivalent to the complementarity problem

$$\langle p_{y_0}, g(w_{y_0}) - y_0\rangle = 0, \quad g(w_{y_0}) - y_0 \le 0. \tag{11}$$

Specifically, when $p = p_y$, relation (10), combined with (11), yields

$$\langle p_y, g(w_{y_0}) - y_0\rangle \le \langle p_{y_0}, g(w_{y_0}) - y_0\rangle = 0. \tag{12}$$

Similarly, when $p = p_{y_0}$, from (4) in view of (5), we have

$$\langle p_{y_0}, g(w_y) - y\rangle \le \langle p_y, g(w_y) - y\rangle = 0. \tag{13}$$

When $w = w_y$, the right inequality in (9) in view of (11) yields

$$\langle p_{y_0}, y_0 - g(w_y)\rangle \le f(w_y) - f(w_{y_0}).$$

Using condition (13), we rearrange this inequality into

$$\langle p_{y_0}, y_0 - y\rangle \le f(w_y) - f(w_{y_0}). \tag{14}$$

Accordingly, the right inequality in (2) with $w = w_{y_0}$ yields

$$\langle p_y, y - g(w_{y_0})\rangle \le f(w_{y_0}) - f(w_y). \tag{15}$$

In view of (12), we obtain

$$\langle p_y, y - y_0\rangle \le f(w_{y_0}) - f(w_y). \tag{16}$$

From (1), (2), and (9), it is easy to see that $f(w_y) = \varphi(y)$, $f(w_{y_0}) = \varphi(y_0)$. In view of these relations, (14) and (16) can be rewritten as

$$\langle -p_{y_0}, y - y_0\rangle \le \varphi(y) - \varphi(y_0) \le \langle -p_y, y - y_0\rangle. \tag{17}$$

Here, p_y and p_{y_0} are any Lagrange multiplier vectors of problem (1) that satisfy system (2) or (9). As was mentioned above, the collection of such vectors corresponding to any parameter value $y \ge 0$ is a convex closed bounded set.

Introducing the notation $\nabla\varphi(y) = -p_y$ and $\nabla\varphi(y_0) = -p_{y_0}$, we call any of these vectors a subgradient of $\varphi(y)$ at $y \in R^m$. The set of all subgradients at y is called a subdifferential; moreover,

$$\nabla\varphi(y) \in \frac{\partial\varphi(y)}{\partial y},\ \nabla\varphi(y_0) \in \frac{\partial\varphi(y)}{\partial y}|_{y=y^0}.$$

By using the notation introduced, (17) can be rewritten in the form of (8).

Property 3. The sensibility function $\varphi(y)$ is convex in the sense of Jensen's inequality [10].

Let $y(\alpha) = \alpha y + (1-\alpha)y_0$. Then (8) implies

$$\langle \nabla\varphi(y(\alpha)), y - y(\alpha)\rangle \le \varphi(y) - \varphi(y(\alpha)),$$
$$\langle \nabla\varphi(y(\alpha), y_0 - y(\alpha)\rangle \le \varphi(y_0) - \varphi(y(\alpha)).$$

Multiplying the first inequality by α and the second by $(1-\alpha)$ and summing them up, we obtain

$$0 = \langle \nabla f(y(\alpha)), y(\alpha) - y(\alpha) \le \alpha f(y) + (1-\alpha)f(y_0) - f(y(\alpha)).$$

This yields

$$f(\alpha y + (1-\alpha)y_0) \le \alpha f(y) + (1-\alpha)f(y_0),\ y \ge 0, \quad y_0 \ge 0. \tag{18}$$

Property 4. The subdifferential of the sensibility function is a monotone set-valued mapping.

System (8) is represented in the form

$$\langle \nabla\varphi(y_0), y - y_0\rangle \le \varphi(y) - \varphi(y_0),\ \ \varphi(y) - \varphi(y_0) \le \langle \nabla\varphi(y), y - y_0\rangle.$$

Summing up both inequalities gives

$$\langle \nabla\varphi(y) - \nabla\varphi(y_0), y - y_0\rangle \ge 0 \tag{19}$$

for all $y \ge 0$ and $y_0 \ge 0$.

Property 5. The epigraph of the sensibility function is a convex closed set.

Let $\mathrm{epi}\varphi = \{(y, \mu) \mid y \in \mathrm{dom}\varphi, \mu \ge \varphi(y)\}$ be the epigraph of $\varphi(y), y \ge 0$. If the points μ_0, y_0 and μ_1, y_1 belong to $\mathrm{epi}\varphi$, then $\mu_0 \ge \varphi(y_0), y_0 \in \mathrm{dom}\varphi$ and $\mu_1 \ge \varphi(y_1), y_1 \in \mathrm{dom}\varphi$ at these points. Multiplying the first condition by α and the second by $(1-\alpha)$ and summing them up, we obtain $\alpha\mu_0 + (1-\alpha)\mu_1 \ge \alpha\varphi(w_0) + (1-\alpha)\varphi(w_1) \ge \varphi(\alpha w_0 + (1-\alpha)w_1), \alpha y_0 + (1-\alpha)y_1$. Thus, if the points μ_0, y_0 and μ_1, y_1 belong to the epigraph of $\varphi(y)$, then the entire segment joining them belongs to the epigraph as well. This means that the epigraph of $\varphi(y)$ is a convex set.

Property 6. The graph of the sensibility function coincides with the subset of positive-orthant Pareto solutions to the multicriteria optimization problem generated by the objective function and the functional constraints.

Define the vector function $F(w) = (f(w), g(w))$ and consider the vector optimization problem

$$F(w^*) = \min\{F(w) \mid w \in W_0\}. \tag{20}$$

The solution set of this problem is a large set of Pareto optimal, or Pareto effective, points. All of them are determined by the following formal condition: $F(w^*)$ is called a Pareto optimal point if there is no vector v such that

$$F(v) \le F(w^*) \text{ and } F(v) \ne F(w^*),$$

i.e., the negative (closed) orthant $K(F(w^*))$ with its vertex at $F(w^*)$ contains no points of the set $F = \{F(w), w \in W_0\}$ other than $F(w^*)$. Stated differently, any point of $F(w^*)$ is such that the intersection of the set F (which is the image of W_0 under the mapping $F(w)$) and $K(F(w^*))$ with its vertex at $F(w^*)$ contains the single point $F(w^*)$.

Recall that the Kuhn–Tucker theorem in the regular case implies that every $y \ge 0$ in problem (1) is associated with a vector of Lagrange multipliers $p_y \ge 0$. According to property 2, every Lagrange multiplier vector is a subgradient $\nabla\varphi(y) = p_y$ of sensibility function (1) (see (8)). Moreover, every $y \ge 0$ is associated with a vector $(f(w_y), g(w_y))$ such that

$$f(w_y) + \langle p_y, g(w_y)\rangle \le f(w) + \langle p_y, g(w)\rangle,\ w \in W_0,$$
$$\langle p, g(w_y) - y\rangle \le \langle p_y, g(w_y) - y\rangle,\ p \ge 0. \tag{21}$$

The first inequality in this system implies that $(f(w_y),\ g(w_y))$ is a Pareto optimal point, while $(1,\ p_y)$ is the normal vector to its linear support functional. Note that the domain of the mapping $\nabla\varphi(y) = p_y$ depends substantially on $f(w)$ and $g(w)$: This domain can include the entire positive orthant $Y = R^m_+$, its proper subset of lower dimension, or the origin $Y = 0$. The last case is possible if the minimizer in the convex programming problem satisfies the Slater condition. Then the domain of the sensibility function for this problem shrinks to a point (to the origin) and the image of $\nabla\varphi(y) = p_y$ is also the origin. If the minimizer of the problem coincides with the intersection point of m functional constraints, i.e., the minimizer solves a system of m equations, then the domain of the sensibility function is the entire orthant $Y = R^m_+$, and, if the minimizer is an interior point for some constraints, then the domain is an orthant of lower dimension.

Accordingly, the range of $\nabla\varphi(y) = p_y$ has a similar structure: It can be the entire orthant, its proper subset, or the origin. Indeed, given a vector $p \ge 0$ with nonzero components such that all the components of $g(w_y)$ in the first inequality in (21) are strictly negative. Then the linear functional in the second inequality in (21) has a normal vector all of whose components are negative (for any $y \ge 0$, which can always be assumed to be zero). However, a linear functional with strictly negative normal components can reach a maximum on the positive orthant only at the origin. Thus, assuming that all the components of p are initially nonzero, we obtain a contradiction. This means that some points of the positive orthant are not the images of $\nabla\varphi(y) = p_y$.

Let us return to the second inequality in (21). We see that the linear functional is bounded above by a constant. This is possible if its normal is zero (i.e., $g(w_y) - y = 0$, which gives $g(w_y) = y$) or if some components of the normal are strictly negative, in which case the corresponding components of p (Lagrange multipliers) are zero and the first inequality in (21) holds as well. Thus, we have

$$g(w_y) - y = 0.$$

Here, if some of the components of $g(w_y)$ are negative, then the corresponding components of y are also zero and this equality holds on a subspace of a lower dimension than m. Note that this subspace contains the graph of the sensibility function, which coincides with the set of Pareto optimal solutions to problem (20). Thus, taking into account $\varphi(y) = f(w_y)$ and $g(w_y) = y$, we conclude that the point $(\varphi(y), y)$ on the graph of the sensibility function corresponds to the Pareto optimal point $(f(w_y), g(w_y))$, which is in the positive orthant.

The converse is also true. Pareto optimal points in the positive orthant lie on the graph of the sensibility function. It was shown above that the image of $\nabla\varphi(y) = p_y$ is not the entire positive orthant but rather a subset of it. Denote

this image by $P_0 = R^n_+$ and consider the inverse mapping $\nabla\psi(p) : P_0 \to R^n$. Given a fixed weight vector $p \in P_0$, it has at least one nonzero component and satisfies the inequality

$$f(w_p) + \langle p, g(w_p)\rangle \le f(w) + \langle p, g(w)\rangle,\ w \in W_0. \tag{22}$$

Here, $f(w_p), g(w_p)$ is a Pareto optimal point as a minimizer of a linear function on the image of the vector criterion $(f(w), g(w))$. To each vector $p \in P_0$, we assign the vector $\nabla\psi(p) = y$ according to the following rule: $y_i = g_i(w_p)$ if $g_i(w_p) \ge 0$ and $y_i = 0$ if $g_i(w_p) < 0$, where $i = 1, 2, \ldots, m$. This rule can be written as the relations

$$\langle p, g(w_p) - y\rangle = 0,\ g(w_p) - y \le 0, \tag{23}$$

which are equivalent to the variational inequality

$$\langle p' - p^*, g(w_p) - y\rangle \le 0,\ p' \ge 0. \tag{24}$$

Combining (22) and (24), we formulate the convex programming problem

$$f(w_p) + \langle p, g(w_p)\rangle \le f(w) + \langle p, g(w)\rangle, \quad w \in W_0, \tag{25}$$

$$\langle p' - p, g(w_p) - y\rangle \le 0, \quad p' \ge 0. \tag{26}$$

These inequalities are equivalent to the problem

$$w_p \in \mathrm{Argmin}\{f(w) \mid g(w) \le y, w \in W_0\}. \tag{27}$$

Here, some of the components of y are zero if they correspond to zero Lagrange multipliers. Thus, each Pareto optimal point in the vector optimization problem (20) is associated with a point lying on the graph of the sensibility function (1).

2 Optimization Problems for the Sensibility Function

Problem (6), (7) or its equivalent (3), (4) is a system of two optimization problems with no additional constraints imposed on the variable $y \ge 0$. However, in mathematical (more exactly, economic) simulation, such constraints are needed to describe the interaction between two agents, of which one offers a vector of resources, while the other sets the price to purchase them. Modification (6), (7) leads to a problem that can be viewed as a model of this situation:

$$w^* \in \mathrm{Argmin}\{f(w) \mid g(w) \le y^*,\ w \in W_0\}, \tag{28}$$

$$\langle p - p^*, g(w^*) - y^*\rangle \le 0, \quad p \ge 0, \tag{29}$$

$$y^* \in \mathrm{Argmin}\{\langle p^*, y\rangle \mid y \in Y\}. \tag{30}$$

Here, the goal is to choose a right-hand side vector of functional constraints $y = y^*$ and the corresponding Lagrange multiplier vector $p = p^*$ such that the linear function $\langle p^*, y\rangle,\ y \in Y$ reaches its minimal value on Y at the point y^*. The first two components of the vector p^*, w^*, y^* are called the primal and dual solutions to problem (28), (29) and comprise a saddle point of the Lagrangian

$$L(p, w, y^*) = f(w) + \langle p, g(w) - y^*\rangle, \quad p \geq 0, \quad w \in W_0. \tag{31}$$

This point satisfies the system of inequalities

$$f(w^*) + \langle p, g(w^*) - y^*\rangle \leq f(w^*) + \langle p^*, g(w^*) - y^*\rangle \leq f(w) + \langle p^*, g(w) - y^*\rangle, \tag{32}$$

where $p \geq 0, w \in W_0$, and $y = y^*$ is a fixed parameter.

However, in this work we consider a problem more complicated than (28), (29), and (30), namely, [11, 12]

$$w^* \in \operatorname{Argmin}\{f_1(w) \mid g(w) \leq h(y^*),\ w \in W_0\}, \tag{33}$$

$$\langle p - p^*, g(w^*) - h(y^*)\rangle \leq 0, \quad p \geq 0, \tag{34}$$

$$y^* \in \operatorname{Argmin}\{f_2(y) - \langle p^*, h(y)\rangle \mid y \in Y\}. \tag{35}$$

Here, $f_1(w)$ and $f_2(y)$ are scalar convex functions; $g(w)$ and $h(y)$ are vector functions all of whose components are convex and concave functions, respectively; $p \in R^m_+$ is the positive orthant; and $W_0 \subset R^n$ and $Y \subset R^m_+$ are convex closed sets (specifically, Y can be a bounded polyhedral set).

In (33), (34), the goal is to choose a right-hand side vector of functional constraints $y = y^*$ such that the dual solution to this problem, i.e., the vector $p = p^*$, generates optimization problem (35) whose objective function reaches a minimum on Y at the point $y^* \in Y$ and, additionally, $h(y^*)$ coincides with the right-hand side vector of functional constraints in problem (33). As is customary, the vectors $p^* \geq 0$ and $w^* \in W_0$ are called dual and primal solutions to the convex programming problem (33), (34). This means that this pair is a saddle point of this problem's Lagrangian

$$L(p, w, y^*) = f_1(w) + \langle p, g(w) - h(y^*)\rangle, \quad p \geq 0, \quad w \in W_0, \tag{36}$$

where the variable $y \in Y$, which takes the value $y = y^* \in Y$ in (36), is a parameter in problem (33), (34). The term "saddle point" always means that

$$f_1(w^*) + \langle p, g(w^*) - h(y^*)\rangle \leq f_1(w^*) + \langle p^*, g(w^*) - h(y^*)\rangle \leq f_1(w) + \\ + \langle p^*, g(w) - h(y^*)\rangle, \tag{37}$$

where $p \geq 0, w \in W_0$, and $y = y^*$ is a fixed parameter.

Using (37), we rewrite (33), (34), and (35) in a different form, namely, as a system consisting of two optimization problems and a variational inequality:

$$w^* \in \operatorname{Argmin}\{f_1(w) + \langle p^*, g(w)\rangle \ \mid \ w \in W_0\},$$

$$\langle p - p^*, g(w^*) - h(y^*)\rangle \le 0, \quad p \ge 0,$$
$$y^* \in \mathrm{Argmin}\{f_2(y) - \langle p^*, h(y)\rangle \mid y \in Y\}. \tag{38}$$

Since the variational inequality is defined on the positive orthant, it splits into two relations that make up a complementarity problem. To see this, it suffices to set $p = 0$ and, then, $p = 2p^*$ in the inequality. Then

$$\langle p^*, g(w^*) - h(y^*)\rangle = 0, \quad g(w^*) - h(y^*) \le 0. \tag{39}$$

Using conditions (39), we can rewrite (38) in the form of (33), (34), and (35). The first two conditions in (38) correspond to (33) and (34). Thus, the equivalence of (38) to (33), (34), and (35) is obvious.

The variational inequality in (38) can also be written as a linear optimization problem. Then this system can be represented as a three-person game with a Nash equilibrium:

$$w^* \in \mathrm{Argmin}\{f_1(w) + \langle p^*, g(w)\rangle \ \mid \ w \in W_0\},$$
$$w^* \in \mathrm{Argmax}\{\langle p, g(w^*) - h(y^*)\rangle \mid p \ge 0\},$$
$$y^* \in \mathrm{Argmin}\{f_2(y) - \langle p^*, h(y)\rangle \mid y \in Y\}. \tag{40}$$

It is easy to see that the first and third problems in this system can be represented as a single optimization problem with a separable objective function with respect to $w \in W_0, y \in Y$. Then system (40) becomes

$$w^*, y^* \in \mathrm{Argmin}\{f_1(w) + f_2(y) + \langle p^*, g(w) - h(y)\rangle \ \mid \ w \in W_0, \ y \in Y\},$$
$$w^* \in \mathrm{Argmax}\{\langle p, g(w^*) - h(y^*)\rangle \mid p \ge 0\}. \tag{41}$$

In turn, system (41) is a zero-sum two-person game, which is equivalent to finding a saddle point of the function

$$\mathcal{L}(w, y, p) = f_1(w) + f_2(y) + \langle p, g(w) - h(y)\rangle, \quad w \in W_0, \quad y \in Y, \quad p \ge 0,$$

where the saddle point satisfies the system of inequalities

$$f_1(w^*) + f_2(y^*) + \langle p, g(w^*) - h(y^*)\rangle \le f_1(w^*) + f_2(y^*) +$$
$$+\langle p^*, g(w^*) - h(y^*)\rangle \le f_1(w) + f_2(y) + \langle p^*, g(w) - h(y)\rangle \tag{42}$$

for all $w \in W_0, y \in Y, p \ge 0$. Thus, we have shown that the original problem (33), (34), and (35) is reduced to saddle-point problem (42) or (41).

Conversely, if (42) holds, then the left inequality in this system yields

$$\langle p - p^*, g(w^*) - h(y^*)\rangle \le 0, \ p \ge 0,$$

which implies (39). From the right inequality in (42), we have

$$f_1(w^*) + f_2(y^*) \le f_1(w) + f_2(y) + \langle p^*, g(w) - h(y)\rangle.$$

If $w \in W_0$ and $y \in Y$ satisfy the constraint $\langle p^*, g(w) - h(y)\rangle \le 0$, then the above inequality is reduced to the optimization of $f_1(w) + f_2(y)$ on the set $W_0 \times Y$ with a single scalar constraint, i.e.,

$$f_1(w^*) + f_2(y^*) \le f_1(w) + f_2(y),\ \langle p^*, g(w) - h(y)\rangle \le 0,\ w \in W_0, y \in Y.$$

Taking into account (39), we reduce this problem to

$$f_1(w^*) + f_2(y^*) \le f_1(w) + f_2(y),\ g(w) - h(y) \le 0,\ w \in W_0, \quad y \in Y.$$

Specifically, if $y = y^*$, we obtain (33), (34), and

$$f_1(w^*) \le f_1(w),\ g(w) \le h(y^*),\ w \in W_0.$$

Now, setting $w = w^*$ in (42), we obtain (35) and

$$f_2(y^*) - \langle p^*, h(y^*)\rangle \le f_2(y) - \langle p^*, h(y)\rangle, \quad y \in Y.$$

Thus, we have proved the following result.

Theorem 1. *Let $f_1(w), f_2(y), g(w)$ be convex functions; $h(y)$ be concave; and W_0, Y be closed and convex sets. Then the systems of problems (33), (34), (35), (38), (41), and (42) are equivalent.*

Note that problem (28), (29), and (30) is a special case of (38). That is why the role of the sensibility function in (38) is especially clearly seen in this problem. According to (17), p^* in (30) is a subgradient of the sensibility function (1). Therefore, problem (30), which is given by the variational inequality $\langle p^*, y - y^*\rangle \ge 0, y \in Y$, is a necessary and sufficient condition for the sensibility function $\varphi(y)$ to have a minimum on Y. This means that complicated system (28), (29), and (30) (and, accordingly, (38)) is reduced to the simple problem of minimizing a convex sensibility function on the simple set Y. In fact, the sensibility function is a scalarization or convolution of the complicated problem and a reduction of the latter to a simple clear form. From an economic point of view, systems (28), (29), and (30) and (38) can be interpreted as follows. In the general case, they describe the interaction between two agents in various economic situations. Specifically, the logic of these systems can be traced in the well-known Arrow–Debreu model [13] in the case when the consumer and the producer are both represented by a single agent. These constructions can be independently viewed as mathematical models for describing demand-equal-to-supply balance interrelations for consumers and producers at different levels [12]. On the other hand, (28), (29), and (30) and (38) can be treated as a type of inverse optimization problems [14].

Now we discuss one interpretation of model (28), (29), and (30) in more detail. Let it be treated as a wholesale market model consisting of two agents, each seeking a maximum profit. In this case, all the partial problems in system (28), (29), and (30) are reduced to the maximization of concave functions, and the system as a whole becomes

$$w^* \in \operatorname{Argmax}\{f_1(w) \mid g(w) \le y^*, \quad w \in W_0\}, \tag{43}$$

$$\langle p - p^*, g(w^*) - y^* \rangle \ge 0, \quad p \le 0, \tag{44}$$

$$y^* \in \operatorname{Argmax}\{\langle p^*, y \rangle \mid y \in Y\}. \tag{45}$$

The first agent (45) provides the second one (43), (44) with the resource vector $y = y^* \in Y$, while the second, as a commodity producer, sets the price vector $p = p^* \ge 0$ (i.e., a Lagrange multiplier vector). The prices play the role of feedback. If the optimum $w^* \in W_0$ in (43) is strongly restricted by the ith constraint y_i^*, then the ith Lagrange multiplier p_i is sufficiently large, which means that the resource is in short supply and, therefore, is significantly needed. The first agent's profit $\langle p^*, y \rangle$ then grows substantially at the expense of y_i^*, because its weight coefficient is sufficiently large. In other words, the production of the scarcest commodities is automatically stimulated in system (43), (44), and (45), since a resource deficit (shortage) leads to an increase in the supplier's possible profit. A similar logic lies behind the more complicated problem (38). Here, the objective function of the first agent can be treated as the Lagrangian of a convex programming problem used as a model of a resource vector producer for the second agent.

3 Primal Extraproximal Method

Now we discuss methods for solving general system (33), (34), and (35). It was shown in the previous section that this problem is reduced to the computation of a saddle point of system (41) or (42). For illustrative purposes, we write this system once again:

$$w^*, y^* \in \operatorname{Argmin}\{f_1(w) + f_2(y) + \langle p^*, g(w) - h(y) \rangle \mid w \in W_0, y \in Y\},$$
$$\langle p - p^*, g(w^*) - h(y^*) \rangle \le 0, \quad p \ge 0. \tag{46}$$

The objective function of the first problem is separable. Consequently, it splits into two independent subproblems (see (40) and (41)):

$$f_1(w^*) + \langle p^*, g(w^*) \rangle \le f_1(w) + \langle p^*, g(w) \rangle, \quad w \in W_0,$$
$$f_2(y^*) - \langle p^*, h(y^*) \rangle \le f_2(y) - \langle p^*, h(y) \rangle, \quad y \in Y. \tag{47}$$

Taking into account this decomposition and the fact that the variational inequality in this problem can be equivalently represented as an operator equation, we rewrite system (46) in the form

$$w^* \in \operatorname{Argmin}\{f_1(w) + \langle p^*, g(w) \rangle \mid w \in W_0\},$$
$$y^* \in \operatorname{Argmin}\{f_2(y) - \langle p^*, h(y) \rangle \mid y \in Y\},$$
$$p^* = \pi_+(p^* + \alpha(g(w^*) - h(y^*))),$$

where $\pi_+(\cdots)$ is the projector of a vector onto the positive orthant. For the extremal mappings of this system to be nonexpansive operators in their

domains, we represent them in an equivalent form of proximal operators. Then the system becomes

$$w^* \in \operatorname{Argmin}\left\{\frac{1}{2}|w-w^*|^2 + \alpha(f_1(w) + \langle p^*, g(w)\rangle) \;\mid\; w \in W_0\right\},$$
$$y^* \in \operatorname{Argmin}\left\{\frac{1}{2}|y-y^*|^2 + \alpha(f_2(y) - \langle p^*, h(y)\rangle) \;\mid\; y \in Y\right\},$$
$$p^* = \pi_+(p^* + \alpha(g(w^*) - h(y^*))). \tag{48}$$

The simple iteration method is a natural approach to solving this system:

$$w^{n+1} \in \operatorname{Argmin}\left\{\frac{1}{2}|w-w^n|^2 + \alpha(f_1(w) + \langle p^n, g(w)\rangle) \;\mid\; w \in W_0\right\},$$
$$y^{n+1} \in \operatorname{Argmin}\left\{\frac{1}{2}|y-y^n|^2 + \alpha(f_2(y) - \langle p^n, h(y)\rangle) \;\mid\; y \in Y\right\},$$
$$p^{n+1} = \pi_+(p^n + \alpha(g(w^n) - h(y^n))).$$

However, in contrast to optimization problems, in equilibrium problems this method does not converge to the solution of the original system. For this reason, the solution is computed by the extraproximal methods described in [15, 16]. They can be treated as simple iteration methods with feedback [17].

3.1 Primal method

$$\bar{w}^n \in \operatorname{Argmin}\left\{\frac{1}{2}|w-w^n|^2 + \alpha(f_1(w) + \langle p^n, g(w)\rangle) \;\mid\; w \in W_0\right\},$$
$$\bar{y}^n \in \operatorname{Argmin}\left\{\frac{1}{2}|y-y^n|^2 + \alpha(f_2(y) - \langle p^n, h(y)\rangle) \;\mid\; y \in Y\right\},$$
$$p^{n+1} = \pi_+(p^n + \alpha(g(\bar{w}^n) - h(\bar{y}^n))),$$
$$w^{n+1} \in \operatorname{Argmin}\left\{\frac{1}{2}|w-w^n|^2 + \alpha(f_1(w) + \langle p^{n+1}, g(w)\rangle) \;\mid\; w \in W_0\right\},$$
$$y^{n+1} \in \operatorname{Argmin}\left\{\frac{1}{2}|y-y^n|^2 + \alpha(f_2(y) - \langle p^{n+1}, h(y)\rangle) \;\mid\; y \in Y\right\}. \tag{49}$$

For simplicity, the parameter $0 < \alpha < \alpha_0$ is chosen from a fixed interval. In the general case, the right-hand boundary of the interval can be estimated in the course of the iteration by using the technique described in [18].

To prove the convergence of process (49), it is equivalently represented in the form of inequalities that are convenient for deriving various estimates. Specifically, we use the inequality

$$\frac{1}{2}|z^* - x|^2 + \alpha_n f(z^*) \le \frac{1}{2}|z-x|^2 + \alpha_n f(z) - \frac{1}{2}|z-z^*|^2 \quad \forall z \in Z, \tag{50}$$

which is satisfied by any function of the form $\frac{1}{2}|z-x|^2+\alpha_n f(z)$. Here, $f(z)$ is a convex not necessarily differentiable function defined on the convex set Z, where $z\in Z$ and z^* is a minimizer of $\varphi(z)=\frac{1}{2}|z-x|^2+\alpha_n f(z)$ on Z for any x [18].

Since the objective functions in process (49) have the structure of function (50), this process can be written in the equivalent form

$$\begin{gathered}|\bar{w}^n-w^n|^2+2\alpha(f_1(\bar{w}^n)+\langle p^n,g(\bar{w}^n)\rangle)\le\\ \le|w-w^n|^2+2\alpha(f_1(w)+\langle p^n,g(w)\rangle)-|w-\bar{w}^n|^2,\\ |\bar{y}^n-y^n|^2+2\alpha(f_2(\bar{y}^n)-\langle p^n,h(\bar{y}^n)\rangle)\le\\ \le|y-y^n|^2+2\alpha(f_2(y)-\langle p^n,h(y)\rangle)-|y-\bar{y}^n|^2\end{gathered}\tag{51}$$

and

$$\begin{gathered}|w^{n+1}-w^n|^2+2\alpha(f_1(w^{n+1})+\langle p^{n+1},g(w^{n+1})\rangle)\le\\ \le|w-w^n|^2+2\alpha(f_1(w)+\langle p^{n+1},g(w)\rangle)-|w-w^{n+1}|^2,\\ |y^{n+1}-y^n|^2+2\alpha(f_2(y^{n+1})-\langle p^{n+1},h(y^{n+1})\rangle)\le\\ \le|y-y^n|^2+2\alpha(f_2(y)-\langle p^{n+1},h(y)\rangle)-|y-y^{n+1}|^2.\end{gathered}\tag{52}$$

According to [10], the operator equation in (49) is represented as the variational inequality

$$\langle p^{n+1}-p^n-\alpha(g(\bar{w}^n)-h(\bar{y}^n)),p-p^{n+1}\rangle\ge 0,\quad p\ge 0.\tag{53}$$

To prove the convergence of the processes, we use the following Lipschitz conditions for the vector functions $g(w),h(y)$:

$$|g(w+k)-g(w)|\le|g||k|,\ \ |h(y+k)-h(y)|\le|h||k|\tag{54}$$

for all $w+k\in W_0, y+k\in Y$, $k\in R^n$, where $|g|,|h|$ are Lipschitz constants.

To estimate the deviation of the vectors $\bar{w}^n$, w^{n+1}, $\bar{y}^n$, and y^{n+1} at every step in (49), we set $w=w^{n+1}$, $w=\bar{w}^n$ and $y=y^{n+1}$, $y=\bar{y}^n$ in (51) and (52), respectively. Then

$$\begin{gathered}|\bar{w}^n-w^n|^2+2\alpha(f_1(\bar{w}^n)+\langle p^n,g(\bar{w}^n)\rangle)\le\\ \le|w^{n+1}-w^n|^2+2\alpha(f_1(w^{n+1})+\langle p^n,g(w^{n+1})\rangle)-|w^{n+1}-\bar{w}^n|^2,\\ |\bar{y}^n-y^n|^2+2\alpha(f_2(\bar{y}^n)-\langle p^n,h(\bar{y}^n)\rangle)\le\\ \le|y^{n+1}-y^n|^2+2\alpha(f_2(y^{n+1})-\langle p^n,h(y^{n+1})\rangle)-|y^{n+1}-\bar{y}^n|^2\end{gathered}$$

and

$$\begin{gathered}|w^{n+1}-w^n|^2+2\alpha(f_1(w^{n+1})+\langle p^{n+1},g(w^{n+1})\rangle)\le\\ \le|\bar{w}^n-w^n|^2+2\alpha(f_1(\bar{w}^n)+\langle p^{n+1},g(\bar{w}^n)\rangle)-|\bar{w}^n-w^{n+1}|^2,\\ |y^{n+1}-y^n|^2+2\alpha(f_2(y^{n+1})-\langle p^{n+1},h(y^{n+1})\rangle)\le\\ \le|\bar{y}^n-y^n|^2+2\alpha(f_2(\bar{y}^n)-\langle p^{n+1},h(\bar{y}^n)\rangle)-|\bar{y}^n-y^{n+1}|^2.\end{gathered}$$

Summing up the resulting inequalities yields

$$|\bar{w}^n - w^{n+1}|^2 \le \alpha\langle p^{n+1} - p^n, g(\bar{w}^n) - g(w^{n+1})\rangle,$$
$$|\bar{y}^n - y^{n+1}|^2 \le \alpha\langle p^{n+1} - p^n, h(y^{n+1}) - h(\bar{y}^n)\rangle.$$

In view of (54), we finally obtain

$$|\bar{w}^n - w^{n+1}| \le \alpha|g||p^{n+1} - p^n|,\ |\bar{y}^n - y^{n+1}| \le \alpha|h||p^{n+1} - p^n|. \tag{55}$$

Let us prove the following convergence theorem for method (49).

Theorem 2. *If equilibrium problem (33), (34), and (35) has a solution, $f_1(w), f_2(y)$, $g(w)$ are convex functions, $h(y)$ is a concave function, the vector functions satisfy Lipschitz conditions (54), and W_0 and Y are convex closed sets, then the sequence p^n, w^n, y^n generated by the primal extraproximal method (49) with α satisfying $0 < \alpha < 1/\sqrt{2(|g|^2 + |h|^2)}$ converges monotonically in norm to one of the solutions of the problem.*

Proof. The iterations of process (49) with respect to w and y have an identical structure and form. Therefore, any transformation of the formulas with respect to w gives a similar result to that with respect to y. Below are some transformations of (51) and (52) with respect to w. Setting $w = w^*$ in (52) and $w = w^{n+1}$ in (51) yields

$$|w^{n+1} - w^n|^2 + 2\alpha(f_1(w^{n+1}) + \langle p^{n+1}, g(w^{n+1})\rangle) \le$$
$$\le |w^* - w^n|^2 + 2\alpha(f_1(w^*) + \langle p^{n+1}, g(w^*)\rangle) - |w^{n+1} - w^*|^2$$

and

$$|\bar{w}^n - w^n|^2 + 2\alpha(f_1(\bar{w}^n) + \langle p^n, g(\bar{w}^n)\rangle) \le$$
$$\le |w^{n+1} - w^n|^2 + 2\alpha(f_1(w^{n+1}) + \langle p^n, g(w^{n+1})\rangle) - |\bar{w}^n - w^{n+1}|^2.$$

Adding the relation

$$\langle p^{n+1}, g(\bar{w}^n)\rangle - \langle p^{n+1}, g(\bar{w}^n)\rangle = 0$$

to both inequalities and summing them up, we obtain

$$|w^{n+1} - w^*|^2 + |w^{n+1} - \bar{w}^n|^2 + |\bar{w}^n - w^n|^2 +$$
$$+2\alpha(\langle p^n, g(\bar{w}^n)\rangle - \langle p^{n+1}, g(\bar{w}^n)\rangle - \langle p^n, g(w^{n+1})\rangle + \langle p^{n+1}, g(w^{n+1})\rangle) +$$
$$+2\alpha(f_1(\bar{w}^n) - f_1(w^*)) + 2\alpha(\langle p^{n+1}, g(\bar{w}^n)\rangle - \langle p^{n+1}, g(w^*)\rangle) \le |w^n - w^*|^2$$

or

$$|w^{n+1} - w^*|^2 + |w^{n+1} - \bar{w}^n|^2 + |\bar{w}^n - w^n|^2 + 2\alpha\langle p^n - p^{n+1}, g(\bar{w}^n) - g(w^{n+1})\rangle +$$
$$+2\alpha(f_1(\bar{w}^n) - f_1(w^*)) + 2\alpha(\langle p^{n+1}, g(\bar{w}^n)\rangle - \langle p^{n+1}, g(w^*)\rangle) \le |w^n - w^*|^2. \tag{56}$$

The same argument applied to the inequalities in (51) and (52) with respect to y gives a similar estimate

$$|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2\alpha\langle p^n-p^{n+1},h(\bar{y}^n)-h(y^{n+1})\rangle+ \\ +2\alpha(f_2(\bar{y}^n)-f_2(y^*))-2\alpha(\langle p^{n+1},h(\bar{y}^n)\rangle-\langle p^{n+1},h(y^*)\rangle)\le|y^n-y^*|^2. \tag{57}$$

Setting $w=\bar{w}^n$ in (47) gives

$$f_1(w^*)+\langle p^*,g(w^*)\rangle\le f_1(\bar{w}^n)+\langle p^*,g(\bar{w}^n)\rangle.$$

Adding this inequality to (56),

$$|w^{n+1}-w^*|^2+|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2+2\alpha\langle p^n-p^{n+1},g(\bar{w}^n)- \\ -g(w^{n+1})\rangle+2\alpha\langle p^{n+1}-p^*,g(\bar{w}^n)-g(w^*)\rangle\le|w^n-w^*|^2. \tag{58}$$

In view of (54), the fourth term in (58) is estimated as

$$|w^{n+1}-w^*|^2+|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2- \\ -2(\alpha|g|)^2|p^n-p^{n+1}|^2+2\alpha\langle p^{n+1}-p^*,g(\bar{w}^n)-g(w^*)\rangle\le|w^n-w^*|^2. \tag{59}$$

Returning to estimate (57), we repeat similar manipulations. Specifically, setting $y=\bar{y}^n$ in (47) produces

$$f_2(y^*)-\langle p^*,h(y^*)\rangle\le f_2(\bar{y}^n)-\langle p^*,h(\bar{y}^n)\rangle.$$

Adding this inequality to (57), we obtain

$$|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2\alpha\langle p^n-p^{n+1},h(\bar{y}^n)-h(y^{n+1})\rangle- \\ -2\alpha\langle p^{n+1}-p^*,h(\bar{y}^n)-h(y^*)\rangle\le|y^n-y^*|^2. \tag{60}$$

In view of (54), the fourth term in (60) is estimated as

$$|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2(\alpha|h|)^2|p^n-p^{n+1}|^2- \\ -2\alpha\langle p^{n+1}-p^*,h(\bar{y}^n)-h(y^*)\rangle\le|y^n-y^*|^2. \tag{61}$$

Adding (59) and (61) gives

$$|w^{n+1}-w^*|^2+|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2-2(\alpha|g|)^2|p^n-p^{n+1}|^2+ \\ +|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2(\alpha|h|)^2|p^n-p^{n+1}|^2+ \\ +2\alpha\langle p^{n+1}-p^*,g(\bar{y}^n)-g(y^*)-h(\bar{y}^n)+h(y^*)\rangle\le|w^n-w^*|^2+|y^n-y^*|^2. \tag{62}$$

A similar estimate is derived for the iteration with respect to $p\ge 0$ in (49). Specifically, setting $p=p^*$ in (53) and $p=p^{n+1}$ in (46) and summing the resulting inequalities, we obtain

$$\langle p^{n+1}-p^n,p^*-p^{n+1}\rangle-\alpha\langle g(\bar{w}^n)-h(\bar{y}^n),p^*-p^{n+1}\rangle+\alpha\langle g(w^*)- \\ -h(y^*),p^*-p^{n+1}\rangle\ge 0$$

or

$$-2\langle p^{n+1}-p^n, p^*-p^{n+1}\rangle - 2\alpha\langle g(w^*)-g(\bar{w}^n), p^*-p^{n+1}\rangle - \\ -2\alpha\langle h(\bar{y}^n)-h(y^*), p^*-p^{n+1}\rangle \le 0. \tag{63}$$

Adding (62) and (63),

$$|w^{n+1}-w^*|^2+|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2-2(\alpha|g|)^2|p^n-p^{n+1}|^2+ \\ +|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2(\alpha|h|)^2|p^n-p^{n+1}|^2+ \\ -2\langle p^{n+1}-p^n, p^*-p^{n+1}\rangle \le |w^n-w^*|^2+|y^n-y^*|^2. \tag{64}$$

Next, using the identity

$$|x_1-x_3|^2=|x_1-x_2|^2+2\langle x_1-x_2, x_2-x_3\rangle+|x_2-x_3|^2, \tag{65}$$

we rearrange the scalar product into the sum of squares

$$|w^{n+1}-w^*|^2+|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2-2(\alpha|g|)^2|p^n-p^{n+1}|^2+ \\ +|y^{n+1}-y^*|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2-2(\alpha|h|)^2|p^n-p^{n+1}|^2+ \\ |p^{n+1}-p^*|^2+|p^{n+1}-p^n|^2 \le |w^n-w^*|^2+|y^n-y^*|^2+|p^n-p^*|^2. \tag{66}$$

Therefore,

$$|w^{n+1}-w^*|^2+|y^{n+1}-y^*|^2+|p^{n+1}-p^*|^2+(1-2\alpha^2(|g|^2+|h|^2))|p^{n+1}-p^n|^2+ \\ +|w^{n+1}-\bar{w}^n|^2+|\bar{w}^n-w^n|^2+|y^{n+1}-\bar{y}^n|^2+|\bar{y}^n-y^n|^2 \le \\ \le |w^n-w^*|^2+|y^n-y^*|^2+|p^n-p^*|^2. \tag{67}$$

Summing up this inequality from $n=0$ to $n=N$ gives

$$|w^{N+1}-w^*|^2+|y^{N+1}-y^*|^2+|p^{N+1}-p^*|^2+d\sum_{k=0}^{k=N}|p^{k+1}-p^k|^2+ \\ +\sum_{k=0}^{k=N}(|w^{k+1}-\bar{w}^k|^2+|\bar{w}^k-w^k|^2+|y^{k+1}-\bar{y}^k|^2+|\bar{y}^k-y^k|^2) \le \\ \le |w^0-w^*|^2+|y^0-y^*|^2+|p^0-p^*|^2,$$

where $d=1-2\alpha^2(|g|^2+|h|^2)>0$. The resulting inequality implies that the trajectory is bounded, i.e.,

$$|w^{N+1}-w^*|^2+|y^{N+1}-y^*|^2+|p^{N+1}-p^*|^2 \le |w^0-w^*|^2+|y^0-y^*|^2+|p^0-p^*|^2,$$

and it also implies the convergence of the series: $\sum_{k=0}^{\infty}|p^{k+1}-p^k|^2<\infty, \sum_{k=0}^{\infty}|w^{k+1}-\bar{w}^k|^2<\infty$, $\sum_{k=0}^{\infty}|\bar{w}^k-w^k|^2<\infty$, $\sum_{k=0}^{\infty}|y^{k+1}-\bar{y}^k|^2<\infty$, $\sum_{k=0}^{\infty}|\bar{y}^k-y^k|^2<\infty$. Therefore,

$$|p^{n+1}-p^n|^2\to 0,\ |w^{n+1}-\bar{w}^n|^2\to 0,\ |\bar{w}^n-w^n|^2\to 0,$$
$$|y^{n+1}-\bar{y}^n|^2\to 0,\ |\bar{y}^n-y^n|^2\to 0,\ n\to\infty.$$

Since the sequence p^n, w^n, y^n is bounded, there exists an element p', w', y' such that $p^{n_i}\to p',\ w^{n_i}\to w',\ y^{n_i}\to y'$ as $n_i\to\infty$. Moreover,

$$|p^{n_i+1}-p^{n_i}|^2\to 0,\ |w^{n_i+1}-\bar{w}^{n_i}|^2\to 0,\ |\bar{w}^{n_i}-w^{n_i}|^2\to 0,$$
$$|y^{n_i+1}-\bar{y}^{n_i}|^2\to 0,\ \bar{y}^{n_i}-y^{n_i}|^2\to 0.$$

Passage to the limit as $n_i\to\infty$ in (51), (52), and (53) yields

$$f_1(w')+\langle p', g(w')\rangle\le f_1(w)+\langle p', g(w)\rangle,$$
$$f_2(y')+\langle p', g(w')\rangle\le f_2(y)+\langle p', g(y)\rangle,$$
$$\langle g(w')-h(y'), p-p'\rangle\ge 0$$

for all $w\in\Omega$, $y\in Y$, $p\ge 0$.

Since these relations are equivalent to (46), we conclude that $w'=w^*\in\Omega^*$, $p'=p^*\ge 0$, and $y'=y^*\in Y$, i.e., any limit point of the sequence p^n, w^n, y^n is a solution to the problem. Since $|w^n-w^*|+|p^n-p^*|+|y^n-y^*|$ decreases monotonically, the limit point is unique; i.e., $p^n\to p^*, w^n\to w^*,\ y^n\to y^*$ as $n\to\infty$. The theorem is proved.

4 Dual Extraproximal Method

Along with the primal method considered in the previous section, the dual extraproximal approach [15, 16] can be used to solve system (33), (34), and (35). Its formulas are as follows.

4.1 Dual Method

$$\bar{p}^n=\pi_+(p^n+\alpha(g(w^n)-h(y^n))),$$
$$w^{n+1}\in\operatorname{argmin}\left\{\frac{1}{2}|w-w^n|^2+\alpha(f_1(w)+\langle\bar{p}^n, g(w)\rangle)\ \mid\ w\in W_0\right\},$$
$$y^{n+1}\in\operatorname{argmin}\left\{\frac{1}{2}|y-y^n|^2+\alpha(f_2(y)-\langle\bar{p}^n, h(y)\rangle)\ \mid\ y\in Y\right\},$$
$$p^{n+1}=\pi_+(p^n+\alpha(g(w^{n+1})-h(y^{n+1}))). \tag{68}$$

If the vector $y^*\in Y$ in original problem (33), (34), and (35) is a constant, i.e., Y is a singleton, then there are no iterative formulas with respect to y. In this case, we have formulas for computing a saddle point of (33), (34). On the contrary, if the convex programming problem degenerates and is absent, then (68) contains the formulas only with respect to y, and this subprocess converges to a solution of problem (30), i.e., to a boundary point of Y that

is a support point for the linear functional $\langle p^*, y\rangle,\ y \in Y$, where p^* is an a priori given vector.

Process (48) can be represented in the form of the inequalities

$$\langle \bar{p}^n - p^n - \alpha(g(w^n) - h(y^n)), p - \bar{p}^n\rangle \geq 0, \quad p \geq 0, \tag{69}$$

$$|w^{n+1} - w^n|^2 + 2\alpha(f_1(w^{n+1}) + \langle \bar{p}^n, g(w^{n+1})\rangle) \leq$$
$$\leq |w - w^n|^2 + 2\alpha(f_1(w) + \langle \bar{p}^n, g(w)\rangle) - |w - w^{n+1}|^2, \quad w \in W_0, \tag{70}$$

$$|y^{n+1} - y^n|^2 + 2\alpha(f_2(y^{n+1}) - \langle \bar{p}^n, h(y^{n+1})\rangle) \leq$$
$$\leq |y - y^n|^2 + 2\alpha(f_2(y) - \langle \bar{p}^n, h(y)\rangle) - |y - y^{n+1}|^2, \quad y \in Y, \tag{71}$$

$$\langle p^{n+1} - p^n - \alpha(g(w^{n+1}) - h(y^{n+1})), p - p^{n+1}\rangle \geq 0, \quad p \geq 0. \tag{72}$$

To estimate the deviations of $\bar{p}^n$ from p^{n+1}, we compare the first and last equations in (68) to obtain

$$|\bar{p}^n - p^{n+1}| \leq \alpha|g(w^n) - h(y^n) - g(w^{n+1}) + h(y^{n+1})|. \tag{73}$$

Let us prove a convergence theorem for method (68).

Theorem 3. *If equilibrium problem (33), (34), and (35) has a solution, $f_1(w), f_2(y)$, $g(w)$ are convex functions, $h(y)$ is a concave function, the vector functions satisfy Lipschitz condition (73), and W_0 and Y are convex closed sets, then the sequence p^n, w^n, y^n generated by the dual extraproximal method (68) with α satisfying $0 < \alpha < \min\{1/(2|g|), 1/(2|h|)\}$ converges monotonically in norm to one of the solutions of the problem.*

Proof. To estimate the deviations of the residuals at the point w^*, y^*, we set $w = w^*$ in (70) and $y = y^*$ in (71). Then

$$|w^{n+1} - w^n|^2 + 2\alpha(f_1(w^{n+1}) + \langle \bar{p}^n, g(w^{n+1})\rangle) \leq$$
$$\leq |w^* - w^n|^2 + 2\alpha(f_1(w^*) + \langle \bar{p}^n, g(w^*)\rangle) - |w^* - w^{n+1}|^2$$

and

$$|y^{n+1} - y^n|^2 + 2\alpha(f_2(y^{n+1}) - \langle \bar{p}^n, h(y^{n+1})\rangle) \leq$$
$$\leq |y^* - y^n|^2 + 2\alpha(f_2(y^*) - \langle \bar{p}^n, h(y^*)\rangle) - |y^* - y^{n+1}|^2.$$

Summing both inequalities gives

$$|w^* - w^{n+1}|^2 + |y^* - y^{n+1}|^2 + |w^{n+1} - w^n|^2 + |y^{n+1} - y^n|^2 +$$
$$+2\alpha(f_1(w^{n+1}) + f_2(y^{n+1}) + \langle \bar{p}^n, g(w^{n+1}) - h(y^{n+1})\rangle) \leq$$
$$\leq |w^* - w^n|^2 + |y^* - y^n|^2 + 2\alpha(f_1(w^*) + f_2(y^*) + \langle \bar{p}^n, g(w^*) - h(y^*)\rangle).$$

A similar estimate of the deviation can be obtained at the point w^{n+1}, y^{n+1}. For this purpose, we set $w = w^{n+1}$ and $y = y^{n=1}$ in the right inequality in (42). Then

$$f_1(w^*) + f_2(y^*) + \langle p^*, g(w^*) - h(y^*)\rangle \leq$$
$$\geq f_1(w^{n+1}) + f_2(y^{n+1}) + \langle p^*, g(w^{n+1}) - h(y^{n+1})\rangle.$$

Summing the last two inequalities, we obtain

$$|w^* - w^{n+1}|^2 + |y^* - y^{n+1}|^2 + |w^{n+1} - w^n|^2 + |y^{n+1} - y^n|^2 +$$
$$+2\alpha(\langle \bar{p}^n - p^*, g(w^{n+1}) - h(y^{n+1}) - g(w^*) + h(y^*)\rangle) \leq$$
$$\leq |w^* - w^n|^2 + |y^* - y^n|^2. \tag{74}$$

Now we consider the inequalities with respect to p. Setting $p = p^*$ in (72) and $p = p^{n+1}$ in (69) yields

$$\langle p^{n+1} - p^n - \alpha(g(w^{n+1}) - h(y^{n+1})), p^* - p^{n+1}\rangle \geq 0,$$
$$\langle \bar{p}^n - p^n - \alpha(g(w^n) - h(y^n)), p^{n+1} - \bar{p}^n\rangle \geq 0.$$

Summing these inequalities, we have

$$\langle p^{n+1} - p^n, p^* - p^{n+1}\rangle + \langle \bar{p}^n - p^n, p^{n+1} - \bar{p}^n\rangle - \alpha\langle g(w^{n+1}) - h(y^{n+1}), p^* -$$
$$-p^{n+1}\rangle + \alpha\langle g(w^{n+1}) - g(w^n) - h(y^{n+1}) + h(y^n), p^{n+1} - \bar{p}^n\rangle -$$
$$-\alpha\langle g(w^{n+1}) - h(y^{n+1}), p^{n+1} - \bar{p}^n\rangle \geq 0.$$

The third term is added to the fifth one, while the fourth is estimated with the help of (73) to obtain

$$\langle p^{n+1} - p^n, p^* - p^{n+1}\rangle + \langle \bar{p}^n - p^n, p^{n+1} - \bar{p}^n\rangle -$$
$$-\alpha\langle g(w^{n+1}) - h(y^{n+1}), p^* - \bar{p}^n\rangle + \alpha^2|g(w^{n+1}) - g(w^n) - h(y^{n+1}) +$$
$$+h(y^n)|^2 \geq 0.$$

Setting $p = \bar{p}^n$ in (34),

$$-\langle \bar{p}^n - p^*, g(w^*) - h(y^*)\rangle \geq 0.$$

Summing the last two inequalities gives

$$2\langle p^{n+1} - p^n, p^* - p^{n+1}\rangle + 2\langle \bar{p}^n - p^n, p^{n+1} - \bar{p}^n\rangle +$$
$$+2\alpha^2|g(w^{n+1}) - g(w^n) - h(y^{n+1}) + h(y^n)|^2 - 2\alpha\langle g(w^{n+1}) - h(y^{n+1}) -$$
$$-g(w^*) + h(y^*), p^* - \bar{p}^n\rangle \geq 0. \tag{75}$$

Finally, adding (74) to (75), we obtain

$$|w^{n+1} - w^*|^2 + |w^{n+1} - w^n|^2 + |y^{n+1} - y^*|^2 + |y^{n+1} - y^n|^2 +$$
$$-2\langle p^{n+1} - p^n, p^* - p^{n+1}\rangle - 2\langle \bar{p}^n - p^n, p^{n+1} - \bar{p}^n\rangle -$$
$$-2\alpha^2|g(w^{n+1}) - g(w^n) - h(y^{n+1}) + h(y^n)|^2 \leq |w^n - w^*|^2 + |y^n - y^*|^2. \tag{76}$$

By using identity (65), the fifth and sixth terms are rearranged into

$$|w^{n+1}-w^*|^2+|y^{n+1}-y^*|^2+|w^{n+1}-w^n|^2+|y^{n+1}-y^n|^2+|p^{n+1}-p^*|^2,$$
$$|p^{n+1}-\bar{p}^n|^2+|\bar{p}^n-p^n|^2-2\alpha^2|g(w^{n+1})-g(w^n)-h(y^{n+1})+h(y^n)|^2\leq$$
$$\leq|w^n-w^*|^2+|p^n-p^*|^2+|y^n-y^*|^2. \tag{77}$$

The last term on the left-hand side of inequality (77) is estimated using $2\langle x,y\rangle\leq|x|^2+|y|^2$ and (55):

$$|g(w^{n+1})-g(w^n)-h(y^{n+1})+h(y^n)|^2=|g(w^{n+1})-g(w^n)|^2-$$
$$-2\langle g(w^{n+1})-g(w^n),h(y^{n+1})-h(y^n)\rangle+|h(y^n)-h(y^{n+1})|^2.$$

Rewriting (77) once again and using this estimate, we have

$$|w^{n+1}-w^*|^2+|y^{n+1}-y^*|^2+|p^{n+1}-p^*|^2+d_1|w^{n+1}-w^n|^2+$$
$$+d_2|y^{n+1}-y^n|^2+|p^{n+1}-\bar{p}^n|^2+|\bar{p}^n-p^n|^2\leq|w^n-w^*|^2+|p^n-p^*|^2+$$
$$+|y^n-y^*|^2, \tag{78}$$

where $d_1=1-4\alpha^2|g|^2>0$, $d_2=1-4\alpha^2|h|^2>0$. Both conditions are satisfied if $0<\alpha<\min\{1/(2|g|),1/(2|h|)\}$. All the terms on the left-hand side of (78) are then positive, and the resulting inequality is similar to (67). The rest of the proof is analogous to that of Theorem 1.

5 Conclusions

In this chapter we investigated the properties of sensitivity function for convex programming problem more detailed, proposed a new view to this function as a natural convolution for system of optimization problems. For solving this system it is offered primal and dual solution methods. The convergence of them is proved.

References

1. Gale, D.: The Theory of Linear Economic Models, McGraw-Gill, New York, NY (1960)
2. Williams, A.C.: Marginal values in linear programming. J. Soc. Indust. Appl. Math. 11, 82–94 (1963)
3. Zlobec, S.: Stable Parametric Programming, Kluwer, Dordrecht (2001)
4. Eremin, I.I., Astafiev, N.N.: Introduction to Linear and Convex Programming Theory, Nauka, Moscow (1976)
5. Elster, K.-H., Reinhardt, R., Schaubel, M., Donath, G.: Einfuhrung in die nicht-lineare optimierung, BSB B.G. Teubner, Leipzig (1977)
6. Rzhevskii, S.V.: Monotone Methods of Convex Programming, Naukova Dumka, Kiev (1993)

7. Golikov, A.I.: Characterization of optimal estimates set for multicriterial optimization problems. U.S.S.R. Comput. Math. Math. Phys. 28(10), 1285–1296 (1988)
8. Zhadan, V.G.: Modified Lagrangian method for multicriterial optimization. U.S.S.R. Comput. Math. Math. Phys. 28(11), 1603–1617 (1988)
9. Rockafellar, R.T., Wets, R.J-B.: Variational Analysis, Springer, Berlin (1998)
10. Vasil'ev, F.P.: Optimization Methods, Factorial Press, Moscow (2002)
11. Antipin, A.S.: Methods of solving systems of convex programming problems. Zhurnal Vychisl. Mat. Mat. Phiziki. 27(3), 368–376 (1987) English transl. U.S.S.R. Comput. Math. Math. Phys. 27(2), 30–35 (1987)
12. Antipin, A.S.: Models of interaction between manufacturers, consumers, and the transportation system. Avtomatika i telemechanika. 10, 105–113 (1989). English transl. Autom Remote Control., 1391–1398 (1990)
13. Karlin, S.: Mathematical Methods and Theory in Games, Programming, and Economics, Pergamon Press, London (1959)
14. Antipin, A.S.: Inverse optimization problem. Economic-mathematical cyclopaedic dictionary. "Large Russian Encyclopaedia". "Infra-M", 346–347 (2003)
15. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games. Comput. Math. Math. Phys. 45(11), 1893–1914 (2005)
16. Antipin, A.S.: An extraproximal method for solving equilibrium programming problems and games with coupled constraints. Comput. Math. Math. Phys. 45(12), 2020–2022 (2005)
17. Antipin, A.S.: Controlled proximal differential systems for saddle problems. The Differentsial'nye Uravneniya. 28(11), 1846–1861 (1992). English transl.: Differ. Equ. 28(11), 1498–1510 (1992)
18. Antipin, A.S.: The convergence of proximal methods to fixed points of extremal mappings and estimates of their rate of convergence. Zhurnal Vychisl. Mat. Mat. Fiz. 35(5), 688–704 (1995) English transl. Comput. Math. Math. Phys. 35(5), 539–551 (1995)

Post-optimal Analysis of Linear Semi-infinite Programs

Miguel A. Goberna

Department of Statistics and Operations Research, University of Alicante, 03080 Alicante, Spain
mgoberna@ua.es

Summary. Linear semi-infinite programming (LSIP) deals with linear optimization problems in which either the dimension of the decision space or the number of constraints (but not both) is infinite. In most applications of LSIP to statistics, electronics, telecommunications, and other fields, all the data (or at least part of them) are uncertain. Post-optimal analysis provides answer to questions about the quantitative impact on the optimal value of small perturbations of the data (sensitivity analysis) and also about the continuity properties of the optimal value, the optimal set, and the feasible set (stability analysis) around the nominal problem. This chapter surveys the state of the art in sensitivity and stability analysis in LSIP.

Key words: linear semi-infinite programming, linear inequality systems, stability analysis, sensitivity analysis

1 Introduction

Let T be an infinite set, $a : T \mapsto \mathbb{R}^n$, $b : T \mapsto \mathbb{R}$, and $c \in \mathbb{R}^n$. Then

$$P : \begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & c'x := \sum\limits_{i=1}^{n} c_i x_i \\ \text{s.t.} & a_t' x \geq b_t,\ t \in T, \end{array} \tag{1}$$

is called a (primal) *linear semi-infinite programming* (LSIP in brief) *problem* because the number of variables is finite whereas the set of constraints is infinite. The mappings a and b are called *left-* and *right-hand side* (LHS and RHS) functions whereas c is called *cost vector*. We denote by σ, F, and F^* the constraint system, the feasible set, and the optimal set of P, respectively. By definition, the optimal value of P is $v(P) = +\infty$ when $F = \emptyset$ (in which case σ and P are called *inconsistent*). P is *solvable* when $F^* \neq \emptyset$ and it is *bounded* if $v(P) \in \mathbb{R}$. P is said to be *continuous* when T is a compact Hausdorff

This work was supported by MICINN of Spain, Grant MTM2008-06695- C03-01.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_2,

topological space, $a \in \mathcal{C}(T)^n$ and $b \in \mathcal{C}(T)$. A classical application of LSIP consists in the best approximation of a given function $f \in \mathcal{C}([\alpha, \beta])$ from the linear hull of a finite family $\{v_1, \ldots, v_n\} \subset \mathcal{C}([\alpha, \beta])$, this linear space being equipped with either the L_1 or the L_∞ norm.

Example 1. Let us consider the L_1 approximation of f from above. Taking into account the constraint $f(t) \leq \sum_{i=1}^n v_i(t)x_i$ for all $t \in [\alpha, \beta]$,

$$\left\| f - \sum_{i=1}^n x_i v_i \right\|_1 = \int_\alpha^\beta \left[\sum_{i=1}^n v_i(t)x_i - f(t) \right] dt$$
$$= \sum_{i=1}^n \left(\int_\alpha^\beta v_i(t)dt \right) x_i - \int_\alpha^\beta f(t)dt.$$

Let $c_i := \int_\alpha^\beta v_i(t)dt, i = 1, \ldots, n$. A best L_1 approximation to f from above is given by $\sum_{i=1}^n \overline{x}_i v_i$, where $\overline{x}$ is an optimal solution of the continuous LSIP problem

$$P: \operatorname*{Min}_{x \in \mathbb{R}^n} c'x$$
$$\text{s.t.} \quad \sum_{i=1}^n v_i(t)x_i \geq f(t), \quad t \in [\alpha, \beta].$$

Example 2. A best uniform approximation to f is obtained by minimizing the uniform error for the linear combinations of $\{v_1, \ldots, v_n\}$, i.e., solving

$$P: \operatorname*{Min}_{(x,y) \in \mathbb{R}^{n+1}} y$$
$$\text{s.t.} \quad -y \leq f(s) - \sum_{i=1}^n v_i(s)x_i \leq y, \quad s \in [\alpha, \beta].$$

Since P can be written in the form of (1), with $T := [\alpha, \beta] \times \{1, 2\}$ compact Hausdorff in $\mathbb{R}^2$ and the functions $a_{(s,j)} := \left((-1)^j v_1(s), \ldots, (-1)^j v_n(s), 1 \right)$ and $b_{(s,j)} := (-1)^j f(s)$, $(s, j) \in T$, continuous on T, P turns out to be a continuous LSIP problem.

Getting stopping rules before optimality requires the availability of some dual problem maximizing lower bounds for $c'x$, $x \in F$. The easiest way to do that consists of considering the *space of generalized finite sequences*

$$\mathbb{R}^{(T)} := \left\{ \lambda \in \mathbb{R}^T \mid |\operatorname{supp} \lambda| < \infty \right\},$$

where

$$\operatorname{supp} \lambda := \{t \in T \mid \lambda_t \neq 0\}$$

denotes the *supporting set* of λ. We represent by $\mathbb{R}_+^{(T)}$ the positive cone in $\mathbb{R}^{(T)}$. Given $\lambda \in \mathbb{R}_+^{(T)}$ such that $c = \sum_{t \varepsilon T} \lambda_t a_t$, multiplying both members of this equation by $x \in F$ we get

$$c'x = \sum_{t\varepsilon T} \lambda_t a_t' x \geq \sum_{t\varepsilon T} \lambda_t b_t.$$

The *Haar's dual problem* of P is then

$$\begin{aligned} D: \max_{\lambda \in \mathbb{R}_+^{(T)}} & \sum_{t\in T} \lambda_t b_t \\ \text{s.t.} & \sum_{t\varepsilon T} \lambda_t a_t = c, \end{aligned}$$

with feasible and optimal sets Λ and Λ^*, respectively, and optimal value $v(D) = -\infty$ when $\Lambda = \emptyset$. In contrast to ordinary linear programming (LP), even though both problems of the pair $P - D$ are bounded, the *duality gap* is possibly non-zero, i.e.,

$$\delta(P, D) := v(P) - v(D) \geq 0.$$

It can be shown that D is equivalent to other well-known dual problems as the Lagrange and the Rockafellar ones (which are the result of aggregating to Λ dominated solutions). If P is continuous, then another dual problem, called *continuous dual*, can be obtained by replacing, in D, $\lambda \in \mathbb{R}_+^{(T)}$ with $\mu \in \mathcal{C}_+'(T)$ (the cone of non-negative regular Borel measures on T) and $\sum_{t\varepsilon T}$ with $\int_T$:

$$\begin{aligned} D_0: \max_{\mu \in \mathcal{C}_+'(T)} & \int_T b_t \, d\mu(t) \\ \text{s.t.} & \int_T a_t d\mu(t) = c. \end{aligned}$$

Because the elements of $\mathbb{R}_+^{(T)}$ such that $|\operatorname{supp} \lambda| = 1$ can be interpreted as atomic measures, $0 \leq \delta(P, D_0) \leq \delta(P, D)$. We could have $\delta(P, D_0) < \delta(P, D)$ for some particular problem but all the known duality theorems guaranteeing the existence of a zero duality gap have the same hypotheses for both dual problems, D_0 and D. Thus the LSIP problems D_0 and D (observe that they have finitely many constraints and infinitely many variables) are also equivalent in practice.

The first LSIP problem (a dual one) was formulated by George Dantzig, in 1939, in order to solve a problem related with the Neyman–Pearson lemma (for details, see [50]). Dantzig understood that Λ is polyhedral-like and conceived the way (his famous geometry of columns) to improve the objective functional by jumping from one of its extreme points to an adjacent one. Dantzig re-started his research in 1945, when he was asked to mechanize the planning of the postwar Pentagon activities. In 1947 he discussed his simplex method (inspired in the geometry of columns) and the duality theory with von Neumann in Princeton and 1 year later he presented the new ideas to the mathematical community in the meeting of the Econometric Society held in Wisconsin, 1948 (the so-called MP0 conference). Although with some precedents such as Haar's seminal works on the constraint system of P published in Hungarian in the 1920s, the research carried out by Dantzig on D, and the

optimality conditions of John for differentiable nonlinear SIP [76], the first papers on LSIP, conceived as a natural extension of LP, are due to Charnes, Cooper, and Kortanek. In [30–32] these authors coined the term LSIP and gave the first duality theorem. The development of LSIP during the 1960s is described in detail in [80].

Concerning the numerical treatment of LSIP problems, with the precedent of Remez method for the Chebyshev approximation problem of Example 2, the first numerical methods were proposed by Gustafson and Kortanek during the 1970s [65, 66, 68]. According to the literature (see [50] and references therein), the most efficient numerical approach to LSIP combines a discretization method (phase 1) with the reduction of P to a nonlinear ordinary system by using the KKT conditions for LSIP problems (phase 2). Discretization consists of solving a sequence of finite subproblems (replacing the index set T in (1) by a finite subset at each iteration) and terminates at some approximate (generally infeasible) optimal solution which is sufficient in many engineering applications that do not require an accurate optimal solution. The subsets of T either can be the terms of some predetermined sequence of grids (e.g., regular grids) or can be obtained by adding a new cutting plane at each step (a constraint violated by the optimal solution of the current LP subproblem) and eliminating constraints detected as irrelevant. At the moment, the most efficient method for phase 1 seems to be the LSIP version of Elzinga–Moore LP method proposed in [6], where the current iterate is the center of the greatest ball contained in the current polytope (which includes some level set of P). Discretization methods converge fast when P has a *strongly unique optimal solution*, i.e., there exists $x^* \in F$ and $\alpha > 0$ such that $c'x \geq c'x^* + \alpha \|x - x^*\|$ for all $x \in F$. A common hypothesis of the convergence theorems for discretization methods is the continuity of P (without this assumption phase 1 can be performed with simplex-like methods whose convergence is dubious). Reduction requires strong assumptions, e.g., the existence of a suitable representation of T (in many real applications, T is a box in some finite dimensional Euclidean space) and the continuity of the coefficients of the constraints with respect to the index t. The nonlinear system arising in phase 2 is usually solved by means of some Newton-like method with quadratic or at least superlinear convergence starting from the approximate optimal solution computed in phase 1 [67].

LSIP methods have been successfully applied during the last years in order to solve LSIP (generally primal continuous) problems arising in statistics [40], machine learning [3, 74, 86, 95], optimal design [73, 101, 102], functional approximation [36, 37, 103], spectrometry [33], control problems [75], variational problems [39], semi-definite programming [82, 83], combinatorial optimization [84], environmental sciences [72, 100], different types of ordinary optimization problems with uncertain data [4, 63, 85, 89], and finance [99]. Authors working in the last field have also numerically solved LSIP dual problems [87, 88]. This chapter is motivated by the observation that, although in most of the

mentioned applications all the data (or at least part of them) are uncertain, no paper has taken this fact into account. Thus the primary purpose of this chapter is to fill the existing gap between the theoretical works on uncertain LSIP problems (most of them about stability theory) and LSIP applications.

Optimization problems with uncertain data can be handled from different perspectives: post-optimal analysis (which deals with the behavior of the optimal value, the optimal set, and the feasible set when some of the data in the *nominal problem* P are the object of small perturbations), robust optimization (which provides risk-averse decisions; see, e.g., [5]), stochastic optimization (where the perturbable data are interpreted as random variables; see, e.g., [94], and references therein), fuzzy optimization (which interprets such perturbable data as fuzzy numbers; see, e.g., [91], and interval optimization (which considers that the perturbable data could take arbitrary values on given intervals; see, e.g., [69]). The viability of these and other alternative approaches depends on the tractability of the auxiliary problems to be solved. Although we are primarily interested in the post-optimal approach, we discuss here the tractability of certain robust, stochastic, fuzzy, and interval models for the LSIP problem P in (1), when the source of uncertainty is the cost vector $c \in \mathbb{R}^n$, the constraint system $\{a_t'x \geq b_t, t \in T\}$, and both:

♦ If $c \in C \subset \mathbb{R}^n$ whereas the constraints remain fixed, the robust counterpart of P consists of minimizing the worst possible value of $c'x$, i.e.,

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{Min}} & \max\limits_{c\in C} c'x \\ \text{s.t.} & a_t'x \geq b_t, t \in T, \end{array}$$

or, equivalently, embedding the problem in higher dimension,

$$\begin{array}{ll} \min\limits_{(x,y)\in\mathbb{R}^{n+1}} & y \\ \text{s.t.} & y - c'x \geq 0, \quad c \in C, \\ & a_t'x \geq b_t, \quad t \in T. \end{array} \tag{2}$$

Thus, if P is continuous (for any $c \in C$) and C is a compact subset of $\mathbb{R}^n$, then the robust counterpart of P is also a continuous LSIP problem.

In the interval optimization approach, the uncertain set is $C = \prod_{i=1}^{n} [l_i, u_i]$, with $l_i < u_i$, $i = 1, \ldots, n$, and the problem consists of determining the range of the optimal value for all the instances of the uncertain problem P. In other words, we have to solve both the optimistic and pessimistic counterparts of P. Because C is the convex hull of its extreme points, $\{y - c'x \geq 0, c \in C\}$ can be replaced, in the pessimistic counterpart (2), by a subsystem of 2^n linear constraints whereas the optimistic counterpart of P reads

$$\min_{x\in F, z\in C} z'x,$$

which can be reformulated as the non-convex quadratic SIP problem

$$\begin{array}{ll} \min\limits_{(x,z)\in\mathbb{R}^{2n}} & z'x \\ \text{s.t.} & l_i \le z_i \le u_i, \quad i=1,\ldots,n, \\ & a_t'x \ge b_t, \quad t\in T. \end{array}$$

The natural stochastic interpretation of P is the uncertain LSIP problem

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & \gamma'x \\ \text{s.t.} & a_t'x \ge b_t, \quad t\in T, \end{array} \tag{3}$$

where $\gamma=(\gamma_1,\ldots,\gamma_n)$ is a random vector taking values on C with a given probability $\mathbb{P}$. Then each realization of γ (usually generated via simulation) provides a different LSIP problem called *scenario* which is continuous whenever the nominal problem is continuous. Solving a large number of these scenario programs it is possible to get an empirical probabilistic distribution of the optimal value of P.

In the fuzzy perspective it is again assumed that $C=\prod_{i=1}^n [l_i,u_i]$, with $l_i<u_i$, $i=1,\ldots,n$; moreover, the random variables γ_i in (3) are assumed to have special types of distributions on $[l_i,u_i]$ called *fuzzy numbers* (e.g., either trapezoidal or triangular distributions). So, the fuzzy counterpart of P can be seen as a particular class of stochastic counterpart.

♦If $(a_t,b_t)\in S_t\subset\mathbb{R}^{n+1}$ for all $t\in T$, whereas c is fixed, the robust approach requires to guarantee the feasibility of the selected decision under any conceivable circumstance, i.e., the robust counterpart of P is now the LSIP problem

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & c'x \\ \text{s.t.} & a'x\ge b, \quad (a,b)\in \bigcup\limits_{t\in T} S_t, \end{array} \tag{4}$$

whose index set is, in general, non-compact even though P is continuous and each set S_t is compact in $\mathbb{R}^{n+1}$. This is the case in the interval approach, where each S_t is assumed to be a box in $\mathbb{R}^{n+1}$.

♦ From the stochastic perspective the robust counterpart (4) can be interpreted as the uncertain LSIP problem

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & c'x \\ \text{s.t.} & \delta'\begin{pmatrix} x \\ -1 \end{pmatrix} \ge 0, \quad \delta\in\Delta, \end{array}$$

where δ is a random vector taking values on $\Delta=\bigcup_{t\in T} S_t\subset\mathbb{R}^{n+1}$ with probability $\mathbb{P}$. Taking N values of δ at random on Δ with probability $\mathbb{P}$, say $\delta_{(1)},\ldots,\delta_{(N)}$, we get the *scenario program*

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & c'x \\ \text{s.t.} & \delta_{(i)}'\begin{pmatrix} x \\ -1 \end{pmatrix} \ge 0, \quad i=1,\ldots,N, \end{array}$$

which is an ordinary LP problem (obviously, each scenario program can be seen as a discretization of the robust counterpart of P, (4), generated at random instead of by using grids or cutting planes). In general, the optimal solution, x_N^*, of a scenario program is not necessarily a feasible solution of (4). Let

$$V\left(x_N^*\right) = \mathbb{P}\left\{\delta \in \Delta \mid \delta' \begin{pmatrix} x_N^* \\ -1 \end{pmatrix} < 0\right\}$$

be the *violation probability* of x_N^*. According to Theorem 1 in [9], the inequality

$$\mathbb{P}\left\{V\left(x_N^*\right) > \varepsilon\right\} \leq \sum_{i=0}^{n-1} \binom{N}{i} \varepsilon^i \left(1-\varepsilon\right)^{N-i} \tag{5}$$

holds for any $\varepsilon > 0$ (this admissible violation probability is selected by the decision maker). Moreover, Campi and Garatti [9] also show that (5) holds with equality under mild conditions.

♦ Finally in this discussion, if $c \in C \subset \mathbb{R}^n$ and $(a_t, b_t) \in S_t \subset \mathbb{R}^{n+1}$ for all $t \in T$, combining the previous ideas, the robust counterpart of P is the LSIP problem

$$\begin{array}{ll} \min\limits_{(y,x)\in\mathbb{R}^{1+n}} & y \\ \text{s.t.} & y - c'x \geq 0, \quad c \in C, \\ & a'x \geq b, \quad (a,b) \in \bigcup\limits_{t\in T} S_t, \end{array} \tag{6}$$

which does not retain the desirable continuity property of P. Program (6) can be interpreted as the uncertain LSIP problem

$$\begin{array}{ll} \min\limits_{(y,x)\in\mathbb{R}^{1+n}} & \begin{pmatrix} 1 \\ 0_n \end{pmatrix}' \begin{pmatrix} y \\ x \end{pmatrix} \\ \text{s.t.} & \delta' \begin{pmatrix} y \\ x \\ -1 \end{pmatrix} \geq 0, \quad \delta \in \Delta, \end{array}$$

where δ is a random vector taking values on

$$\Delta = \left[\{1\} \times (-C) \times \{0\}\right] \cup \left[\{0\} \times \left(\bigcup_{t\in T} S_t\right)\right]$$

with probability $\mathbb{P}$. Taking N values of δ at random on Δ with probability $\mathbb{P}$, we get the corresponding scenario program and, maintaining the notation of the previous case, we have, again by [9], the tight bound

$$\mathbb{P}\left\{V\left(x_N^*\right) > \varepsilon\right\} \leq \sum_{i=0}^{n} \binom{N}{i} \varepsilon^i \left(1-\varepsilon\right)^{N-i},$$

for any given $\varepsilon > 0$ and for any x_N^* optimal solution of the scenario program.

♦ From now on we focus on the main purpose of this chapter, which is intended to survey the state of the art in post-optimal analysis (stability and sensitivity) of LSIP problems arising in practice. Frequently a proper subset of the triple (a, b, c) can be perturbed due to either measurement errors or rounding errors occurring during the computation process. The practitioner should identify the sources of uncertainty and then apply the known results for the corresponding model. For instance, let us analyze the possible sources of uncertainty of problem P in Example 1, where $T = [\alpha, \beta]$, $a_t = (v_1(t), \ldots, v_n(t))$, $b_t = f(t)$, and $c_i = \int_\alpha^\beta v_i(t)dt$, $i = 1, \ldots, n$, in the notation of (1). If $v_1, \ldots, v_n$ are polynomials and α and β are integer numbers, then the only source of uncertainty is f, i.e., the RHS function b. If f is also polynomial but α and β are irrational numbers, then the source of uncertainty is the cost vector c (whose components are computed with quadrature rules). Most commonly, if all the involved functions are non-polynomial, all the data in P can be perturbed, so that the perturbations could affect all the elements of the triple (a, b, c). Analogously, the uncertainty in the LSIP problem of Example 2 can be caused by the LHS function a, by the RHS function b, or by the pair (a, b). Nevertheless, in this example the perturbations are linked, e.g., concerning b we must have $b_{(s,1)} + b_{(s,2)} = 0$ for all $s \in [\alpha, \beta]$.

Sensitivity analysis allows the prediction (or at least the estimation) of the quantitative impact on the optimal value of small perturbations of the data. Stability analysis informs about the continuity properties of the optimal value, the optimal set, and the feasible set as functions of the data. The works on parametric LSIP published during the 1980s ([8, 41, 96, 97], etc.) dealt with the continuity properties of the primal optimal value, the optimal set, and the feasible set in continuous LSIP. The first extension of these results to general LSIP was obtained in the mid-1990s. These results have been completed during the 2000s. The recent research in this area is mostly focussed on the obtainment of quantitative information related with computational issues (well posedness and error bounds), developing stability analysis of special LSIP models arising in practice and extending sensitivity analysis tools from LP to LSIP.

The main aim of this survey is to convince the practitioners that it is desirable (and frequently possible) to include post-optimal analysis in real applications of LSIP involving uncertain data and the secondary aim to encourage the theoretical research in this area of optimization.

The chapter is organized as follows. Section 2 introduces the necessary notation and a relatively exhaustive list of concepts about LSIP, extended functions, and set-valued mappings allowing the understanding of the survey by non-specialists. In Sections 3–6 we suppose that the admissible perturbations preserve the structure of the nominal problem, i.e., that they provide LSIP problems posed in the same space of variables $\mathbb{R}^n$ and having the same index set T. Moreover, we also assume that, if the nominal problem P is continuous, then the admissible perturbations also provide continuous problems (in each section we consider first results on general LSIP and then the continuous

counterparts). We only consider the four perturbation models corresponding to the types of admissible perturbations more frequently encountered in practice: perturbations of all the data (Section 3), simultaneous perturbations of the RHS function and the cost vector (Section 4), and separate perturbations of the RHS function and the cost vector (Sections 5 and 6, respectively). Finally, Section 7 contains a list of open problems, a sketch of other models, and the conclusions.

2 Preliminaries

First we introduce some notation. 0_n and 0_T denote the null vectors in $\mathbb{R}^n$ and $\mathbb{R}^{(T)}$, respectively. The Euclidean, the l_∞ (or Chebyshev), and the l_1 norms in $\mathbb{R}^n$ are represented by $\|\cdot\|$, $\|\cdot\|_\infty$, and $\|\cdot\|_1$, respectively, with associated distances d, d_∞, and d_1. $|X|$ denotes the cardinality of a set X. Given $X \neq \emptyset$ contained in a real linear space, by $\operatorname{aff} X$, $\operatorname{span} X$, and $\operatorname{conv} X$ we denote the affine hull, the linear hull, and the convex hull of X, respectively. The conical convex hull of $X \cup \{0_n\}$ is represented by $\operatorname{cone} X$. Moreover, if X is convex, $\dim X$ and $\operatorname{extr} X$ denote the dimension and the set of extreme points of X, respectively. From the topological side, if X is a subset of some topological space, $\operatorname{int} X$, cl X, and $\operatorname{bd} X$ represent the interior, the closure, and the boundary of X, respectively. If $X \neq \emptyset$ is a subset of some topological vector space, $\operatorname{rint} X$ denotes the relative interior of X (i.e., the interior of X in the topology induced on $\operatorname{aff} X$) and $X_\infty := \left\{\lim_k \mu_k x^k \mid \left\{x^k\right\} \subset X, \{\mu_k\} \downarrow 0\right\}$ its asymptotic cone. Finally, if $(X, \|\cdot\|)$ is a normed space, the *dual norm* on its topological dual X^* is $\|u\|^* = \sup_{\|x\| \leq 1} |u(x)|$.

2.1 Basic Concepts on Sets and Mappings

Let $\{X_r\}$ be a sequence of non-empty sets in $\mathbb{R}^n$. We denote by $\liminf_r X_r$ ($\limsup_r X_r$) the set formed by all the possible limits (cluster points, respectively) of sequences $\{x_r\}$ such that $x_r \in X_r$ for all $r \in \mathbb{N}$. When these two limit sets are non-empty and coincide, then it is said that $\{X_r\}$ converges in the *Painlevé–Kuratowski sense* to the set

$$\lim_r X_r := \liminf_r X_r = \limsup_r X_r.$$

Let X be a topological space and let $f : X \mapsto \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. The *domain* of f is $\operatorname{dom} f := \{x \in X \mid f(x) \in \mathbb{R}\}$. f is called *lower semicontinuous* (lsc) at $x_0 \in X$ if for each scalar $\gamma < f(x_0)$ there exists an open set $V \subset X$, containing x_0, such that $\gamma < f(x)$ for each $x \in V$. f is *upper semicontinuous* (usc) at $x_0 \in X$ if $-f$ is lsc at x_0.

The *directional derivative* of f at $x_0 \in X$ (linear space) in the direction $v \in X$ is

$$f'(x_0;v) := \lim_{\varepsilon \searrow 0} \frac{f(x_0+\varepsilon v) - f(x_0)}{\varepsilon}.$$

The (convex) *subdifferential* of $f : X \mapsto \overline{\mathbb{R}}$ at $x_0 \in X$ (a topological vector space) such that $f(x_0) \in \mathbb{R}$ is

$$\partial f(x_0) := \{u \in X^* \mid f(x) \geq f(x_0) + u(y - x_0) \ \ \forall x \in X\},$$

where X can be replaced by $\operatorname{dom} f$.

The subdifferential of a concave function f is the (convex) subdifferential of $-f$.

Now consider a given set-valued mapping $\mathcal{M} : X \rightrightarrows Y$, where X and Y are pseudometric spaces equipped with pseudometrics d_X and d_Y, respectively. The *domain* of $\mathcal{M}$ is $\operatorname{dom} \mathcal{M} := \{x \in X \mid \mathcal{M}(x) \neq \emptyset\}$.

$\mathcal{M} : X \rightrightarrows Y$ is *lower semicontinuous* at $x_0 \in X$ *in the Berge–Kuratowski sense* (lsc in brief) if, for each open set $W \subset Y$ such that $W \cap \mathcal{M}(x_0) \neq \emptyset$, there exists an open set $V \subset X$, containing x_0, such that $W \cap \mathcal{M}(x) \neq \emptyset$ for each $x \in V$. The intuitive meaning is that $\mathcal{M}$ does not implode in the proximity of x_0.

$\mathcal{M}$ is *upper semicontinuous* at $x_0 \in X$ *in the Berge–Kuratowski sense* (usc) if, for each open set $W \subset Y$ such that $\mathcal{M}(x_0) \subset W$, there exists an open set $V \subset X$, containing x_0, such that $\mathcal{M}(x) \subset W$ for each $x \in V$. This means that $\mathcal{M}$ does not burst in the proximity of x_0.

$\mathcal{M}$ is *closed* at $x_0 \in \operatorname{dom} \mathcal{M}$ if for all sequences $\{x_r\}_{r=1}^{\infty} \subset X$ and $\{y_r\}_{r=1}^{\infty} \subset Y$ satisfying $y_r \in \mathcal{M}(x_r)$ for all $r \in \mathbb{N}$, $\lim_{r\to\infty} x_r = x_0$ and $\lim_{r\to\infty} y_r = y_0$ and one has $y_0 \in \mathcal{M}(x_0)$.

$\mathcal{M}$ is lsc (usc, closed) if it is lsc (usc, closed) at x for all $x \in X$. Obviously, $\mathcal{M}$ is closed if and only if its graph

$$\operatorname{gph} \mathcal{M} := \{(x,y) \in X \times Y \mid y \in \mathcal{M}(x)\}$$

is a closed set.

$\mathcal{M}$ is *metrically regular* (or *pseudo-Lipschitz*) at $(x_0, y_0) \in \operatorname{gph} \mathcal{M}$ if there exists $L > 0$ and two open sets V and W such that $x_0 \in V \subset X$ and $y_0 \in W \subset Y$, and

$$d_X\left(x, \mathcal{M}^{-1}(y)\right) \leq L d_Y(y, \mathcal{M}(x)) \tag{7}$$

for all $x \in V, y \in W$. This means that $\mathcal{M}^{-1}(y)$ does not change abruptly when y changes slightly in the proximity of y_0. For this reason, in that case, $\mathcal{M}^{-1}$ is said to be *Aubin continuous* at (y_0, x_0). The smallest L satisfying (7) is called the *regularity modulus* of $\mathcal{M}$ at (x_0, y_0), represented by $\operatorname{reg} \mathcal{M}(x_0 \mid y_0)$.

2.2 Basic Concepts and Results on LSIP

We associate with problem P in (1) (actually with its constraint system σ) its *set of coefficients*

$$C := \{(a_t, b_t), t \in T\} \subset \mathbb{R}^{n+1},$$

its *characteristic cone*

$$K := \operatorname{cone}\{C \cup \{(0_n, -1)\}\},$$

and its *first moment cone*

$$M := \operatorname{cone}\{a_t, t \in T\}$$

(which is the orthogonal projection of K on the hyperplane $x_{n+1} = 0$).

$\overline{x} \in \mathbb{R}^n$ is a *strong Slater* (SS in brief) element for P if there exists $\varepsilon > 0$ such that $a_t'\overline{x} \geq b_t + \varepsilon$ for all $t \in T$. If P is continuous, then $\overline{x} \in \mathbb{R}^n$ is an SS element for P if and only if it is a *Slater element* (i.e., $a_t'\overline{x} > b_t$ for all $t \in T$). P satisfies the Slater (SS) condition if there exists some Slater (SS) element. P satisfies the SS condition if and only if $v(P_{\mathrm{SS}}) > 0$, where

$$P_{\mathrm{SS}}: \begin{array}{ll} \max\limits_{(x,y)\in\mathbb{R}^{n+1}} & y \\ \text{s.t.} & a_t'x \geq b_t + y, \quad t \in T, \end{array}$$

so that we can conclude that $v(P_{\mathrm{SS}}) > 0$ from any feasible solution of P_{SS} such that $y > 0$ (i.e., it is not necessary to solve the associated LSIP problem P_{SS} until optimality). Observe that P_{SS} is continuous if and only if P is continuous.

The following results are well known (see, e.g., [50]): σ is inconsistent if and only if $(0_n, 1) \in \operatorname{cl} K$ and it contains a finite inconsistent subsystem if and only if $(0_n, 1) \in K$. The non-homogeneous Farkas lemma establishes that a linear inequality $w'x \geq \gamma$ is a consequence of a consistent system σ (i.e., $w'x \geq \gamma$ for all $x \in F$) if and only if $(w, \gamma) \in \operatorname{cl} K$. If P is continuous and satisfies the Slater condition then K is closed. The first duality theorem establishes that, if P is consistent and K is closed, then $\delta(P, D) = 0$ and D is solvable whereas the second one asserts that, if $c \in \operatorname{rint} M$, then $\delta(P, D) = 0$ and P is solvable.

The classical approach to sensitivity analysis for LP problems, based on the computation of optimal basis by means of some variant of the simplex method, allows to predict the optimal value under separate perturbations of the cost and the RHS vectors, and its extension to LSIP is possible by using the duality theorems. The modern approach is based on the computation of optimal partitions by means of interior point methods, and it allows to predict the optimal value under simultaneous perturbations of the costs and the RHS vectors. Its generalization to LSIP requires suitable extensions of the concept of optimal and maximal partitions from LP to LSIP [56]. A couple $(x, \lambda) \in F \times \Lambda$ is called a *complementary solution of* $P - D$ if $x \in F^*$, $\lambda \in \Lambda^*$, and $\delta(P, D) = 0$, i.e., if $\operatorname{supp} x \cap \operatorname{supp} \lambda = \emptyset$, where $\operatorname{supp} x := \{t \in T \mid a_t'x > b_t\}$ is the *supporting set* of $x \in F$. Consequently, given a point $\overline{x} \in F$, there exists $\overline{\lambda} \in \Lambda$ such that $\left(\overline{x}, \overline{\lambda}\right)$ is a complementary solution of

$P - D$ if and only if $\overline{x}$ is an optimal solution for some finite subproblem of P.

A triple $(B, N, Z) \in \left(2^T\right)^3$ is called an *optimal partition* if there exists a complementary solution (x, λ) such that $B = \operatorname{supp} x$, $N = \operatorname{supp} \lambda$, and $Z = T \setminus (B \cup N)$. Obviously, the non-empty elements of the *tripartition* (B, N, Z) give a partition of T. We say that a tripartition $\left(\overline{B}, \overline{N}, \overline{Z}\right)$ is *maximal* if

$$\overline{B} = \bigcup_{x \in F^*} \operatorname{supp} x, \qquad \overline{N} = \bigcup_{\lambda \in \Lambda^*} \operatorname{supp} \lambda, \quad \text{and} \quad \overline{Z} = T \setminus (\overline{B} \cup \overline{N}).$$

The uniqueness of the maximal partition is a straightforward consequence of the definition. If there exists an optimal solution pair $\left(\overline{x}, \overline{\lambda}\right) \in F^* \times \Lambda^*$ such that $\operatorname{supp} \overline{x} = \overline{B}$ and $\operatorname{supp} \overline{\lambda} = \overline{N}$, then the maximal partition is called the *maximal optimal partition*. If $\left(\overline{B}, \overline{N}, \overline{Z}\right)$ is an optimal partition such that $\overline{Z} = \emptyset$, then it is a maximal optimal partition. Assuming the existence of a complementary solution (i.e., that $\delta(P, D) = 0$ and $F^* \neq \emptyset \neq \Lambda^*$), then there exists a maximal optimal partition if and only if the sets of extreme points and extreme directions of Λ^* are finite.

2.3 Perturbed LSIP Problems

The stability analysis of the nominal problem P, identified with the nominal data $\pi := (a, b, c)$, requires the embedding of π in some space of all admissible perturbations (called *parameters*), Π, and to equip it with some topology. Since the 1980s, there is a consensus about the convenience of equipping Π with the *topology of the uniform convergence*, corresponding to the following pseudodistance on Π: given two parameters, $\pi_1 = \left(a^1, b^1, c^1\right)$ and $\pi_2 = \left(a^2, b^2, c^2\right)$, the pseudodistance between π_1 and π_2 is

$$d_\infty(\pi_1, \pi_2) := \max \left\{ \left\| c^1 - c^2 \right\|_\infty, \sup_{t \in T} \left\| \begin{pmatrix} a_t^1 \\ b_t^1 \end{pmatrix} - \begin{pmatrix} a_t^2 \\ b_t^2 \end{pmatrix} \right\|_\infty \right\}.$$

Observe that we can have $d_\infty(\pi_1, \pi_2) = +\infty$. From now on the parameters and the corresponding primal and dual problems will be distinguished with the same index (e.g., the problems associated with π_1 are P_1 and D_1).

In particular, if all the data are uncertain (as in Section 3), $\Pi = \left(\mathbb{R}^n \times \mathbb{R}\right)^T \times \mathbb{R}^n$ in the general case and $\Pi = \mathcal{C}(T)^n \times \mathcal{C}(T) \times \mathbb{R}^n$ in the continuous case. If only b and c are variable (Section 4), $\Pi = \mathbb{R}^T \times \mathbb{R}^n$ in the general case and $\Pi = \mathcal{C}(T) \times \mathbb{R}^n$ in the continuous case. If only b is variable (Section 5), then $\Pi = \mathbb{R}^T$ in the general case and $\Pi = \mathcal{C}(T)$ in the continuous case. Finally, if only c is variable (Section 6), then $\Pi = \mathbb{R}^n$ (in this model there is no distinction between the general and the continuous cases). In general LSIP Π is a linear space equipped with the pseudometric d_∞ whereas it is a Banach space in the continuous case.

Important subspaces of Π are those formed by the consistent (inconsistent, solvable, bounded, unbounded) primal problems, which are denoted by $\Pi_{\mathrm{C}}^{\mathrm{P}}$ ($\Pi_{\mathrm{I}}^{\mathrm{P}}$, $\Pi_{\mathrm{S}}^{\mathrm{P}}$, $\Pi_{\mathrm{B}}^{\mathrm{P}}$, $\Pi_{\mathrm{U}}^{\mathrm{P}}$, respectively). Similarly, the sets of consistent (inconsistent, solvable, bounded, unbounded) dual problems are denoted by $\Pi_{\mathrm{C}}^{\mathrm{D}}$ ($\Pi_{\mathrm{I}}^{\mathrm{D}}$, $\Pi_{\mathrm{S}}^{\mathrm{D}}$, $\Pi_{\mathrm{B}}^{\mathrm{D}}$, $\Pi_{\mathrm{U}}^{\mathrm{D}}$, respectively). A desirable property is *generic* on one of these sets when it holds on some open dense subset. In other words, if the nominal parameter belongs to the first set, then there exist arbitrary close parameters satisfying the corresponding property stably (i.e., in some neighborhood).

An optimization problem is *ill-posed* relative to certain desirable property when sufficiently small perturbations of the data provide problems enjoying this property and others, where the property fails. In some fields of mathematical programming as LP or conic programming, the distance to ill-posedness (the supremum of the size of the perturbations preserving certain property as consistency or solvability) is related with measures of conditioning, complexity analysis of numerical algorithms, and metric regularity (see, e.g., [38, 92]). The following sets of ill-posed problems have been considered in the literature on LSIP: $\mathrm{bd}\,\Pi_{\mathrm{C}}^{\mathrm{P}}$ is the set of *ill-posed problems in the feasibility sense*, $\mathrm{bd}\,\Pi_{\mathrm{SI}}^{\mathrm{P}}$ (where $\Pi_{\mathrm{SI}}^{\mathrm{P}}$ denotes the set of problems which have a finite inconsistent subproblem) is the set of *generalized ill-posed problems in the feasibility sense*, and $\mathrm{bd}\,\Pi_{\mathrm{S}}^{\mathrm{P}} = \mathrm{bd}\,\Pi_{\mathrm{B}}^{\mathrm{P}}$ is the set of *ill-posed problems in the optimality sense* (other sets of ill-posed problems in LSIP are discussed in [58]). The distance from the nominal problem π to a set of ill-posed parameters can be interpreted as a measure of well-posedness.

We associate with each $\pi_1 \in \Pi$ the primal problem P_1 and its dual problem D_1. The primal and dual optimal value mappings are ϑ^{P}, $\vartheta^{\mathrm{D}} : \Pi \mapsto \overline{\mathrm{R}}$ such that $\vartheta^{\mathrm{P}}(\pi_1) = v(P_1)$ and $\vartheta^{\mathrm{D}}(\pi_1) = v(D_1)$. The relevant set-valued mappings are the primal feasible set and the primal optimal set mappings, $\mathcal{F}$, $\mathcal{F}^* : \Pi \rightrightarrows \mathbb{R}^n$ such that $\mathcal{F}(\pi_1)$ and $\mathcal{F}^*(\pi_1)$ are the feasible and the optimal sets of P_1, respectively, and their dual counterparts, $\mathcal{L}$, $\mathcal{L}^* : \Pi \rightrightarrows \mathbb{R}^{(T)}$ such that $\mathcal{L}(\pi_1)$ and $\mathcal{L}^*(\pi_1)$ are the feasible and the optimal sets of D_1. The inconvenience with $\mathcal{L}$ and $\mathcal{L}^*$ is the non-intrinsic character of the topologies on $\mathbb{R}^{(T)}$. The few papers dealing with the stability of the dual mappings consider the L_1 and the L_∞ norms because they provide results that are symmetric to those of the primal mappings $\mathcal{F}$ and $\mathcal{F}^*$.

♦ From (7), $\mathcal{F}$ is Aubin continuous at (π, x) if and only if there exists $L > 0$ and two open sets V and W such that $x \in V \subset \mathbb{R}^n$ and $\pi \in W \subset \Pi$,

$$d_\infty(x_1, \mathcal{F}(\pi_1)) \leq L d_\infty\left(\pi_1, \mathcal{F}^{-1}(x_1)\right)$$

(where $\mathcal{F}^{-1}(x) = \{\pi_1 \in \Pi \mid x \in \mathcal{F}(\pi_1)\}$) for all $x_1 \in V$ and $\pi_1 \in W$.

The following definition of well-posedness orientated toward the stability of the primal optimal value function $\vartheta^{\mathrm{P}} : \{x^r\} \subset \mathbb{R}^n$ is an *asymptotically minimizing sequence* for $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ associated with $\{\pi_r\} \subset \Pi_{\mathrm{B}}^{\mathrm{P}}$ if $x^r \in \mathcal{F}(\pi_r)$ for all r, $\lim_r \pi_r = \pi$, and $\lim_r \left[(c^r)' x^r - \vartheta^{\mathrm{P}}(\pi_r)\right] = 0$. In particular, $\pi \in \Pi_{\mathrm{S}}^{\mathrm{P}}$ is *Hadamard well-posed* (Hwp in brief) if for every $x^* \in F^*$ and for every

$\{\pi_r\} \subset \Pi_{\mathrm{B}}^{\mathrm{P}}$ such that $\lim_r \pi_r = \pi$ there exists an asymptotically minimizing sequence converging to x^*.

Obviously, if we fix successive elements of the triple (a, b, c), the sufficient conditions for the stability properties of $\mathcal{F}$, $\mathcal{F}^*$, $\mathcal{L}$, or $\mathcal{L}^*$ at the nominal data are still sufficient. For instance, if $\mathcal{F}$ is lsc at $\pi = (a, b, c)$ under arbitrary perturbations of all the data, the same is true for perturbations limited to the RHS function and/or the cost vector. In particular, if we fix a and b (a and c), $\mathcal{F}$ ($\mathcal{L}$, respectively) is constant and, so, it is trivially lsc, usc, and closed. Conversely, any necessary condition for one of the above mappings to be lsc at π under perturbations of a part of the data is also necessary for the lsc property of the corresponding set-valued mapping under arbitrary perturbations of all the data.

3 Perturbing All the Data

This model is the most frequently encountered in the recent literature on stability in LSIP. One of the reasons is that the characterizations of different stability properties in this model become sufficient conditions for the remaining models and sometimes these conditions are also necessary. Analogously, the formulae providing the distance to ill-posedness are at least upper bounds in other models whereas the error bound is still valid (although they could be improved). Few sensitivity analysis results have been published on this model (it is difficult to predict the behavior of the optimal value under perturbations of the LHS function even in ordinary LP). Recall that $\mathcal{F}(\pi) = F$ is the feasible set of the nominal problem and $\mathcal{F}^*(\pi) = F^*$ is its optimal set.

3.1 Stability of the Feasible Set

It is easy to prove that $\mathcal{F}$ is closed everywhere whereas the lsc and the usc properties are satisfied or not at $\pi = (a, b, c) \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ depending on the nominal data a and b.

The basic result on stability analysis in LSIP is the following (non-exhaustive) list of characterizations of the lower semicontinuity of $\mathcal{F}$ [49–51, 61]:

- $\blacklozenge$ $\mathcal{F}$ is lsc at $\pi \in \operatorname{dom} \mathcal{F} = \Pi_{\mathrm{C}}^{\mathrm{P}}$
- $\Leftrightarrow$ $\pi \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}$ (*stable consistency*)
- $\Leftrightarrow$ the SS condition holds
- $\Leftrightarrow 0_{n+1} \notin \operatorname{cl} \operatorname{conv} C$
- $\Leftrightarrow \forall \{\pi_r\} \subset \Pi$ such that $\lim_r \pi_r = \pi$ $\exists r_0 \in \mathbb{N}$ such that $\lim_{r \geq r_0} \mathcal{F}(\pi_r) = \mathcal{F}(\pi)$
- $\Leftrightarrow \exists$ an open set V, $\pi \in V \subset \Pi$, such that $\dim \mathcal{F}(\pi_1) = \dim F$ $\forall \pi_1 \in V$
- $\Leftrightarrow \exists$ an open set V, $\pi \in V \subset \Pi$, such that $\operatorname{aff} \mathcal{F}(\pi_1) = \operatorname{aff} F$ $\forall \pi_1 \in V$

In the case that $0_n \notin \operatorname{bd} \operatorname{conv}\{a_t, t \in T\}$, the following condition is also equivalent to the lower semicontinuity of $\mathcal{F}$ at π: $\exists$ an open set V, $\pi \in V \subset \Pi$ such that $\mathcal{F}(\pi_1)$ is homeomorphic to F for all $\pi_1 \in V$.

It has also been proved [23] that, if $\mathcal{F}$ is lsc at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$, then $\mathcal{F}^{-1}$ is metrically regular at (x, π) for all $x \in F$. The converse is not true.

The characterization of the usc property of $\mathcal{F}$ at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ in [18] (refining previous results in [52]) requires some additional notation. Let K^R be the characteristic cone of the linear system

$$\sigma^R := \{a'x \geq b, (a, b) \in (\operatorname{conv} C)_\infty\}. \tag{8}$$

Observe that any inequality $a'x \geq b$ in (8) is a consequence of σ because $\binom{a}{b} \in \operatorname{cl} K^R$. If F is bounded, then $\mathcal{F}$ is usc at π. Otherwise two cases are possible:

- If F contains at least one line, then $\mathcal{F}$ is usc at π if and only if $K^R = \operatorname{cl} K$.
- Otherwise (i.e., if $\dim \operatorname{span}\{a_t, t \in T\} = n$), selecting some vector w that is the sum of some basis of $\mathbb{R}^n$ contained in $\{a_t, t \in T\}$, $\mathcal{F}$ is usc at π if and only if there exists $\beta \in \mathbb{R}$ such that

$$\operatorname{cone}\left(K^R \cup \{(w, \beta)\}\right) = \operatorname{cone}\left(\operatorname{cl} K \cup \{(w, \beta)\}\right).$$

The stability properties of $\mathcal{F}$ are closely related with those corresponding to the boundary and the extreme point set mappings: $\mathcal{B}, \mathcal{E} : \Pi \rightrightarrows \mathbb{R}^n$ such that $\mathcal{B}(\pi_1) := \operatorname{bd} \mathcal{F}(\pi_1)$ and $\mathcal{E}(\pi_1) := \operatorname{extr} \mathcal{F}(\pi_1)$ for all $\pi_1 \in \Pi$. The transmission of stability properties between $\mathcal{F}$, $\mathcal{B}$, and $\mathcal{E}$ [45, 46, 55] has been used in order to provide a sufficient condition for the stable containment of solution sets of LSISs of two [44] in the following sense: let π and τ be two given two LSISs with associated feasible sets $\mathcal{F}$ and $\mathcal{G}$; $\mathcal{F}$ is said to be contained in $\mathcal{G}$ stably at (π, τ) if $\mathcal{F}(\pi_1) \subset \mathcal{G}(\tau_1)$ for π_1 and τ_1 close enough to π and τ, respectively. Analogously, we say that $\mathcal{F}$ intersects $\mathcal{G}$ stably at (π, τ) if $\mathcal{F}(\pi_1) \cap \mathcal{G}(\tau_1) \neq \emptyset$ for π_1 and τ_1 close enough to π and τ, respectively. The stability of the containment of the feasible set of a given linear (convex) system in the feasible set of a similar system [62] and the stability of their intersection [47] have been analyzed.

Concerning the stability of the dual feasible set mapping $\mathcal{L}$, each of the following conditions is equivalent to the lsc property of $\mathcal{L}$ at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{D}}$ [53]: $\pi \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{D}}$, $c \in \operatorname{int} M$, and dual consistency under sufficiently small perturbations of c (among others). There, it is also shown that $\mathcal{L}$ is closed or not depending on the norm considered on the image space $\mathbb{R}^{(T)}$ (it is closed for L_1 but not for L_∞).

We finish this section considering the continuous case. Concerning the stability of $\mathcal{F}$, in the early 1980s it was proved that $\mathcal{F}$ is lsc at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ if and only if π satisfies the Slater condition and that $\mathcal{F}$ is usc at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ if and only if F is either the whole space $\mathbb{R}^n$ or a compact set [3, 41]. In [53] it is also shown that $\pi \in \left(\operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}\right) \cap \left(\operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{D}}\right)$ if and only if F^* and Λ^* are non-empty bounded sets if and only if $\pi \in \operatorname{int}\left(\Pi_{\mathrm{S}}^{\mathrm{P}} \cap \Pi_{\mathrm{S}}^{\mathrm{D}}\right)$. This result extends Robinson's theorem from LP to LSIP [93].

Most characterizations of the lsc property of $\mathcal{F}$ are valid for convex and some particular classes of non-convex systems posed in locally convex topological vector spaces (see [34]).

3.2 Stability of the Optimal Set

In [22] (see also [50]) it is proved that, if $\pi \in \Pi_{\mathrm{S}}^{\mathrm{P}}$, then the following statements hold:

- $\mathcal{F}^*$ is closed at $\pi \Leftrightarrow$ either $\mathcal{F}$ is lsc at π or $F = F^*$.
- $\mathcal{F}^*$ is lsc at $\pi \Leftrightarrow \mathcal{F}$ is lsc at π and $|F^*| = 1$ (uniqueness).
- If $\mathcal{F}^*$ is usc at π, then $\mathcal{F}^*$ is closed at π (and the converse is true if F^* is bounded).

Exploiting a suitable concept of extended active constraint, it has been shown in [54] that the strong uniqueness is a generic property on the intersection of $\Pi_{\mathrm{S}}^{\mathrm{P}}$ with the (open and closed) classes of those elements of Π which have bounded LHS function.

The continuous versions of the above characterizations of the semicontinuity and closedness of $\mathcal{F}^*$ appeared in [8, 41]. Concerning the mentioned generic result for general LSIP, it is an extension of a generic result in [96] for continuous LSIP (where any problem has bounded LHS function). Always in the continuous case, Todorov [97] proved that the majority (in the Baire sense) of the elements of $\Pi_{\mathrm{S}}^{\mathrm{P}}$ have an associated Lagrange function with a unique saddle point.

3.3 Stability of the Optimal Value and Well-Posedness

The following statements are proved in [22] (see also [50]):

- If F^* is a non-empty compact set, then ϑ^{P} is lsc at π. The converse statement holds if $\pi \in \Pi_{\mathrm{B}}^{\mathrm{P}}$.
- ϑ^{P} is usc at $\pi \Leftrightarrow \mathcal{F}$ is lsc at π.
- If π is Hwp, then the restriction of ϑ^{P} to $\Pi_{\mathrm{B}}^{\mathrm{P}}$is continuous.
- If F^* is bounded, then π is Hwp $\Leftrightarrow$ either $\mathcal{F}$ is lsc at π or $|F| = 1$.
- If F^* is unbounded and π is Hwp, then $\mathcal{F}$ is lsc at π.

A similar analysis has been carried out in [22] with other Hwp concepts.

In the particular case that $\pi \in \operatorname{int} \Pi_{\mathrm{S}}^{\mathrm{P}}$, Cánovas et al. [25] provide an expression for α (called *Lipschitz constant*), in terms of the data, such that

$$\left|\vartheta^{\mathrm{P}}(\pi_1) - \vartheta^{\mathrm{P}}(\pi_2)\right| \leq \alpha d_\infty(\pi_1, \pi_2)$$

for all π_1, π_2 in some neighborhood of π. The Lipschitz inequality

$$\left|\vartheta^{\mathrm{P}}(\pi_1) - \vartheta^{\mathrm{P}}(\pi)\right| \leq \alpha d_\infty(\pi_1, \pi)$$

for π_1 in some neighborhood of π provides bounds for the variation of ϑ^P in the proximity of π, i.e., this inequality can be seen as sensitivity analysis result.

3.4 Distance to Ill-Posedness

The following formulae [22, 24, 27] reduce the calculus of pseudodistances from π to the sets of ill-posed problems to the calculus of distances from the origin to a suitable set in certain Euclidean space:

- $d_\infty\left(\pi, \operatorname{bd}\Pi_{\mathrm{C}}^{\mathrm{P}}\right) = \left|\sup_{x\in\mathbb{R}^n}\inf_{t\in T}\frac{a_t'x - b_t}{\|(x,-1)\|_\infty^*}\right|$.
- If $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ and $H := \operatorname{conv} C + \operatorname{cone}\{(0_n, -1)\}$, then

$$d_\infty\left(\pi, \operatorname{bd}\Pi_{\mathrm{SI}}^{\mathrm{P}}\right) = d_\infty\left(0_{n+1}, \operatorname{bd} H\right).$$

- If $\pi \in (\operatorname{cl}\Pi_{\mathrm{S}}^{\mathrm{P}}) \cap (\operatorname{int}\Pi_{\mathrm{C}}^{\mathrm{P}})$ and $Z^- := \operatorname{conv}\{a_t, t \in T; -c\}$, then

$$d_\infty(\pi, \operatorname{bd}\Pi_{\mathrm{S}}^{\mathrm{P}}) = \min\{d_\infty(0_{n+1}, \operatorname{bd} H), d_\infty(0_n, \operatorname{bd} Z^-)\}.$$

- If $\pi \in (\operatorname{cl}\Pi_{\mathrm{S}}^{\mathrm{P}}) \cap (\operatorname{bd}\Pi_{\mathrm{C}}^{\mathrm{P}})$ and $Z^+ := \operatorname{conv}\{a_t, t \in T; c\}$, then

$$d_\infty(\pi, \operatorname{bd}\Pi_{\mathrm{S}}^{\mathrm{P}}) \geq \min\{d_\infty(0_{n+1}, \operatorname{bd} H), d_\infty(0_n, \operatorname{bd} Z^+)\}.$$

In [28] a subclass of $\left(\operatorname{bd}\Pi_{\mathrm{C}}^{\mathrm{P}}\right) \cap \left(\operatorname{bd}\Pi_{\mathrm{S}}^{\mathrm{P}}\right)$, called set of *totally ill-posed problems* (problems that are simultaneously ill posed in both feasibility and optimality senses), was identified. The totally ill-posed problems have been characterized, initially (in [26]) in terms of a set of parameters whose definition does not involve the data (so that it is hard to be checked) and recently (in [70]) in terms of the data.

3.5 Error Bounds

The *residual function* is $r : \mathbb{R}^n \times \Pi \mapsto \overline{\mathbb{R}}$ such that

$$r\left(x, \pi_1\right) := \sup_{t\in T}\left(b_t^1 - \left(a_t^1\right)' x\right)^+,$$

where $\alpha^+ := \max\{\alpha, 0\}$. Obviously, $x \in \mathcal{F}\left(\pi_1\right)$ if and only if $r\left(x, \pi_1\right) = 0$.

The scalar $0 \leq \beta < +\infty$ is a *global error bound* for $\pi_1 \in \Pi_{\mathrm{C}}^{\mathrm{P}}$ if

$$d\left(x, \mathcal{F}\left(\pi_1\right)\right) \leq \beta r\left(x, \pi_1\right) \text{ for all } x \in \mathbb{R}^n.$$

If there exists such a β, then *the condition number* of π_1 is

$$0 \leq \tau\left(\pi_1\right) := \sup_{x\in\mathbb{R}^n\setminus F}\frac{d\left(x, \mathcal{F}\left(\pi_1\right)\right)}{r\left(x, \pi_1\right)} < +\infty.$$

An estimation of $\tau(\pi)$ when F is bounded can be found in [29].

The following statements provide global error bounds for the parameters in some neighborhood of π, under the only assumption that C (the set of coefficient vectors) is bounded [71]:

♦ Assume that F is bounded and $\pi = (a, b, c) \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}$, and let ρ, x^0, and $\varepsilon > 0$ be such that $\|x\| \leq \rho$ for all $x \in F$ and $a_t' x^0 \geq b_t + \varepsilon$ for all $t \in T$. Let $0 \leq \gamma < 1$. Then, if $d(\pi_1, \pi) < \dfrac{\varepsilon\gamma}{2\rho\sqrt{n}}$, we have

$$\tau(\pi_1) \leq \frac{2\rho}{\varepsilon}\left[\frac{1+\gamma}{(1-\gamma)^2}\right].$$

♦ Assume that F is unbounded and $(a, 0, c) \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}$, and let u and $\eta > 0$ such that $a_t' u \geq \eta$ for all $t \in T$, $\|u\| = 1$. Let $0 < \delta < n^{-1/2}\eta$. Then, if $d(\pi_1, \pi) < \delta$, we have

$$\tau(\pi_1) \leq \left(\eta - \delta n^{1/2}\right)^{-1}.$$

Improved error bounds for arbitrary π have been given in [23].

3.6 Primal–Dual Stability

In the same way that $\operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}$ is interpreted as the set of primal stable consistent parameters (in the sense that sufficiently small perturbations provide primal consistent problems), the topological interior of the main subsets of Π can be seen as the sets of stable parameters in the corresponding sense. Some of these interiors have been characterized in the continuous case [57, 59], e.g., those corresponding to the primal partition $\left\{\Pi_{\mathrm{I}}^{\mathrm{P}}, \Pi_{\mathrm{B}}^{\mathrm{P}}, \Pi_{\mathrm{U}}^{\mathrm{P}}\right\}$, the dual partition $\left\{\Pi_{\mathrm{I}}^{\mathrm{D}}, \Pi_{\mathrm{B}}^{\mathrm{D}}, \Pi_{\mathrm{U}}^{\mathrm{D}}\right\}$, and their non-empty intersections (the so-called primal–dual partition):

♦ $\pi \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}} \Leftrightarrow \pi$ satisfies the Slater condition.

♦ $\pi \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{D}} \Leftrightarrow c \in \operatorname{int} M$.

♦ $\pi \in \operatorname{int} \Pi_{\mathrm{B}}^{\mathrm{P}} \Leftrightarrow \pi \in \operatorname{int}\left(\Pi_{\mathrm{B}}^{\mathrm{P}} \cap \Pi_{\mathrm{B}}^{\mathrm{D}}\right) \Leftrightarrow \pi \in \operatorname{int} \Pi_{\mathrm{B}}^{\mathrm{D}} \Leftrightarrow \pi$ satisfies the Slater condition and $c \in \operatorname{int} M$.

♦ $\pi \in \operatorname{int} \Pi_{\mathrm{I}}^{\mathrm{P}} \Leftrightarrow \pi \in \operatorname{int}\left(\Pi_{\mathrm{I}}^{\mathrm{P}} \cap \Pi_{\mathrm{U}}^{\mathrm{D}}\right) \Leftrightarrow \pi \in \operatorname{int} \Pi_{\mathrm{U}}^{\mathrm{D}} \Leftrightarrow (0_n, 1) \in \operatorname{int} K$.

♦ $\pi \in \operatorname{int} \Pi_{\mathrm{I}}^{\mathrm{D}} \Leftrightarrow \pi \in \operatorname{int}\left(\Pi_{\mathrm{U}}^{\mathrm{P}} \cap \Pi_{\mathrm{I}}^{\mathrm{D}}\right) \Leftrightarrow \pi \in \operatorname{int} \Pi_{\mathrm{U}}^{\mathrm{P}} \Leftrightarrow \exists y \in \mathbb{R}^n$ such that $c'y < 0$ and $a_t' y > 0$ for all $t \in T$.

Moreover,

$$\operatorname{int}\left(\Pi_{\mathrm{I}}^{\mathrm{P}} \cap \Pi_{\mathrm{I}}^{\mathrm{D}}\right) = \operatorname{int}\left(\Pi_{\mathrm{B}}^{\mathrm{P}} \cap \Pi_{\mathrm{I}}^{\mathrm{D}}\right) = \operatorname{int}\left(\Pi_{\mathrm{I}}^{\mathrm{P}} \cap \Pi_{\mathrm{B}}^{\mathrm{D}}\right) = \emptyset.$$

The above results have been extended [60] to the refined primal-dual partitions obtained by splitting the sets of parameters having bounded problems in the primal and the dual partitions, $\Pi_{\mathrm{B}}^{\mathrm{P}}$ and $\Pi_{\mathrm{B}}^{\mathrm{D}}$, into those which have compact optimal sets and those where this desirable property fails. The above characterizations of thetopological interiors of the main subsets of Π have been used

in order to prove that most parameters having either primal or dual bounded associated problems have primal and dual compact optimal sets [60]. This generic property does not hold in general LSIP despite almost all the characterizations of the topological interior of above subsets of Π being still valid in general LSIP [24, 26, 53].

4 Perturbing the Cost Vector and the RHS Function

In this section the parameter space is $\Pi = \mathbb{R}^T \times \mathbb{R}^n$ (general case) or the Banach space $\Pi = \mathcal{C}(T) \times \mathbb{R}^n$ (continuous case). This model is the most general one for which some sensitivity analysis with exact formulae can be performed at the moment of writing this chapter. Because the admissible perturbations of π are of the form $\pi_1 = (a, w, z)$, $w \in \mathbb{R}^T$, and $z \in \mathbb{R}^n$, we can identify π_1 with (z, w) (called *rim data* in the LP literature).

4.1 Stability Analysis

Because $\mathcal{F}$ is closed under perturbations of all the data, $\mathcal{F}$ is also closed under perturbations of some data.

According to [50], the characterizations of the lower semicontinuity of $\mathcal{F}$ at π in Section 3 remain valid for any model (in both general and continuous LSIP) allowing arbitrary perturbations of the RHS function.

The characterization of the upper semicontinuity of $\mathcal{F}$ at π also remains valid because the argument given for arbitrary perturbations of all the data in [18] only involves perturbations of the RHS function.

Concerning the stability of $\mathcal{F}^*$ and well-posedness, the proofs given in [22, 50] used perturbations of the LHS function, so that all can be asserted at present is that

- if F^* is a non-empty compact set, then ϑ^{P} is lsc at π;
- if $\mathcal{F}$ is lsc at π, then ϑ^{P} is usc at π; and
- if F^* is bounded and either $\mathcal{F}$ is lsc at π or $|F| = 1$, then π is Hwp.

Characterizing the stability properties of ϑ^{P} and $\mathcal{F}$ and the well-posedness in this model are open problems.

In the continuous case, it has been proved [12] that, given $(\pi, x) \in \operatorname{gph} \mathcal{F}^*$, $(\mathcal{F}^*)^{-1}$ is metrically regular at x if and only if $\mathcal{F}^*$ is single-valued in some neighborhood of π. In that case, $\mathcal{F}^*$ is also Lipschitz continuous on that neighborhood of π and $\operatorname{reg}(\mathcal{F}^*)^{-1}(x, \pi)$ can be calculated under a mild condition that always holds if $n \leq 3$. The latter results have been extended to CSIP under linear perturbations of the objective functions in [16], which give conditions for the metric regularity of $(\mathcal{F}^*)^{-1}$, and [13–15], which provide lower and upper bounds for $\operatorname{reg}(\mathcal{F}^*)^{-1}$ in terms of the problem's data; in LSIP the upper bound (or exact modulus) adopts a notably simplified expression.

4.2 Sensitivity Analysis

Consider the parametric problem

$$P(z,w): \min_{x\in\mathbb{R}^n} z'x \quad \text{s.t.} \quad a_t'x \geq w_t, \quad t\in T$$

and its corresponding dual

$$D(z,w): \max_{\lambda\in\mathbb{R}_+^{(T)}} \sum_{t\in T}\lambda_t w_t \quad \text{s.t.} \quad \sum_{t\in T}\lambda_t a_t = z.$$

Observe that $P(z,w)$ is continuous when P is continuous (recall that, in that case, we take $w\in\mathcal{C}(T)$).

In order to describe the behavior of the optimal value functions ϑ^{P} and ϑ^{D} we define a class of functions after giving a brief motivation. Let V be a linear space and let $\varphi: V^2\mapsto\mathbb{R}$ be a bilinear form on V. Let $X=\operatorname{conv}\{v_i, i\in I\}\subset V$ and let $q_{ij}:=\varphi(v_i,v_j)$, $(i,j)\in I^2$. Then any $v\in X$ can be expressed as

$$v=\sum_{i\in I}\mu_i v_i,\ \sum_{i\in I}\mu_i=1,\ \mu\in\mathbb{R}_+^{(I)}. \tag{9}$$

Then we have

$$\varphi(v,v)=\sum_{i,j\in I}\mu_i\mu_j q_{ij}. \tag{10}$$

Accordingly, given $q: X\mapsto\mathbb{R}$, where $X=\operatorname{conv}\{v_i, i\in I\}\subset V$, we say that q is quadratic on X if there exist real numbers q_{ij}, $i,j\in I$, such that $q(v)$ satisfies (10) for all $v\in X$ satisfying (9). The following result is proved in [56]:

♦ Let $\{(c^i,b^i), i\in I\}\subset\mathbb{R}^n\times\mathbb{R}^T$ be such that there exists a common optimal partition for the family of problems $\{P(c^i,b^i), i\in I\}$. Then $P(z,w)$ and $D(z,w)$ are solvable and $\vartheta^{\mathrm{P}}(z,w)=\vartheta^{\mathrm{D}}(z,w)$ on $\operatorname{conv}\{c^i, i\in I\}\times\operatorname{conv}\{b^i, i\in I\}$ and ϑ^{P} is quadratic on $\operatorname{conv}\{(c^i,b^i), i\in I\}$. Moreover, if $(c,b)\in\operatorname{conv}\{c^i, i\in I\}\times\operatorname{conv}\{b^i, i\in I\}$, then $\vartheta^{\mathrm{P}}(\cdot,b)$ and $\vartheta^{\mathrm{P}}(c,\cdot)$ are affine on $\operatorname{conv}\{c^i, i\in I\}$ and $\operatorname{conv}\{b^i, i\in I\}$, respectively.

Obviously, if $(c,b)\in\operatorname{int}\operatorname{conv}\{(c^i,b^i), i\in I\}$, then ϑ^{P} and ϑ^{D} coincide and are quadratic on a neighborhood of (c,b). In particular, if the problems $P(z,w)$ have a common optimal partition when (z,w) ranges on a certain neighborhood of (c,b), then we can assert that P has a strongly unique solution (and D has a unique solution).

5 Perturbing the RHS Function

We consider here that a and c are fixed whereas the RHS function b can be perturbed. For simplicity we use the variable w instead of b[1]. Thus, we write $\vartheta^{\mathrm{P}}(w)$ instead of $\vartheta^{\mathrm{P}}(\pi_1)$.

5.1 Stability Analysis

As in the previous section, $\mathcal{F}$ is closed and the characterizations of the lower and upper semicontinuities of $\mathcal{F}$ are the same as in Section 3 due to the same reasons. The condition $\pi \in \operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}$ means now that the consistency of the problem is preserved by sufficiently small perturbations of the RHS function. This property was called *regularity* by Robinson [93].

In the continuous case, the following formula for the regularity modulus of $\mathcal{F}^{-1}$ at $(\pi, x) \in \operatorname{gph} \mathcal{F}^{-1}$ has been obtained appealing to the distance to ill-posedness in feasibility sense [10]:

$$\operatorname{reg} \mathcal{F}^{-1}(x \mid \pi) = \sup \left\{ \left(\|u\|_{\infty}^{*} \right)^{-1} \mid (u, u'x) \in \operatorname{conv} C \right\}.$$

Concerning the stability of the primal optimal value function ϑ^{P}, according to [35] (which deals with convex infinite programs), if $\pi \in \Pi_{\mathrm{B}}^{\mathrm{P}}$ and K is closed, then ϑ^{P} is lsc at π, there exists an affine minorant of the directional derivative of ϑ^{P} at b (i.e., there exists $\lambda \in \mathbb{R}^{(T)}$ such that $\left(\vartheta^{\mathrm{P}}\right)'(b; w) \geq \lambda(w - b) \forall w \in \mathbb{R}^{T}$), and ϑ^{P} is subdifferentiable at b (i.e., $\partial \vartheta^{\mathrm{P}}(b) \neq \emptyset$). The first two properties are called *inf-stability* and *inf–dif-stability* in Laurent's sense, whereas the third one is equivalent to *calmness* in Clarke's sense [7].

A Lipschitz constant for ϑ^{D} at π in terms of the data has been given recently in [98, Theorem 1] under the assumption that $\pi \in \Pi_{\mathrm{C}}^{\mathrm{D}} \cap \left(\operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{P}}\right)$.

The open problems enumerated in Section 4.1 are also open problems for this model.

5.2 Sensitivity Analysis

Here we consider the parametric problems

$$\begin{aligned} P(w): \min_{x \in \mathbb{R}^n} \; & c'x \\ \text{s.t.} \quad & a_t'x \geq w_t, \quad t \in T \end{aligned}$$

and

$$\begin{aligned} D(w): \max_{\lambda \in \mathbb{R}_{+}^{(T)}} \; & \sum_{t \in T} \lambda_t w_t \\ \text{s.t.} \quad & \sum_{t \varepsilon T} \lambda_t a_t = c, \end{aligned}$$

with respective optimal values $\vartheta^{\mathrm{P}}(w)$ and $\vartheta^{\mathrm{D}}(w)$ (observe that $P(w)$ is continuous when P is continuous). Obviously, the optimal values of the nominal problem P and its dual D are $\vartheta^{\mathrm{P}}(b) = v(P)$ and $\vartheta^{\mathrm{D}}(b) = v(D)$, respectively.

The following sensitivity results have been shown:

♦ If ϑ^{P} is affine on a certain neighborhood of b, then D has at most one optimal solution and the converse is true under strong assumptions [43].

♦ ϑ^{P} is affine on a segment emanating from b in the direction of a bounded function $d \in \mathbb{R}^T \setminus \{0_T\}$ if P and D are solvable with the same optimal value, the problem

$$P_d : \min_{(x,y)\in\mathbb{R}^{n+1}} c'x + \vartheta^{\mathrm{P}}(b)\, y$$
$$\text{s.t.} \qquad a_t'x + b_t y \geq d_t, \quad t \in T$$

is also solvable and has zero duality gap, and P_d satisfies certain additional condition [43]. Once again, observe that P_d is continuous when P is continuous (provided $d \in \mathcal{C}(T)$).

♦ Let $\operatorname{conv}\{b^i i \in I\}$ be such that all the problems $P(b^i)$, $i \in I$, have a common optimal partition. Then ϑ^{P} and ϑ^{D} coincide and are affine on $\operatorname{conv}\{b^i, i \in I\}$ [56]. This result can be seen as the LSIP version of the optimal partition perspective of LP (see [64]).

Cánovas et al. [29] provide a lower bound for ϑ^{D} under the only assumption that ϑ^{P} is lsc.

6 Perturbing the Cost Vector

Now we consider a and b given (fixed) functions whereas c can be perturbed, i.e., the elements of Π are the triples $\pi_1 = (a, b, c^1)$, with $c^1 \in \mathbb{R}^n$. The theoretical advantage of this model is that the space of parameters is finite dimensional. For the sake of simplicity we write z instead of c^1.

6.1 Stability Analysis

The following result describes the local behavior of the optimal value functions ϑ^{P} and ϑ^{D} (which is related to the viability of the discretization approach in LSIP):

♦ ϑ^{D} is a proper concave function and ϑ^{P} is its closure, whose hypograph is $\operatorname{cl} K$ [48, 50], and its domain satisfies $\operatorname{rint} M \subset \operatorname{dom} \vartheta^{\mathrm{P}} \subset \operatorname{cl} M$. Thus, ϑ^{P} is positively homogeneous (i.e., $\vartheta^{\mathrm{P}}(\lambda z) = \lambda \vartheta^{\mathrm{P}}(z)$ for all $\lambda \geq 0$) and ϑ^{P} and ϑ^{D} are continuous on $\operatorname{rint} M$.

Concerning the stability, the following statements are true (the proofs in [8], on continuous LSIP, remain valid in general LSIP):

- $\mathcal{F}^*$ is closed.
- If $\mathcal{F}^*$ is lsc at c and F^* contains an exposed point of F, then $|F^*| = 1$.
- If F^* is bounded, then $\mathcal{F}^*$ is usc at c.

The characterization of the lsc and the characterization of the usc properties of $\mathcal{F}^*$ are open problems, whereas its metric regularity has been analyzed in [17], in the more general framework of convex semi-infinite programming. The stability of the dual problem has not been analyzed, except the lsc property of $\mathcal{L}$ at $\pi \in \Pi_{\mathrm{C}}^{\mathrm{D}}$, which is equivalent to $c \in \operatorname{int} M$.

In the particular case that $\pi \in \Pi_{\mathrm{C}}^{\mathrm{P}} \cap \left(\operatorname{int} \Pi_{\mathrm{C}}^{\mathrm{D}}\right)$, [98, Theorem 2] provides a Lipschitz constant for ϑ^{P} at π in terms of the data.

6.2 Sensitivity Analysis

The perturbed problems of P and D to be considered in this section are

$$P(z): \begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & z'x \\ \text{s.t.} & a_t'x \geq b_t, \quad t \in T \end{array}$$

and

$$D(z): \begin{array}{ll} \max\limits_{\lambda \in \mathbb{R}_+^{(T)}} & \sum\limits_{t \in T} \lambda_t b_t \\ \text{s.t.} & \sum\limits_{t \varepsilon T} \lambda_t a_t = z, \end{array}$$

with optimal values $\vartheta^{\mathrm{P}}(z)$ and $\vartheta^{\mathrm{D}}(z)$, respectively (observe that $P(z)$ is continuous when P is continuous). With this notation, the effective domain of ϑ^{D} is the first moment cone, M, and the optimal values of the nominal problem P and its dual D are $\vartheta^{\mathrm{P}}(c)$ and $\vartheta^{\mathrm{D}}(c)$, respectively.

The next three results can be seen as the extension to LSIP of classical results on sensitivity analysis in LP.

- If $c \in \operatorname{rint} M$, then the subdifferential of ϑ^{P} at c is $\partial \vartheta^{\mathrm{P}}(c) = F^* \neq \emptyset$. In particular, if $c \in \operatorname{int} M$, i.e., F^* is compact, the directional derivative of ϑ^{P} at c in the direction of $d \in \mathbb{R}^n \backslash \{0_n\}$ is

$$\left(\vartheta^{\mathrm{P}}\right)'(c; d) = \max_{x \in F^*} d'x,$$

and ϑ^{P} turns out to be differentiable at c if and only if $|F^*| = 1$ (i.e., P has a unique optimal solution). Then, $\nabla \vartheta^{\mathrm{P}}(c) = x^*$ if $F^* = \{x^*\}$ [50].
- ϑ^{P} is linear in a neighborhood of c if and only if P has a strongly unique solution. In such a case, if $F^* = \{x^*\}$, then $\vartheta^{\mathrm{P}}(z) = (x^*)'z$ for z ranging on some open convex cone containing c [43].
- Let P and $d \in \mathbb{R}^n$ be such that P and D are solvable, $\vartheta^{\mathrm{D}}(c) = \vartheta^{\mathrm{P}}(c)$ and

$$D(d): \begin{array}{ll} \max\limits_{\lambda \in \mathbb{R}_+^{(T)}} & \sum\limits_{t \in T} \lambda_t b_t + \mu \vartheta^{\mathrm{P}}(c) \\ \text{s.t.} & \sum\limits_{t \varepsilon T} \lambda_t a_t + \mu c = d \end{array}$$

is also solvable and has zero duality gap. Then there exists $\varepsilon > 0$ such that

$$\vartheta^{\mathrm{P}}(c + \rho d) = \vartheta^{\mathrm{P}}(c) + \rho \min\{d'x \mid x \in F^*\} \quad \text{if } 0 \leq \rho < \varepsilon.$$

Consequently, ϑ^{P} is linear on cone $[c, c + \varepsilon d]$ ([43], extending Gauvin's formula in [42] to LSIP).

♦ Let $\{c^i, i \in I\} \subset \mathbb{R}^n$ be such that there exists a common optimal partition for the family of problems $\{P(c^i), i \in I\}$. Then ϑ^{P} and ϑ^{D} coincide and are affine on conv $\{c^i, i \in I\}$ [56].

7 Conclusions

We have shown in Section 1 that post-optimal analysis is a sensible way to deal with LSIP problems with uncertain data (the other one is robust optimization, but only in the case that the unique uncertain data are the cost coefficients).

The following concerns the post-optimal models surveyed in Sections 3–6:

♦ The stability analysis of $\mathcal{F}$ and ϑ^{P} is almost complete in all cases (except for the Aubin continuity) whereas the analysis corresponding to $\mathcal{F}^*$ is only complete for perturbations of all the data.

♦ The Hadamard well-posedness has not been characterized (although some necessary and some sufficient conditions are known).

♦ The distance to ill-posedness is only computable for perturbations of all the data (the formulae in Section 3.3 only give lower bounds in the remaining models).

♦ The condition number of an arbitrary LSIP problem cannot be computed (although upper bounds can be obtained).

♦ No generic result is available for general LSIP (the existing literature requires the LHS function to be bounded).

♦ No sensitivity analysis with exact formulae can be carried out when perturbations of all the data are admissible although some quantitative information is available, e.g., Lipschitz constants in terms of the data for certain types of stable problems.

♦ Almost nothing is known about the dual problem (only the stability of the feasible set mapping $\mathcal{L}$ has been studied in detail up to now), although this type of problem seldom arises in the real applications of LSIP.

For the sake of simplicity, we have assumed in this chapter (as in most of the published works) that the perturbable data are non-empty subsets of $\{a, b, c\}$ (e.g., that we can perturb the whole cost vector c but not just an individual coefficient c_i, $i = 1, \ldots, n$, all the constraints but not a part of them). There is an active research in progress about models containing linked inequalities to be preserved by any admissible perturbation [1, 2, 11], where each equation can be interpreted as two zero-sum inequalities), models

including imperturbable constraints (e.g., the physical constraints $x_i \geq 0$, $i = 1, \ldots, n$), or both [1, 2].

We have also precluded from this survey models involving a parametrization mapping describing either the coefficients, the index set [81], or both [19, 20]. Three of these alternative approaches are compared in [21]: free perturbation of all the data, perturbations of the data depending on a given parameter (in both cases maintaining the structure of the problem), and perturbations preserving the space of primal variables and the linearity of the system. Under suitable smoothness assumptions on the parametrization mapping it is possible to guarantee strong topological properties of the feasible set and the optimal set mappings or to describe the geometry of the trajectory described by the optimal solution, if it is unique (see, e.g., [77, 78]. For more information on perturbation analysis in more general contexts, the reader is referred to [7, 79] and references therein.

All the previous models assume that the perturbations preserve the structure of the nominal problem, i.e., that the perturbed problems are linear programs with the same number of variables and constraints as the nominal one. Even more, in the papers dealing with continuous problems, it is also assumed that the coefficients of the constraints of the perturbed problems are continuous functions of the parameter. Other approaches are possible but very unusual in the literature. For instance, because each triple $\pi = (a, b, c)$ could be identified with a couple $(X, c) \in 2^{\mathbb{R}^{n+1}} \times \mathbb{R}^n$, where X is some set associated with the constraints, it is possible to consider a subset of $2^{\mathbb{R}^{n+1}} \times \mathbb{R}^n$ as space of parameters equipped with the product topology of a suitable one on the family of sets representing the systems by the usual topology on $\mathbb{R}^n$. The Hausdorff topology on the space of compact sets in $\mathbb{R}^{n+1}$ is a good choice if these sets are compact (as $\operatorname{cl} K \cap \operatorname{cl} B\,(0_{n+1}; 1)$ in [90], paper devoted to the stability of the feasible set mapping), whereas hypertopologies could be preferable in the case that the sets are closed cones (as $\operatorname{cl} K$). Nevertheless, almost all the specialists prefer to use the topology of the uniform convergence on the parameter space Π introduced in Section 2 for two reasons: first, because this topology makes sense in practice and second because the representation of π in $2^{\mathbb{R}^{n+1}} \times \mathbb{R}^n$ affects the dual problem, i.e., the alternative approach is only suitable for the stability analysis of the primal problem.

In conclusion, post-optimal analysis in LSIP is an active research field which includes different perturbation models covering a variety of structures arising in practice and possible sources of uncertainty.

Acknowledgment The author wishes to thank M.J. Cánovas, M.D. Fajardo, and J. Parra for their valuable comments and suggestions.

References

1. Amaya, J., Bosch, P., Goberna, M.A.: Stability of the feasible set mapping of linear systems with an exact constraint set. Set-Valued Anal. 16, 621–635 (2008)

2. Amaya, J., Goberna, M.A.: Stability of the feasible set of linear systems with an exact constraints set. Math. Methods Oper. Res. 63, 107–121 (2006)
3. Bennett, K.P., Parrado-Hernández, E.: The interplay of optimization and machine learning research. J. Mach. Learn. Res. 7, 1265–1281 (2006)
4. Ben-Tal, A., Goryashko, A., Guslitzer, E., Nemirovski, A.: Adjustable robust solutions of uncertain linear programs. Math. Program. 99A, 351–376 (2004)
5. Ben-Tal, A., Nemirovski, A.: Robust optimization – methodology and applications. Math. Program. 92B, 453–480 (2002)
6. Betró, B.: An accelerated central cutting plane algorithm for linear semi-infinite programming. Math. Program. 101A, 479–495 (2004)
7. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems, Springer, NY (2000)
8. Brosowski, B.: Parametric Semi-Infinite Optimization, Verlag Peter Lang, Frankfurt am Main - Bern (1982)
9. Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM J. Optim. 19, 1211–1230 (2008)
10. Cánovas, M.J., Dontchev, A.L., López, M.A., Parra, J.: Metric regularity of semi-infinite constraint systems. Math. Program. 104B, 329–346 (2005)
11. Cánovas, M.J., Gómez-Senent, F.J., Parra, J.: Stability of systems of linear equations and inequalities: distance to ill-posedness and metric regularity. Optimization 56, 1–24 (2007)
12. Cánovas, M.J., Gómez-Senent, F.J., Parra, J.: On the Lipschitz modulus of the argmin mapping in linear semi-infinite optimization. Set-Valued Anal. 16, 511–538 (2008)
13. Cánovas, M.J., Hantoute, A., López, M.A., Parra, J.: Lipschitz behavior of convex semi-infinite optimization problems: a variational approach. J. Global Optim. 41, 1–13 (2008)
14. Cánovas, M.J., Hantoute, A., López, M.A., Parra, J.: Lipschitz modulus of the optimal set mapping in convex optimization via minimal subproblems. Paci. J. Optim. 4, 411–422 (2008)
15. Cánovas, M.J., Hantoute, A., López, M.A., Parra, J.: Lipschitz modulus in convex semi-infinite optimization via d.c. functions. ESAIM Control, Optim. Cal. Var. (to appear)
16. Cánovas, M.J., Hantoute, A., López, M.A., Parra, J.: Stability of indices in KKT conditions and metric regularity in convex semi-infinite optimization. J. Optim. Theory Appl. 139, 485–500 (2008)
17. Cánovas, M.J., Klatte, D., López, M.A., Parra, J.: Metric regularity in convex semi-infinite optimization under canonical perturbations, SIAM J. Optim. 18, 717–732 (2007)
18. Cánovas, M.J., López, M.A., Parra, J.: Upper semicontinuity of the feasible set mapping for linear inequality systems. Set-Valued Anal. 10, 361–378 (2002)
19. Cánovas, M.J., López, M.A., Parra, J.: Stability of linear inequality systems in a parametric setting. J. Optim. Theory Appl. 125, 275–297 (2005)
20. Cánovas, M.J., López, M.A., Parra, J.: On the continuity of the optimal value in parametric linear optimization. Stable discretization of the Lagrangian dual of nonlinear problems. Set-Valued Anal. 13, 69–84 (2005)
21. Cánovas, M.J., López, M.A., Parra, J.: On the equivalence of parametric contexts for linear inequality systems. J. Comput. Appl. Math. 217, 448–456 (2008)

22. Cánovas, M.J., López, M.A., Parra, J., Todorov, M.I.: Solving strategies and well-posedness in linear semi-infinite programming. Ann. Oper. Res. 101, 171–190 (2001)
23. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Distance to ill-posedness and the consistency value of linear semi-infinite inequality systems. Math. Program. 103A, 95–126 (2005)
24. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Distance to solvability/unsolvability in linear optimization. SIAM J. Optim. 16, 629–649 (2006)
25. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Lipschitz continuity of the optimal value via bounds on the optimal set in linear semi-infinite optimization. Math. Oper. Res. 31, 478–489 (2006)
26. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Ill-posedness with respect to the solvability in linear optimization. Linear Algebra Appl. 416, 520–540 (2006)
27. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Distance to ill-posedness in linear optimization via the Fenchel-Legendre conjugate. J. Optim. Theory Appl. 130, 173–183 (2006)
28. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Sufficient conditions for total ill-posedness in linear semi-infinite optimization. Eur. J. Oper. Res. 181, 1126–1136 (2007)
29. Cánovas, M.J., López, M.A., Parra, J., Toledo, F.J.: Error bounds for the inverse feasible set mapping in linear semi-infinite optimization via a sensitivity dual approach. Optimization 56, 547–563 (2007)
30. Charnes, A., Cooper, W.W. , Kortanek, K.O.: Duality, Haar programs, and finite sequence spaces. Proc. Natl. Acad. Sci. USA 48, 783–786 (1962)
31. Charnes, A., Cooper, W.W. , Kortanek, K.O.: Duality in semi-infinite programs and some works of Haar and Carathéodory. Manage. Sci. 9, 209–228 (1963)
32. Charnes, A., Cooper, W.W. , Kortanek, K.O.: On representations of semi-infinite programs which have no duality gaps. Manage. Sci. 12, 113–121(1965)
33. Coelho, C.J., Galvao, R.K.H. , de Araujo, M.C.U., Pimentel, M.F., da Silva, E.C.: A linear semi-infinite programming strategy for constructing optimal wavelet transforms in multivariate calibration problems, J. Chem. Inf. Comput. Sci. 43, 928–933 (2003)
34. Dinh, N., Goberna, M.A., López, M.A.: On the stability of the feasible set in optimization problems. Technical. Report, Department of Statistics and Operational Research, University of Alicante, Spain (2009)
35. Dinh, N., Goberna, M.A., López, M.A., Son, T.Q.: New Farkas-type constraint qualifications in convex infinite programming. ESAIM Control, Optim. Calc. Var. 13, 580–597 (2007)
36. Dolgin, Y., Zeheb, E.: Model reduction of uncertain FIR discrete-time systems. IEEE Trans. Circuits Syst. 51, 406–411 (2004)
37. Dolgin, Y., Zeheb, E.: Model reduction of uncertain systems retaining the uncertainty structure. Syst. Control Lett. 54, 771–779 (2005)
38. Epelman, M., Freund, R.M.: A new condition measure, preconditioners, and relations between different measures of conditioning for conic linear systems. SIAM J. Optim. 12, 627–655 (2002)
39. Fang, S.C., Wu, S.Y., Sun, J.: An analytic center based cutting plane method for solving semi-infinite variational inequality problems. J. Global Optim. 24, 141–152 (2004)

40. Feyzioglu, O., Altinel, I.K., Ozekici, S.: The design of optimum component test plans for system reliability. Comput. Stat. Data Anal. 50, 3099–3112 (2006)
41. Fischer, T.: Contributions to semi-infinite linear optimization. Meth. Verf. Math. Phys. 27, 175–199 (1983)
42. Gauvin, J.: Formulae for the Sensitivity Analysis of Linear Programming Problems. In: M. Lassonde (Ed.), Approximation, Optimization and Mathematical Economics (pp.117–120). Physica Verlag, Berlin (2001)
43. Goberna, M.A., Gómez, S., Guerra, F., Todorov, M.I.: Sensitivity analysis in linear semi-infinite programming: perturbing cost and right-hand-side coefficients. Eur. J. Oper. Res. 181, 1069–1085 (2007)
44. Goberna, M.A., Jeyakumar, V., Dinh, N.: Dual characterizations of set containments with strict inequalities. J. Global Optim. 34, 33–54 (2006)
45. Goberna, M.A., Larriqueta, M., Vera de Serio, V.N.: On the stability of the boundary of the feasible set in linear optimization. Set-Valued Anal. 11, 203–223 (2003)
46. Goberna, M.A., Larriqueta, M., Vera de Serio, V.N.: On the stability of the extreme point set in linear optimization. SIAM J. Optim. 15, 1155–1169 (2005)
47. Goberna, M.A., Larriqueta, M., Vera de Serio, V.N.: Stability of the intersection of solution sets of semi-infinite systems, J. Comput. Appl. Math. 217, 420–431 (2008)
48. Goberna, M.A., López, M.A.: Optimal value function in semi-infinite programming. J. Optim. Theory Appl. 59, 261–280 (1988)
49. Goberna, M.A., López, M.A.: Topological stability of linear semi-infinite inequality systems. J. Optim. Theory Appl. 89, 227–236 (1996)
50. Goberna, M.A., López, M.A.: Linear Semi-Infinite Optimization, Wiley, Chichester, England (1998)
51. Goberna, M.A., López, M.A., Todorov, M.I.: Stability theory for linear inequality systems. SIAM J. Matrix Anal. Appl. 17, 730–743 (1996)
52. Goberna, M.A., López, M.A., Todorov, M.I.: Stability theory for linear inequality systems. II: upper semicontinuity of the solution set mapping. SIAM J. Optim. 7, 1138–1151 (1997)
53. Goberna, M.A., López, M.A., Todorov, M.I.: On the stability of the feasible set in linear optimization. Set-Valued Anal. 9, 75–99 (2001)
54. Goberna, M.A., López, M.A., Todorov, M.I.: A generic result in linear semi-infinite optimization. Appl. Math. Optim. 48, 181–19 (2003)
55. Goberna, M.A., López, M.A., Todorov, M.I.: On the stability of closed-convex-valued mappings and the associated boundaries. J. Math. Anal. Appl. 306, 502–515 (2005)
56. Goberna, M.A., Terlaky, T., Todorov, M.I.: Sensitivity analysis in linear semi-infinite programming via partitions. Math. Oper. Res. 35, 14–25 (2010)
57. Goberna, M.A., Todorov, M.I.: Primal, dual and primal-dual partitions in continuous linear optimization. Optimization 56, 617–628 (2007)
58. Goberna, M.A., Todorov, M.I.: Ill-posedness in continuous linear optimization via partitions of the space of parameters. Com. Ren. Acad. Bulgare Sci. 60, 357–364 (2007)
59. Goberna, M.A., Todorov, M.I.: Primal-dual stability in continuous linear optimization. Math. Program. 116B, 129–146 (2009)
60. Goberna, M.A., Todorov, M.I.: Generic primal-dual solvability in continuous linear semi-infinite programming. Optimization 57 1–10 (2008)

61. Goberna, M.A., Todorov, M.I., Vera de Serio, V.N.: On stable uniqueness in linear semi-infinite programming, Technical. Report, Department of Statistics and operational Research, University of Alicante, Spain (2009)
62. Goberna, M.A., Vera de Serio, V.N.: On the stable containment of two sets. J. Global Optim. 41, 613–624 (2008)
63. Gómez, J.A., Gómez, W.: Cutting plane algorithms for robust conic convex optimization problems. Optim. Methods Softw. 21, 779–803 (2006)
64. Greenberg, H.J.: The use of the optimal partition in a linear programming solution for postoptimal analysis. Oper. Res. Lett. 15, 179–185 (1994)
65. Gustafson, S.A.: On the computational solution of a class of generalized moment problems. SIAM J. Numer. Anal. 7, 343–357 (1970)
66. Gustafson, S.A.: On semi-infinite programming in numerical analysis. Lect. Notes Control Inf. Sci. 15, 137–153 (1979)
67. Gustafson, S.A.: A three-phase algorithm for semi-infinite programs, semi-infinite programming and applications. Lect. Notes Econ. Math. Syst. 215, 136–157 (1983)
68. Gustafson, S.A., Kortanek, K.O.: Numerical treatment of a class of semi-infinite programming problems. Nav. Res. Logist. Quart. 20, 477–504 (1973)
69. Hansen, E., Walster, G.W.: Global Optimization Using Interval Analysis (2nd edn), Marcel Dekker, NY (2004)
70. Hantoute, A., López, M.A.: Characterization of total ill-posedness in linear semi-infinite optimization. J. Comput. Appl. Math. 217, 350–364 (2008)
71. Hu, H.: Perturbation analysis of global error bounds for systems of linear inequalities. Math. Program. 88B, 277–284 (2000)
72. Huang, G.H., He, L., Zeng, G.M., Lu, H.W.: Identification of optimal urban solid waste flow schemes under impacts of energy prices. Environ. Eng. Sci. 25, 685–696 (2008)
73. Ito, R., Hirabayashi, R.: Design of FIR filter with discrete coefficients based on semi-infinite linear programming method. Pac. J. Optim. 3, 73–86 (2007)
74. Jeyakumar, V., Ormerod, J., Womersly, R.S.: Knowledge-based semi-definite linear programming classifiers. Optim. Methods Softw. 21, 693–706 (2006)
75. Jia, D., Krogh, B.H., Stursberg, O.: LMI approach to robust model predictive control. J. Optim. Theory Appl. 127, 347–365 (2005)
76. John, F.: Extremum problems with inequalities as subsidiary conditions. In Studies and Essays Presented to R. Courant on his 60th Birthday (pp.187–204), Interscience Publishers, NY (1948)
77. Jongen, H.Th., Rückmann, J.J.: On stability and deformation in semi-infinite optimization. In: R. Reemtsen, J.J. Rückmann (Eds.), Semi-infinite Programming, Kluwer, Boston, 29–67 (1998)
78. Jongen, H.Th., Twilt, F., Weber, G.H.: Semi-infinite optimization: structure and stability of the feasible set. J. Optim. Theory Appl. 72, 529–552 (1992)
79. Klatte, D., Henrion, R.: Regularity and stability in nonlinear semi-infinite optimization. In: R. Reemtsen, J.J. Rückmann (Eds.), Semi-infinite Programming, Kluwer, Boston, 69–102 (1998)
80. Kortanek, K.O.: On the 1962–1972 decade of semi-infinite programming: a subjective view. In: M.A. Goberna, M.A. López (Eds.), Semi-Infinite Programming: Recent Advances. Kluwer, Dordrecht, 3–34 (2001)

81. Kostyukova, O.I.: An algorithm constructing solutions for a family of linear semi-infinite problems. J. Optim. Theory Appl. 110, 585–609 (2001)
82. Krishnan, K., Mitchel, J.E.: Semi-infinite linear programming approaches to semidefinite programming problems. In: P. Pardalos, (Ed.) Novel Approaches to Hard Discrete Optimization (pp.121–140), American Mathematical Society, Providence, RI (2003)
83. Krishnan, K., Mitchel, J.E.: A unifying framework for several cutting plane methods for semidefinite programming. Optim. Methods Softw. 21, 57–74 (2006)
84. Krishnan, K., Mitchel, J.E.: A semidefinite programming based polyhedral cut and price approach for the maxcut problem. Comput. Optim. Appl. 33, 51–71 (2006)
85. León, T., Vercher, E.: Solving a class of fuzzy linear programs by using semi-infinite programming techniques. Fuzzy Sets and Systems, 146, 235–252 (2004)
86. Mangasarian, O.L.: Knowledge-based linear programming. SIAM J. Optim. 12, 375–382 (2004)
87. Maruhn, J.H.: Duality in static hedging of barrier options. Optimization 58, 319–333 (2009)
88. Maruhn, J.H., Sachs, E.W.: Robust static super-replication of barrier options in the black Scholes Model. In: A.J. Kurdila, P.M. Pardalos, M. Zabarankin, (Ed.) Robust Optimization-Directed Design (pp.127–143). Springer, NY (2005)
89. Meer, K.: On a refined analysis of some problems in interval arithmetic using real number complexity theory. Reliab. Comput. 10, 209–225 (2004)
90. Mira, J.A., Mora, G.: Stability of linear inequality systems measured by the Hausdorff metric. Set-Valued Anal. 8, 253–266 (2000)
91. Ramík, J., Inuiguchi, M. (Ed.): Fuzzy Mathematical Programming, Elsevier, Amsterdam (2000)
92. Renegar, J.: Linear programming, complexity theory and elementary functional analysis. Math. Program. 70A, 279–351 (1995)
93. Robinson, S.M.: Stability theory for systems of inequalities. Part I: Linear systems. SIAM J. Numer. Anal. 12, 754–769 (1975)
94. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming. Modeling and Theory, MPS/SIAM Series on Optimization 9, Philadelphia, PA (2009)
95. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. J. Mach. Learn. Res. 7, 1531–1565 (2006)
96. Todorov, M.I.: Generic existence and uniqueness of the solution set to linear semi-infinite optimization problems. Numer. Funct. Anal. Optim. 8, 541–556 (1985–1986)
97. Todorov, M.I.: Uniqueness of the saddle points for most of the Lagrange functions of the linear semi-infinite optimization. Numer. Funct. Anal. Optim. 10, 367–382 (1989)
98. Toledo, F.J., Some results on Lipschitz properties of the optimal values in semi-infinite programming. Optim. Meth. Softw. 23, 811–820 (2008)
99. Tuan, H.D., Nam, L.H., Tuy, H., Nguyen, T.Q.: Multicriterion optimized QMF bank design. IEEE Trans. Signal Proces. 51, 2582–2591 (2003)
100. Vaz, A.I.F., Ferreira, E.C.: Air pollution control with semi-infinite programming. Appl. Math. Modelling 33, 1957–1969 (2009)

101. Venkataramani, R., Bresler, Y.: Filter design for MIMO sampling and reconstruction. IEEE Trans. Signal Proces. 51, 3164–3176 (2003)
102. Yu, Y.J., Zhao, G., Teo, K.L., Lim, Y.C.: Optimization of extrapolated impulse response filters using semi-infinite programming, In Control, Communications and Signal Processing, First International Symposium, 397–400 (2004)
103. Zhu, L.M., Ding, Y., Ding, H.: Algorithm for spatial straightness evaluation using theories of linear complex Chebyshev approximation and semi-infinite linear programming. J. Manufact. Sci. Eng.- Trans. of the ASME 128, 167–174 (2006)

On Equilibrium Problems

Gábor Kassay

Faculty of Mathematics and Computer Science, Babes-Bolyai University, Cluj, Romania
kassay@math.ubbcluj.ro

Summary. In this chapter we give an overview of the theory of scalar equilibrium problems. To emphasize the importance of this problem in nonlinear analysis and in several applied fields we first present its most important particular cases as optimization, Kirszbraun's problem, saddlepoint (minimax) problems, and variational inequalities. Then, some classical and new results together with their proofs concerning existence of solutions of equilibrium problems are exposed. The existence of approximate solutions via Ekeland's variational principle – extended to equilibrium problems – is treated within the last part of the chapter.

Key words: equilibrium problem, saddlepoint, variational inequality, intersection theorems, Ekeland's variational principle, approximate solutions

1 Introduction

One of the most important problems in nonlinear analysis is the so-called *equilibrium problem*, which can be formulated as follows. Let A and B be two nonempty sets and $f : A \times B \to \mathbb{R}$ a given function. The problem consists in finding an element $a \in A$ such that

$$f(a, b) \geq 0 \quad \forall b \in B. \tag{EP}$$

(EP) has been extensively studied in recent years (e.g. [6–10, 17–19, 22] and the references therein). One of the reasons is that it has among its particular cases, optimization problems, saddlepoint (minimax) problems, variational inequalities (monotone or otherwise), Nash equilibrium problems, and other problems of interest in many applications (see [10] for a survey).

As far as we know the term "equilibrium problem" was attributed in [10], but the problem itself has been investigated more than 20 years before in

Work supported by the grant PNII, ID 523/2007

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_3,

a paper of Ky Fan [15] in connection with the so-called intersection theorems (i.e., results stating the nonemptiness of a certain family of sets). Ky Fan considered (EP) in the special case $A = B$ a compact convex subset of a Hausdorff topological vector space and termed it "minimax inequality." Within short time (in the same year) Brézis, Nirenberg, and Stampacchia [11] improved Ky Fan's result, extending it to a not necessarily compact set, but assuming instead a so-called coercivity condition, which is automatically satisfied when the set is compact.

Recent result on (EP) emphasizing existence of solutions can be found in [6–8, 28], and many other papers. New necessary (and in some cases also sufficient) conditions for existence of solutions in infinite dimensional spaces were proposed in [18], and later on simplified and further analyzed in [17].

Looking on the proofs given for existence results, one may detect two fundamental methods: fixed point methods (intersection theorems mostly based on Brouwer's fixed point theorem) and separation methods (Hahn–Banach type theorems). It is an old conjecture whether Brouwer's fixed point theorem can be proved using (only) separation results.

The aim of this chapter is to provide an overlook on (EP) by emphasizing its most important particular cases, to expose some classical and recent existence results of it, and to deal with approximate solutions, which, in case the exact solution does not exist, may have an important role.

The chapter is divided into four sections (including Introduction). In Section 2, the most important particular cases of (EP) such as the minimum problem, Kirszbraun's problem, saddlepoint problem (in connection with game theory, duality in optimization, etc.), and variational inequalities are presented. The next section is devoted to several existence results on (EP). First we focus on results which use fixed point tools and show that these results form an equivalent chain which includes Brouwer's and Schauder's fixed point theorems, Knaster–Kuratowski–Mazurkiewitz and Ky Fan's intersection theorems, Ky Fan's minimax inequality theorem. Then we expose some recent results on (EP) using separation tools. Finally, in Section 4 (EP) and its more general case, the system of equilibrium problems (abbreviated (SEP)), are discussed in connection with the famous Ekeland's variational principle. The latter has been established for optimization problems and guarantees the existence of the so-called approximate minimum points. Based on recent results of the author, the extensions of Ekeland's variational principle for (EP) and (SEP) are given under suitable conditions. These results are useful tools in obtaining new existence results for (EP) and (SEP) without any convexity assumptions on the sets and functions involved.

2 The Equilibrium Problem and Its Important Particular Cases

To underline the importance of (EP) we present in this section some of its various particular cases which have been extensively studied in the literature.

The most of them are important models of real-life problems originated from mechanics, economy, biology, etc.

2.1 The Minimum Problem

For $A = B$ and $F : A \to \mathbb{R}$, let $f(a, b) := F(b) - F(a)$. Then each solution of (EP) is a minimum point of F and vice versa.

2.2 The Kirszbraun's Problem

Let m and n be two positive integers and consider two systems of closed balls in $\mathbb{R}^n$: (B_i) and (B_i'), $i \in \{1, 2, \ldots, m\}$. Denote by $r(B_i)$ and $d(B_i, B_j)$ the radius of B_i and the distance between the centers of B_i and B_j, respectively. The following result is known in the literature as *Kirszbraun's theorem* (see [24]).

Theorem 1. *Suppose that*

(a) $\cap_{i=1}^m B_i \neq \emptyset$;
(b) $r(B_i) = r(B_i')$, for all $i \in \{1, 2, \ldots, m\}$;
(c) $d(B_i', B_j') \leq d(B_i, B_j)$, for all $i, j \in \{1, 2, \ldots, m\}$.

Then $\cap_{i=1}^m B_i' \neq \emptyset$.

To relate this result to (EP), let $A := \mathbb{R}^n$, $B := \{(x_i, y_i) | \, i \in \{1, 2, \ldots, m\}\} \subseteq \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\|y_i - y_j\| \leq \|x_i - x_j\| \quad \forall i, j \in \{1, 2, \ldots, m\}. \tag{1}$$

Choose an arbitrary element $x \in \mathbb{R}^n$ and put

$$f(y, b_i) := \|x - x_i\|^2 - \|y - y_i\|^2 \tag{2}$$

for each $y \in \mathbb{R}^n$ and $b_i = (x_i, y_i) \in B$. Then $y \in \mathbb{R}^n$ is a solution of (EP) if and only if

$$\|y - y_i\| \leq \|x - x_i\| \quad \forall i \in \{1, 2, \ldots, m\}. \tag{3}$$

It is easy to see by Theorem 1 that the equilibrium problem given by the function f defined in (2) has a solution. Indeed, let $x \in \mathbb{R}^n$ be fixed and put $r_i := \|x - x_i\|$ for $i := 1, 2, \ldots, m$. Take B_i the closed ball centered at x_i with radius r_i and B_i' the closed ball centered at y_i with radius r_i. Obviously, by (1), the assumptions of Theorem 1 are satisfied, hence there exists an element $y \in \mathbb{R}^n$ which satisfies (3).

Observe that, by compactness (i.e., the closed balls in $\mathbb{R}^n$ are compact sets), Theorem 1 of Kirszbraun remains valid for an arbitrary family of balls. More precisely, instead of the finite set $\{1, 2, \ldots, m\}$, one can take an arbitrary set I of indices. Using this observation, it is easy to derive the following result concerning the extensibility of an arbitrary nonexpansive function to the whole space. Let $D \subseteq \mathbb{R}^n$, $D \neq \mathbb{R}^n$, and $f : D \to \mathbb{R}^n$ a given *nonexpansive* function, i.e.,

$$\|f(x) - f(y)\| \leq \|x - y\| \quad \forall x, y \in D.$$

Then there exists a nonexpansive function $\bar{f} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\bar{f}(x) = f(x)$, for each $x \in D$. Indeed, let $z \in \mathbb{R}^n \setminus D$ and take for each $x \in D$ the number $r_x := \|z - x\|$. Let B_x be the closed ball centered at x with radius r_x and let B'_x be the closed ball centered at $f(x)$ with radius r_x. Then we obtain that the set $\cap_{x \in D} B'_x$ is nonempty. Now for $\bar{f}(z) \in \cap_{x \in D} B'_x$, the conclusion follows.

2.3 The Saddlepoint (Minimax Theorems)

Next we turn to show a situation where the solution of the equilibrium problem reduces to a *saddlepoint* of a bifunction. Let X, Y be two nonempty sets and $h : X \times Y \to \mathbb{R}$ be a given function. The pair $(x_0, y_0) \in X \times Y$ is called a *saddlepoint of h on the set $X \times Y$* if

$$h(x, y_0) \leq h(x_0, y_0) \leq h(x_0, y) \quad \forall (x, y) \in X \times Y. \tag{4}$$

Let $A = B = X \times Y$ and let $f : A \times B \to \mathbb{R}$ defined by

$$f(a, b) := h(x, v) - h(u, y) \quad \forall a = (x, y),\ b = (u, v). \tag{5}$$

Then each solution of the equilibrium problem (EP) is a saddlepoint of h and vice versa.

The saddlepoint can be characterized as follows. Suppose that for each $x \in X$ there exists $\min_{y \in Y} h(x, y)$ and for each $y \in Y$ there exists $\max_{x \in X} h(x, y)$. Then we have the following result.

Proposition 1. *f admits a saddlepoint on $X \times Y$ if and only if there exist $\max_{x \in X} \min_{y \in Y} f(x, y)$ and $\min_{y \in Y} \max_{x \in X} f(x, y)$ and they are equal.*

Proof. Suppose first that h admits a saddlepoint $(x_0, y_0) \in X \times Y$. Then by relation (4) one obtains

$$\min_{y \in Y} h(x, y) \leq h(x, y_0) \leq h(x_0, y_0) = \min_{y \in Y} h(x_0, y) \quad \forall x \in X$$

and

$$\max_{x \in X} h(x, y) \geq h(x_0, y) \geq h(x_0, y_0) = \max_{x \in X} h(x, y_0) \quad \forall y \in Y.$$

Therefore,

$$\min_{y \in Y} h(x_0, y) = \max_{x \in X} \min_{y \in Y} h(x, y)$$

and

$$\max_{x \in X} h(x, y_0) = \min_{y \in Y} \max_{x \in X} h(x, y),$$

and both equal to $h(x_0, y_0)$. For the reverse implication take $x_0 \in X$ such that

$$\min_{y\in Y} h(x_0, y) = \max_{x\in X} \min_{y\in Y} h(x, y)$$

and $y_0 \in Y$ such that

$$\max_{x\in X} h(x, y_0) = \min_{y\in Y} \max_{x\in X} h(x, y).$$

Then by our assumption we obtain

$$\min_{y\in Y} h(x_0, y) = \max_{x\in X} h(x, y_0);$$

therefore, in the obvious relations

$$\min_{y\in Y} h(x_0, y) \le h(x_0, y_0) \le \max_{x\in X} h(x, y_0)$$

one obtains equality in both sides. This completes the proof. □

Remark 1. Observe that, for arbitrary nonempty sets X, Y and function $h : X \times Y \to \mathbb{R}$, the inequality

$$\sup_{x\in X} \inf_{y\in Y} h(x, y) \le \inf_{y\in Y} \sup_{x\in X} h(x, y)$$

always holds. Therefore,

$$\max_{x\in X} \min_{y\in Y} h(x, y) \le \min_{y\in Y} \max_{x\in X} h(x, y)$$

holds either, provided these two values exist.

One of the main issues in minimax theory is to find sufficient and/or necessary conditions for the sets X, Y and function h, such that the reverse inequality in the above relations also holds. Such results are called *minimax theorems*.

Minimax theorems or, in particular, the existence of a saddlepoint, is important in many applied fields of mathematics. One of them is the *game theory*.

2.3.1 Two-Player Zero-Sum Games

To introduce a static two-player zero-sum (noncooperative) game (for more details and examples, see [2, 3, 20, 26, 27, 32]) and its relation to a minimax theorem we consider two players called 1 and 2 and assume that the set of pure strategies (also called actions) of player 1 is given by some nonempty set X, while the set of pure strategies of player 2 is given by a nonempty set Y. If player 1 chooses the pure strategy $x \in X$ and player 2 chooses the pure strategy $y \in Y$, then player 2 has to pay player 1 an amount $h(x, y)$ with

$h : A \times B \to R$ a given function. This function is called the payoff function of player 1. Since the gain of player 1 is the loss of player 2 (this is a so-called zero-sum game) the payoff function of player 2 is $-h$. Clearly player 1 likes to gain as much profit as possible. However, at the moment he does not know how to achieve this and so he first decides to compute a lower bound on his profit. To compute this lower bound player 1 argues as follows: if he decides to choose action $x \in X$, then it follows that his profit is at least $\inf_{y \in Y} h(x, y)$, irrespective of the action of player 2. Therefore a lower bound on the profit for player 1 is given by

$$r_* := \sup_{x \in X} \inf_{y \in Y} h(x, y). \tag{6}$$

Similarly player 2 likes to minimize his losses but since he does not know how to achieve this he also decides to compute first an upper bound on his losses. To do so, player 2 argues as follows. If he decides to choose action $y \in Y$, it follows that he loses at most $\sup_{x \in X} h(x, y)$ and this is independent of the action of player 1. Therefore an upper bound on his losses is given by

$$r^* := \inf_{y \in Y} \sup_{x \in X} h(x, y). \tag{7}$$

Since the profit of player 1 is at least r_* and the losses of player 2 are at most r^* and the losses of player 2 are the profits of player 1, it follows directly that $r_* \leq r^*$. In general $r_* < r^*$, but under some properties on the pure strategy sets and payoff function one can show that $r_* = r^*$. If this equality holds and in relations (6) and (7) the suprema and infima are attained, an optimal strategy for both players is obvious. By the interpretation of r_* for player 1 and the interpretation of r^* for player 2 and $r^* = r_* := v$ both players will choose an action which achieves the value v and so player 1 will choose that action $x_0 \in X$ satisfying

$$\inf_{y \in Y} h(x_0, y) = \max_{x \in X} \inf_{y \in Y} h(x, y).$$

Moreover, player 2 will choose that strategy $y_0 \in Y$ satisfying

$$\sup_{x \in X} h(x, y_0) = \min_{y \in Y} \sup_{x \in X} h(x, y).$$

Another field, where the concept of saddlepoint plays an important role, is the so-called *duality in optimization*.

2.3.2 Duality in Optimization

Let X be a nonempty subset of $\mathbb{R}^n$. A subset K of $\mathbb{R}^m$ is called *cone* if, for each $y \in K$ and $\lambda > 0$, it follows that $\lambda y \in K$. The set K is called *convex cone*, if K is a cone and additionally, a convex set. Let $F : \mathbb{R}^n \to \mathbb{R}$ and $G : \mathbb{R}^n \to \mathbb{R}^m$ be given functions. For K, a nonempty convex cone of $\mathbb{R}^m$, define the following optimization problem:

$$v(P) := \inf\{F(x) | \, G(x) \in -K, \, x \in X\}. \tag{8}$$

This (general) problem has many important particular cases.

The Optimization Problem with Inequality and Equality Constraints. Let $X := \mathbb{R}^n$, $K := \mathbb{R}^p_+ \times \{0_{\mathbb{R}^{m-p}}\}$, where $1 \leq p < m$, and $0_{\mathbb{R}^{m-p}}$ denotes the origin of the space $\mathbb{R}^{m-p}$. Then problem (8) reduces to the classical optimization problem with inequality and equality constraints

$$\inf\{F(x) | \, G_i(x) \leq 0, i = 1, 2, \ldots, p, \;\; G_j(x) = 0, j = p+1, \ldots, m\}.$$

The Linear Programming Problem. Let

$$X := \mathbb{R}^n_+, \quad K := \{0_{\mathbb{R}^m}\}, \quad F(x) := c^T x, \quad G(x) := Ax - b,$$

where A is a matrix with m rows and n columns (with all entries real numbers), $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ are given elements. Then (8) reduces to the following linear programming problem:

$$\inf\{c^T x | \, Ax = b, \, x \geq 0\}.$$

The Conical Programming Problem. Let $K \subseteq \mathbb{R}^n$ be a nonempty convex cone, let $X := b + L \subseteq \mathbb{R}^n$, where L is a linear subspace of $\mathbb{R}^n$, and let $F(x) := c^T x$, $G(x) := x$. Then we obtain the so-called *conical programming problem*

$$\inf\{c^T x | \, x \in b + L, \, x \in -K\}.$$

Denote by $\mathcal{F}$ the *feasible set* of problem (8), i.e., the set $\{x \in X | \, G(x) \in -K\}$. The problem

$$v(R) := \inf\{F_R(x) | \, x \in \mathcal{F}_R\}$$

is called a *relaxation* of the initial problem (8), if $\mathcal{F} \subseteq \mathcal{F}_R$ and $F_R(x) \leq F(x)$ for each $x \in \mathcal{F}$. It is obvious that $v(R) \leq v(P)$. Next we show a natural way to construct a relaxation of problem (8). Let $\lambda \in \mathbb{R}^m$ and consider the problem

$$\inf\{F(x) + \lambda^T G(x) | \, x \in X\}.$$

Clearly $\mathcal{F} \subseteq X$ and $F(x) + \lambda^T G(x) \leq F(x)$ for each $x \in \mathcal{F}$ if and only if $\lambda^T G(x) \leq 0$ for each $x \in \mathcal{F}$. Let $K^* := \{y \in \mathbb{R}^m | \, y^T x \geq 0 \;\; \forall x \in K\}$ be the *dual cone* of K. Now it is clear that $\lambda \in K^*$ implies $\lambda^T G(x) \leq 0$, for each $x \in \mathcal{F}$. Define the (Lagrangian) function $L : X \times K^* \to \mathbb{R}$ by $L(x, \lambda) := F(x) + \lambda^T G(x)$ and consider the problem

$$\theta(\lambda) := \inf\{L(x, \lambda) | \, x \in X\}. \tag{9}$$

Clearly $\theta(\lambda) \leq v(P)$ for each $\lambda \in K^*$, and therefore we also have

$$\sup_{\lambda \in K^*} \theta(\lambda) \leq v(P),$$

hence

$$\sup_{\lambda \in K^*} \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in \mathcal{F}} F(x). \tag{10}$$

By this relation it follows that the optimal objective value $v(D)$ of the *dual problem*

$$v(D) := \sup\{\theta(\lambda) | \, \lambda \in K^*\}$$

approximates from below the optimal objective value $v(P)$ of the primal problem (8). From both theoretical and practical points of view, an important issue is to establish sufficient conditions in order to have equality between the optimal objective values of the primal and dual problems. In this respect, observe that for each $x \in \mathcal{F}$ one has

$$\sup_{\lambda \in K^*} L(x, \lambda) = \sup_{\lambda \in K^*} (F(x) + \lambda^T G(x)) = F(x).$$

Therefore,

$$\inf_{x \in \mathcal{F}} F(x) = \inf_{x \in \mathcal{F}} \sup_{\lambda \in K^*} L(x, \lambda) = \inf_{x \in X} \sup_{\lambda \in K^*} L(x, \lambda).$$

Indeed, if $x \in X \setminus \mathcal{F}$, then $G(x) \notin -K$. By the *bipolar theorem* [29] we have $K = K^{**}$, hence it follows that there exists $\lambda^* \in K^*$ such that ${\lambda^*}^T G(x) > 0$. Since $t\lambda^* \in K$ for each $t > 0$, then

$$\sup_{\lambda \in K^*} L(x, \lambda) = \infty \quad \forall x \in X \setminus \mathcal{F}.$$

Combining the latter with relation (10) and taking into account that the "supinf" is always less or equal than the "infsup," one obtains

$$v(D) = \sup_{\lambda \in K^*} \inf_{x \in X} L(x, \lambda) \leq \inf_{x \in X} \sup_{\lambda \in K^*} L(x, \lambda) = v(P). \tag{11}$$

Hence we obtain that $v(D) = v(P)$, if a saddlepoint $(\bar{x}, \bar{\lambda})$ of the Lagrangian L exists. This situation is called *perfect duality*. In this case $\bar{x}$ is the optimal solution of the primal, while $\bar{\lambda}$ is the optimal solution of the dual problem.

2.4 Variational Inequalities

Let E be a real topological vector space and E^* be the dual space of E. Let $K \subseteq E$ be a nonempty convex set and $T : K \to E^*$ a given operator. For $x \in E$ and $x^* \in E^*$, the *duality pairing* between these two elements will be denoted by $\langle x, x^* \rangle$. If $A = B := K$ and $f(x, y) := \langle T(x), y - x \rangle$, for each $x, y \in K$, then each solution of the equilibrium problem (EP) is a solution of the *variational inequality*

$$\langle T(x),\, y - x\rangle \geq 0 \quad \forall y \in K, \tag{12}$$

and vice versa.

Variational inequalities have shown to be important mathematical models in the study of many real problems, in particular in network equilibrium models ranging from spatial price equilibrium problems and imperfect competitive oligopolistic market equilibrium problems to general financial or traffic equilibrium problems.

An important particular case of the variational inequality (12) is the following. Let $E := H$ be a real Hilbert space with inner product $\langle\ ,\ \rangle$. It is well known that in this case the dual space E^* can be identified with H. Consider the bilinear and continuous function $a : H \times H \to \mathbb{R}$, the linear and continuous function $L : H \to \mathbb{R}$, and formulate the problem: find an element $x \in K \subseteq H$ such that

$$a(x,\, y - x) \geq L(y - x) \quad \forall y \in K. \tag{13}$$

By the hypothesis, for each $x \in H$ the function $a(x, \cdot) : H \to \mathbb{R}$ is linear and continuous. Therefore, by the Riesz representation theorem in Hilbert spaces (see, for instance, [30]) there exists a unique element $A(x) \in H$ such that $a(x,\, y) = \langle A(x),\, y\rangle$ for each $y \in H$. It is easy to see that $A : H \to H$ is a linear and continuous operator. Moreover, since L is also linear and continuous, again by the Riesz theorem, there exists a unique element $l \in H$ such that $L(x) = \langle l,\, x\rangle$ for each $x \in H$. Now for $T(x) := A(x) - l$, problem (13) reduces to (12).

In optimization theory, those variational inequalities in which the operator T is a gradient map (i.e., is the gradient of a certain differentiable function) are of special interest since their solutions are (in some cases) the *minimum points* of the function itself. Suppose that $X \subseteq \mathbb{R}^n$ is an open set, $K \subseteq X$ is a convex set, and the function $F : X \to \mathbb{R}$ is differentiable on X. Then each minimum point of F on the set K is a solution of the variational inequality (12), with $T := \nabla F$. Indeed, let $x_0 \in K$ be a minimum point of F on K and $y \in K$ be an arbitrary element. Then we have

$$F(x_0) \leq F(\lambda y + (1 - \lambda)x_0) \quad \forall \lambda \in [0, 1].$$

Therefore,

$$\frac{1}{\lambda}(F(x_0 + \lambda(y - x_0)) - F(x_0)) \geq 0 \quad \forall \lambda \in (0, 1].$$

Now letting $\lambda \to 0$ we obtain $\langle \nabla F(x_0),\, y - x_0\rangle \geq 0$, as claimed.

If we suppose further that F is a *convex function* on the convex set X, then we obtain the reverse implication as well, i.e., each solution of the variational inequality (12), with $T := \nabla F$, is a minimum point of F on the set K. Indeed, let $x_0 \in K$ be a solution of (12) and $y \in K$ be an arbitrary element. Then by convexity

$$F(x_0 + \lambda(y - x_0)) \leq (1 - \lambda)F(x_0) + \lambda F(y) \quad \forall \lambda \in [0, 1],$$

which yields

$$\frac{1}{\lambda}(F(x_0 + \lambda(y - x_0)) - F(x_0)) \leq F(y) - F(x_0) \quad \forall \lambda \in (0, 1].$$

By letting $\lambda \to 0$ one obtains from the latter that

$$\langle \nabla F(x_0),\, y - x_0 \rangle \leq F(y) - F(x_0),$$

which yields the desired implication.

The particular cases presented above shows the importance of the equilibrium problem (EP). Therefore, one of the main issues is to know in advance whether (EP) admits a solution. In the next section we give sufficient conditions for the existence of a solution of this problem.

3 Some Existence Results on Equilibrium Problem

There are many results concerning the existence of solutions of (EP) known in the literature. Usually, regarding their proofs, they can be divided into two classes: results that uses fixed point tools and results using separation tools. There are, however, some results (usually consequences of more general statements) that belong to both classes. The aim of this section is to present two classical results from the first class due to Ky Fan [15] and Brézis, Nirenberg, Stampacchia [11], and a more recent result belonging to the second class due to Kassay and Kolumbán [23].

3.1 Results Based on Fixed Point Theorems

To start, let us first recall the celebrated Brouwer's fixed point theorem.

Theorem 2. *Let $C \subseteq \mathbb{R}^n$ be a convex, compact set and $h : C \to C$ be a continuous function. Then h admits at least one fixed point.*

Since the appearance of this theorem, many different proofs of it have been published. It is still an open question whether there exists an elementary proof of Brouwer's fixed point theorem in case $n \geq 2$, using separation arguments only.

By Theorem 2 one can prove some of the so-called *intersection theorems*, which are useful tools regarding existence results for the equilibrium problem. The first important intersection theorem has been published in 1929: the celebrated *Knaster–Kuratowski–Mazurkiewicz's* theorem [25] (called in the literature *KKM lemma*). This result has been extended by Ky Fan [14] in 1961 to infinite dimensional spaces. We will formulate these results later in this section as particular cases of a recent result obtained by Chang and Zhang [12]. In order to present the latter we first need the following definitions. Let E and E' be two topological vector spaces and let X be a nonempty subset of E.

Definition 1. *The set-valued mapping $F : X \to 2^E$ is called* **KKM mapping**, *if $co\{x_1, \ldots, x_n\} \subseteq \cup_{i=1}^n F(x_i)$ for each finite subset $\{x_1, \ldots, x_n\}$ of X.*

A slightly more general concept was introduced by Chang and Zhang [12]:

Definition 2. *The mapping $F : X \to 2^{E'}$ is called* **generalized KKM mapping,** *if for any finite set $\{x_1, \ldots, x_n\} \subseteq X$, there exists a finite set $\{y_1, \ldots, y_n\} \subseteq E'$, such that for any subset $\{y_{i_1}, \ldots, y_{i_k}\} \subseteq \{y_1, \ldots, y_n\}$, we have*

$$\mathrm{co}\{y_{i_1}, \ldots, y_{i_k}\} \subseteq \bigcup_{j=1}^{k} F(x_{i_j}). \tag{14}$$

In case $E = E'$ it is clear that every KKM mapping is a generalized KKM mapping too. The converse of this implication is not true, as the following example shows.

Example 1. (Chang and Zhang [12]). Let $E := \mathbb{R}$, $X := [-2, 2]$ and $F : X \to 2^E$ be defined by

$$F(x) := [-(1 + x^2/5),\ 1 + x^2/5].$$

Since $\cup_{x \in X} F(x) = [-9/5, 9/5]$, we have

$$x \notin F(x) \quad \forall x \in [-2, -9/5) \cup (9/5, 1].$$

This shows that F is not a KKM mapping. On the other hand, for any finite subset $\{x_1, \ldots, x_n\} \subseteq X$, take $\{y_1, \ldots, y_n\} \subseteq [-1, 1]$. Then for any $\{y_{i_1}, \ldots, y_{i_k}\} \subseteq \{y_1, \ldots, y_n\}$ we have

$$\mathrm{co}\{y_{i_1}, \ldots, y_{i_k}\} \subseteq [-1, 1] = \bigcap_{x \in X} F(x) \subseteq \bigcup_{j=1}^{k} F(x_{i_j}),$$

i.e., F is a generalized KKM mapping.

Theorem 3. (Chang and Zhang [12]). *Suppose that E is a Hausdorff topological vector space, $X \subseteq E$ is nonempty, and $F : X \to 2^E$ is a mapping such that for each $x \in X$ the set $F(x)$ is finitely closed (i.e., for every finite dimensional subspace L of E, $F(x) \cap L$ is closed in the Euclidean topology in L). Then F is a generalized KKM mapping if and only if for every finite subset $I \subseteq X$ the intersection of the subfamily $\{F(x) | \, x \in I\}$ is nonempty.*

Proof. Suppose first that for arbitrary finite set $I = \{x_1, \ldots, x_n\} \subseteq X$ one has

$$\bigcap_{i=1}^{n} F(x_i) \neq \emptyset.$$

Take $x_* \in \cap_{i=1}^n F(x_i)$ and put $y_i := x_*$, for each $i \in \{1, \ldots, n\}$. Then for every $\{y_{i_1}, \ldots, y_{i_k}\} \subseteq \{y_1, \ldots, y_n\}$ we have

$$\mathrm{co}\{y_{i_1}, \ldots, y_{i_k}\} = \{x_*\} \subseteq \bigcap_{i=1}^{n} F(x_i) \subseteq \bigcap_{j=1}^{k} F(x_{i_j}).$$

This implies that F is a generalized KKM mapping.

To show the reverse implication, let $F : X \to 2^E$ be a generalized KKM mapping. Supposing the contrary, there exists some finite set $\{x_1, \ldots, x_n\} \subseteq X$ such that $\cap_{i=1}^n F(x_i\text{`}) = \emptyset$. By the assumption, there exists a set $\{y_1, \ldots, y_n\} \subseteq E$ such that for any $\{y_{i_1}, \ldots, y_{i_k}\} \subseteq \{y_1, \ldots, y_n\}$, relation (14) holds. In particular, we have

$$\mathrm{co}\{y_1, \ldots, y_n\} \subseteq \bigcup_{i=1}^n F(x_i).$$

Let $S := \mathrm{co}\{y_1, \ldots, y_n\}$ and $L := \mathrm{span}\{y_1, \ldots, y_n\}$. Since for each $x \in X$, $F(x)$ is finitely closed, then the sets $F(x_i) \cap L$ are closed. Let d be the Euclidean metric on L. It is easy to verify that

$$d(x,\, F(x_i) \cap L) > 0 \quad \text{if and only if} \quad x \notin F(x_i) \cap L. \tag{15}$$

Define now the function $g : S \to \mathbb{R}$ by

$$g(c) := \sum_{i=1}^n d(c,\, F(x_i) \cap L), \quad c \in S.$$

It follows by (15) and $\cap_{i=1}^n F(x_i) = \emptyset$ that for each $c \in S$, $g(c) > 0$. Let

$$h(c) := \sum_{i=1}^n \frac{1}{g(c)} d(c,\, F(x_i) \cap L) y_i.$$

Then h is a continuous function from S to S. By the Brouwer's fixed point theorem (Theorem 2), there exists an element $c_* \in S$ such that

$$c_* = h(c_*) = \sum_{i=1}^n \frac{1}{g(c_*)} d(c_*,\, F(x_i) \cap L) y_i. \tag{16}$$

Denote

$$I := \{i \in \{1, \ldots, n\} |\, d(c_*,\, F(x_i) \cap L) > 0\}. \tag{17}$$

Then for each $i \in I$, $c_* \notin F(x_i) \cap L$. Since $c_* \in L$, then $c_* \notin F(x_i)$ for each $i \in I$, or, in other words,

$$c_* \notin \bigcup_{i \in I} F(x_i). \tag{18}$$

By (16) and (17) we have

$$c_* = \sum_{i=1}^n \frac{1}{g(c_*)} d(c_*,\, F(x_i) \cap L) y_i \in \mathrm{co}\{y_i |\, i \in I\}.$$

Since F is a generalized KKM mapping, this leads to

$$c_* \in \bigcup_{i \in I} F(x_i),$$

which contradicts (18). This completes the proof. □

By the above theorem one can easily deduce the following result.

Theorem 4. (Chang and Zhang [12]) *Suppose that $F : X \to 2^E$ is a set-valued mapping such that for each $x \in X$, the set $F(x)$ is closed. If there exists an element $x_0 \in X$ such that $F(x_0)$ is compact, then $\cap_{x\in X} F(x) \neq \emptyset$ if and only if F is a generalized KKM mapping.*

The proof of this theorem is an easy consequence of Theorem 3.

As we mentioned in the first part of this section, a particular case of Theorem 3 is the intersection theorem due to Ky Fan, known in the literature as *Ky Fan's lemma*.

Theorem 5. (Ky Fan [14]) *Let E be a Hausdorff topological vector space, $X \subseteq E$ and for each $x \in X$, let $F(x)$ be a closed subset of E, such that*

(a) *there exists $x_0 \in X$, such that the set $F(x_0)$ is compact;*
(b) *for each $x_1, x_2, \ldots, x_n \in X$, $\quad \mathrm{co}\{x_1, x_2, \ldots, x_n\} \subseteq \cup_{i=1}^{n} F(x_i)$.*

Then

$$\bigcap_{x\in X} F(x) \neq \emptyset.$$

To conclude our presentation concerning intersection theorems, let us mention the famous result of Knaster, Kuratowski, and Mazurkiewitz (known as KKM lemma).

Theorem 6. (KKM [25]) *Let $E_i \subseteq \mathbb{R}^n$ be closed sets and $e_i \in E_i$, $i = 1, \ldots, m$. Suppose that for each $J \subseteq \{1, \ldots, m\}$ we have $\mathrm{co}\{e_j | \, j \in J\} \subseteq \cup_{j\in J} E_j$. Then*

$$\bigcap_{i=1}^{m} E_i \neq \emptyset.$$

Now let us turn back to the equilibrium problem (EP). In what follows we need some further definitions.

Definition 3. *Let X be a convex subset of a certain vector space and let $h : X \to \mathbb{R}$ be some function. Then h is said to be* **quasiconvex** *if for every $x_1, x_2 \in X$ and $0 < \lambda < 1$*

$$h(\lambda x_1 + (1-\lambda)x_2) \leq \max\{h(x_1), h(x_2)\}.$$

We say that h is **quasiconcave** *if $-h$ is quasiconvex.*

It is easy to check that h is quasiconvex if and only if the lower level sets $\{x \in X | \, h(x) \leq a\}$ are convex for each $a \in \mathbb{R}$. Similarly, h is quasiconcave if and only if the upper level sets $\{x \in X | \, h(x) \geq a\}$ are convex for each $a \in \mathbb{R}$. It is also easy to see that in the statements above, relations $\leq$ ($\geq$) can be replaced with $<$ ($>$) and the assertions remain valid.

Definition 4. *Let X be a topological space and let $h : X \to \mathbb{R}$ be some function. Then h is said to be* **lower semicontinuous (lsc in short) on** *X if the lower level sets $\{x \in X | \, h(x) \leq a\}$ are closed for each $a \in \mathbb{R}$. h is said to be* **upper semicontinuous (usc in short) on** *X if $-h$ is lsc on X, that is, its upper level sets are all closed.*

By means of Ky Fan's theorem (Theorem 5) one can prove the following existence result for (EP), due also to Ky Fan. This is known in the literature as *Ky Fan's minimax inequality theorem.*

Theorem 7. (Ky Fan [15]) *Let A be a nonempty, convex, compact subset of a Hausdorff topological vector space and let $f : A \times A \to \mathbb{R}$, such that*

$$\forall b \in A, \quad f(\cdot, b) : A \to \mathbb{R} \text{ is usc}, \tag{19}$$

$$\forall a \in A, \quad f(a, \cdot) : A \to \mathbb{R} \text{ is quasiconvex} \tag{20}$$

and

$$\forall a \in A, \quad f(a, a) \geq 0. \tag{21}$$

Then (EP) admits a solution.

Proof. For each $b \in A$, consider the set $F(b) := \{a \in A | \, f(a, b) \geq 0\}$. By (19), these sets are closed, and since A is compact, they are compact too. It is easy to see that the conclusion of the theorem is equivalent to

$$\bigcap_{b \in A} F(b) \neq \emptyset. \tag{22}$$

In order to prove relation (22), let $b_1, b_2, \ldots, b_n \in A$. We shall show that

$$\text{co}\{b_i | \, i \in \{1, 2, \ldots, n\}\} \subseteq \bigcup_{i=1}^{n} F(b_i). \tag{23}$$

Indeed, suppose by contradiction that there exist $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$, $\sum_{j=1}^{n} \lambda_j = 1$, such that

$$\sum_{j=1}^{n} \lambda_j b_j \notin \bigcup_{j=1}^{n} F(b_j).$$

By definition, the latter means

$$f\left(\sum_{j=1}^{n} \lambda_j b_j, \, b_i\right) < 0 \quad \forall i \in \{1, 2, \ldots, n\}.$$

By (20) (quasiconvexity), one obtains

$$f\left(\sum_{j=1}^{n} \lambda_j b_j, \sum_{j=1}^{n} \lambda_j b_j\right) < 0,$$

which contradicts (21). This shows that (23) holds. Now applying Theorem 5, we obtain (22), which completes the proof. □

As we have seen, the basic tool in the proof of Theorem 3 (and 4) of Chang and Zhang was the Brouwer's fixed point theorem (Theorem 2). Moreover, Ky Fan's intersection (and consequently his minimax inequality theorems (Theorems 5 and 7)), follow by Theorem 4. On the other hand, as we show next, by Theorem 7 one can easily reobtain the Brouwer's fixed point theorem, which means that all these mentioned results are equivalent. To do this, we first state the following result.

Theorem 8. *Let E be a normed space, $X \subseteq E$ be a compact convex set, and $g, h : X \to E$ be continuous functions such that*

$$\|x - g(x)\| \geq \|x - h(x)\| \quad \forall x \in X. \tag{24}$$

Then there exists an element $x_0 \in X$, such that

$$\|y - g(x_0)\| \geq \|x_0 - h(x_0)\| \quad \forall y \in X.$$

Proof. Let $f : X \times X \to \mathbb{R}$ defined by $f(x, y) := \|y - g(x)\| - \|x - h(x)\|$. It is clear that this function satisfies the hypothesis of Theorem 7; thus there exists an element $x_0 \in X$ such that

$$\|x_0 - h(x_0)\| \leq \|y - g(x_0)\| \quad \forall y \in X. \tag{25}$$

This completes the proof. □

Observe, in case $g(X) \subseteq X$, we can put $y := g(x_0)$ in (25); in this way we obtain that x_0 is a fixed point of f. Now it is immediate the well-known Schauder's fixed point theorem:

Theorem 9. (Schauder [31]) *Let X be a convex compact subset of a real normed space and $h : X \to X$ a continuous function. Then h has a fixed point.*

Proof. Taking $h = g$ in the previous theorem, we obtain this result by (25), with $y := h(x_0)$. □

Clearly, Brouwer's fixed point theorem (Theorem 2) is a particular case of Theorem 9.

3.2 Results Based on Separation Theorems

As announced at the beginning of this section, we present now some existence results on (EP) which uses separation tools in their proofs.

The result below is a particular case of a theorem due to Kassay and Kolumbán [23].

Theorem 10. *Let A be a nonempty, compact, convex subset of a certain topological vector space, let B be a nonempty convex subset of a certain vector space, and let $f : A \times B \to \mathbb{R}$ be a given function.*

Suppose that the following assertions are satisfied:

(a) *f is usc and concave in its first variable;*
(b) *f is convex in its second variable;*
(c) $\sup_{a\in A} f(a,b) \geq 0$, *for each* $b \in B$.

Then the equilibrium problem (EP) has a solution.

Remark 2. Condition (c) in the previous theorem is satisfied if, for instance, $B \subseteq A$ and $f(a,a) \geq 0$ for each $a \in B$. This condition arises naturally in most of the particular cases presented above.

A similar, but more general existence result for the problem (EP) has been established by Kassay and Kolumbán also in [23], where instead of the convexity (concavity) assumptions upon the function f, certain kind of generalized convexity (concavity) assumptions are supposed.

Theorem 11. *Let A be a compact topological space, let B be a nonempty set, and let $f : A \times B \to \mathbb{R}$ be a given function such that*

(a) *for each $b \in B$, the function $f(\cdot, b) : A \to \mathbb{R}$ is usc;*
(b) *for each $a_1, \ldots, a_m \in A$, $b_1, \ldots, b_k \in B$, $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_{i=1}^m \lambda_i = 1$, the inequality*

$$\min_{1\leq j\leq k} \sum_{i=1}^{m} \lambda_i f(a_i, b_j) \leq \sup_{a\in A} \min_{1\leq j\leq k} f(a, b_j)$$

holds;
(c) *For each $b_1, \ldots, b_k \in B$, $\mu_1, \ldots, \mu_k \geq 0$ with $\sum_{j=1}^k \mu_j = 1$, one has*

$$\sup_{a\in A} \sum_{j=1}^{k} \mu_j f(a, b_j) \geq 0.$$

Then the equilibrium problem (EP) admits a solution.

Proof. Suppose by contradiction that (EP) has no solution, i.e., for each $a \in A$ there exists $b \in B$ such that $f(a,b) < 0$ or, equivalently, for each $a \in A$ there exists $b \in B$ and $c > 0$ such that $f(a,b) + c < 0$. Denote by $U_{b,c}$ the set $\{a \in A |\, f(a,b) + c < 0\}$ where $b \in B$ and $c > 0$. By (a) and our assumption, the family of these sets is an open covering of the compact set A. Therefore, one can select a finite subfamily which covers the same set A, i.e., there exist $b_1, \ldots, b_k \in B$ and $c_1, \ldots, c_k > 0$ such that

$$A = \bigcup_{j=1}^{k} U_{b_j, c_j}. \tag{26}$$

Let $c := \min\{c_1, \ldots, c_k\} > 0$ and define the vector-valued function $H : A \to \mathbb{R}^k$ by

$$H(a) := (f(a, b_1) + c, \ldots, f(a, b_k) + c).$$

We show that

$$\text{co}H(A) \cap \text{int}\mathbb{R}^k_+ = \emptyset, \tag{27}$$

where co$H(A)$ denotes the convex hull of the set $H(A)$ and int$\mathbb{R}^k_+$ denotes the interior of the positive orthant $\mathbb{R}^k_+$. Indeed, supposing the contrary, there exist $a_1, \ldots, a_m \in A$ and $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_{i=1}^m \lambda_i = 1$, such that

$$\sum_{i=1}^m \lambda_i H(a_i) \in \text{int}\mathbb{R}^k_+$$

or, equivalently,

$$\sum_{i=1}^m \lambda_i (f(a_i, b_j) + c) > 0 \quad \forall j \in \{1, \ldots, k\}. \tag{28}$$

By (b), (28) implies

$$\sup_{a \in A} \min_{1 \leq j \leq k} f(a, b_j) > -c. \tag{29}$$

Now using (26), for each $a \in A$ there exists $j \in \{1, \ldots, k\}$ such that $f(a, b_j) + c_j < 0$. Thus, for each $a \in A$ we have

$$\min_{1 \leq j \leq k} f(a, b_j) < -c,$$

which contradicts (29). This shows that relation (27) is true. By the well-known separation theorem of two disjoint convex sets in finite dimensional spaces (see, for instance, [29]), the sets co$H(A)$ and int$\mathbb{R}^k_+$ can be separated by a hyperplane, i.e., there exist $\mu_1, \ldots, \mu_k \geq 0$ such that $\sum_{j=1}^k \mu_j = 1$ and

$$\sum_{j=1}^k \mu_j (f(a, b_j) + c) \leq 0 \quad \forall a \in A,$$

or, equivalently

$$\sum_{j=1}^k \mu_j f(a, b_j) \leq -c \quad \forall a \in A. \tag{30}$$

Observe, the latter relation contradicts assumption (c) of the theorem. Thus the proof is complete. □

4 The Equilibrium Problem and the Ekeland's Principle

Due to its important applications, the problem of solving an equilibrium problem is an important task. However, it often happens, an equilibrium problem

may not have solution even in case when the problem arises from practice. Therefore, it is important to find approximate solutions in some sense or to show their existence in case of an equilibrium problem.

The Ekeland's variational principle (see, for instance, [13]) has been widely used in nonlinear analysis since it entails the existence of approximate solutions of a minimization problem for lower semicontinuous functions on a complete metric space. Since, as we have seen in Section 2, minimization problems are particular cases of equilibrium problems, one is interested in extending Ekeland's theorem to the setting of an equilibrium problem.

Recently, inspired by the study of systems of vector variational inequalities, Ansari, Schaible, and Yao [1] introduced and investigated systems of equilibrium problems, which are defined as follows. Let m be a positive integer. By a system of equilibrium problems we understand the problem of finding $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m) \in A$ such that

$$f_i(\bar{x}, y_i) \geq 0 \qquad \forall i \in I, \quad \forall y_i \in A_i, \tag{SEP}$$

where $f_i : A \times A_i \to \mathbb{R}$, $A = \prod_1^m A_i$, with A_i some given sets.

The aim of this section is to present some recent results concerning existence of approximate equilibria for (EP) and (SEP). We find a suitable set of conditions on the functions that do not involve convexity and lead to an Ekeland's variational principle for equilibrium and system of equilibrium problems. Via the existence of approximate solutions, we are able to show the existence of equilibria on general closed sets. Our setting is an Euclidean space, even if the results could be extended to reflexive Banach spaces, by adapting the assumptions in a standard way.

4.1 The Ekeland's Principle for (EP) and (SEP)

To start, let us recall the celebrated Ekeland's variational principle established within the framework of minimization problems for lower semicontinuous functions on complete metric spaces.

Theorem 12. (Ekeland [13]) *Let (X, d) be a complete metric space and $F : X \to \mathbb{R}$ a lower bounded, lower semicontinuous function. Then for every $\varepsilon > 0$ and $x_0 \in X$ there exists $\bar{x} \in X$ such that*

$$\begin{cases} \varepsilon d(x_0, \bar{x}) \leq F(x_0) - F(\bar{x}) \\ F(\bar{x}) < F(x) + \varepsilon d(\bar{x}, x) \quad \forall x \in X, \quad x \neq x_0. \end{cases} \tag{31}$$

Remark 3. If $X = \mathbb{R}$ with the Euclidean norm, then (31) can be written as

$$\begin{cases} \varepsilon |x_0 - \bar{x}| \leq F(x_0) - F(\bar{x}) \\ F(\bar{x}) < F(x) + \varepsilon |\bar{x} - x| \quad \forall x \in X, \quad x \neq x_0, \end{cases}$$

and this relation has a clear geometric interpretation.

Starting from Theorem 12, in a most recent paper [5] the authors established the following general result which we present here in detail.

Theorem 13. *Let A be a closed set of $\mathbb{R}^n$ and $f : A \times A \to \mathbb{R}$. Assume that the following conditions are satisfied:*

(a) *$f(x, \cdot)$ is lower bounded and lower semicontinuous, for every $x \in A$;*

(b) *$f(t, t) = 0$, for every $t \in A$;*

(c) *$f(z, x) \leq f(z, y) + f(y, x)$, for every $x, y, z \in A$.*

Then, for every $\varepsilon > 0$ and for every $x_0 \in A$, there exists $\overline{x} \in A$ such that

$$\begin{cases} f(x_0, \overline{x}) + \varepsilon \|x_0 - \overline{x}\| \leq 0 \\ f(\overline{x}, x) + \varepsilon \|\overline{x} - x\| > 0 \quad \forall x \in A, \quad x \neq \overline{x}. \end{cases} \tag{32}$$

Proof. Without loss of generality, we can restrict the proof to the case $\varepsilon = 1$. Denote by $\mathcal{F}(x)$ the set

$$\mathcal{F}(x) := \{y \in A : f(x, y) + \|y - x\| \leq 0\}.$$

By (a), $\mathcal{F}(x)$ is closed, for every $x \in A$; by (b), $x \in \mathcal{F}(x)$, hence $\mathcal{F}(x)$ is nonempty for every $x \in A$. Assume $y \in \mathcal{F}(x)$, i.e., $f(x, y) + \|y - x\| \leq 0$, and let $z \in \mathcal{F}(y)$ (i.e., $f(y, z) + \|y - z\| \leq 0$). Adding both sides of the inequalities, we get, by (c),

$$0 \geq f(x, y) + \|y - x\| + f(y, z) + \|y - z\| \geq f(x, z) + \|z - x\|,$$

that is, $z \in \mathcal{F}(x)$. Therefore $y \in \mathcal{F}(x)$ implies $\mathcal{F}(y) \subseteq \mathcal{F}(x)$.

Define

$$v(x) := \inf_{z \in \mathcal{F}(x)} f(x, z).$$

For every $z \in \mathcal{F}(x)$,

$$\|x - z\| \leq -f(x, z) \leq \sup_{z \in \mathcal{F}(x)} (-f(x, z)) = - \inf_{z \in \mathcal{F}(x)} f(x, z) = -v(x)$$

that is,

$$\|x - z\| \leq -v(x) \qquad \forall z \in \mathcal{F}(x).$$

In particular, if $x_1, x_2 \in \mathcal{F}(x)$,

$$\|x_1 - x_2\| \leq \|x - x_1\| + \|x - x_2\| \leq -v(x) - v(x) = -2v(x),$$

implying that

$$\operatorname{diam}(\mathcal{F}(x)) \leq -2v(x) \qquad \forall x \in A.$$

Fix $x_0 \in A$; $x_1 \in \mathcal{F}(x_0)$ exists such that

$$f(x_0, x_1) \leq v(x_0) + 2^{-1}.$$

Denote by x_2 any point in $\mathcal{F}(x_1)$ such that

$$f(x_1, x_2) \leq v(x_1) + 2^{-2}.$$

Proceeding in this way, we define a sequence $\{x_n\}$ of points of A such that $x_{n+1} \in \mathcal{F}(x_n)$ and

$$f(x_n, x_{n+1}) \leq v(x_n) + 2^{-(n+1)}.$$

Notice that

$$\begin{aligned} v(x_{n+1}) &= \inf_{y \in \mathcal{F}(x_{n+1})} f(x_{n+1}, y) \geq \inf_{y \in \mathcal{F}(x_n)} f(x_{n+1}, y) \\ &\geq \inf_{y \in \mathcal{F}(x_n)} (f(x_n, y) - f(x_n, x_{n+1})) \left(\inf_{y \in \mathcal{F}(x_n)} f(x_n, y) \right) - f(x_n, x_{n+1}) \\ &= v(x_n) - f(x_n, x_{n+1}). \end{aligned}$$

Therefore,

$$v(x_{n+1}) \geq v(x_n) - f(x_n, x_{n+1})$$

and

$$-v(x_n) \leq -f(x_n, x_{n+1}) + 2^{-(n+1)} \leq (v(x_{n+1}) - v(x_n)) + 2^{-(n+1)},$$

that entails

$$0 \leq v(x_{n+1}) + 2^{-(n+1)}.$$

It follows that

$$\operatorname{diam}(\mathcal{F}(x_n)) \leq -2v(x_n) \leq 2 \cdot 2^{-n} \to 0, \quad n \to \infty.$$

The sets $\{\mathcal{F}(x_n)\}$ being closed and $\mathcal{F}(x_{n+1}) \subseteq \mathcal{F}(x_n)$, we have that

$$\bigcap_n \mathcal{F}(x_n) = \{\overline{x}\}.$$

Since $\overline{x} \in \mathcal{F}(x_0)$, then

$$f(x_0, \overline{x}) + \|\overline{x} - x_0\| \leq 0.$$

Moreover, $\overline{x}$ belongs to all $\mathcal{F}(x_n)$, and, since $\mathcal{F}(\overline{x}) \subseteq \mathcal{F}(x_n)$, for every n, we get that

$$\mathcal{F}(\overline{x}) = \{\overline{x}\}.$$

It follows that $x \notin \mathcal{F}(\overline{x})$ whenever $x \neq \overline{x}$, implying that

$$f(\overline{x}, x) + \|x - \overline{x}\| > 0.$$

This completes the proof. □

Remark 4. It is easy to see that any function $f(x,y) = g(y) - g(x)$ trivially satisfies (c) (actually with equality). One might wonder whether a bifunction f satisfying all the assumptions of Theorem 13 should be of the form $g(y)-g(x)$, and as such reducing the result above to the classical Ekeland's principle. It is not the case, as the example below shows: let the function $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x,y) = \begin{cases} e^{-\|x-y\|} + 1 + g(y) - g(x) & x \neq y \\ 0 & x = y \end{cases},$$

where g is a lower bounded and lower semicontinuous function. Then all the assumptions of Theorem 13 are satisfied, but clearly f cannot be represented in the above-mentioned form.

Next we shall extend the result above for a system of equilibrium problems. Let m be a positive integer and $I = \{1, 2, \ldots, m\}$. Consider the functions $f_i : A \times A_i \to \mathbb{R}$, $i \in I$, where $A = \prod_{i\in I} A_i$, and $A_i \subseteq X_i$ is a closed subset of the Euclidean space X_i. An element of the set $A^i = \prod_{j\neq i} A_j$ will be represented by x^i; therefore, $x \in A$ can be written as $x = (x^i, x_i) \in A^i \times A_i$. If $x \in \prod X_i$, the symbol $|||x|||$ will denote the Tchebiseff norm of x, i.e., $|||x||| = \max_i ||x_i||_i$ and we shall consider the Euclidean space $\prod X_i$ endowed with this norm.

Theorem 14. (Bianchi et al. [5]) *Assume that*

(a) $f_i(x,\cdot) : A_i \to \mathbb{R}$ *is lower bounded and lower semicontinuous for every* $i \in I$;
(b) $f_i(x, x_i) = 0$ *for every* $i \in I$ *and every* $x = (x_1, \ldots, x_m) \in A$;
(c) $f_i(z, x_i) \leq f_i(z, y_i) + f_i(y, x_i)$, *for every* $x, y, z \in A$, *where* $y = (y^i, y_i)$, *and for every* $i \in I$.

Then for every $\varepsilon > 0$ *and for every* $x^0 = (x_1^0, \ldots, x_m^0) \in A$ *there exists* $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_m) \in A$ *such that for each* $i \in I$ *one has*

$$f_i(x^0, \bar{x}_i) + \varepsilon \|x_i^0 - \bar{x}_i\|_i \leq 0 \tag{33}$$

and

$$f_i(\bar{x}, x_i) + \varepsilon \|\bar{x}_i - x_i\|_i > 0 \quad \forall x_i \in D_i, \quad x_i \neq \bar{x}_i. \tag{34}$$

Proof. As before, we restrict the proof to the case $\varepsilon = 1$. Let $i \in I$ be arbitrarily fixed. Denote for every $x \in A$

$$\mathcal{F}_i(x) := \{y_i \in A_i : f_i(x, y_i) + \|x_i - y_i\|_i \leq 0\}.$$

These sets are closed and nonempty (for every $x = (x_1, \ldots, x_m) \in A$ we have $x_i \in \mathcal{F}_i(x)$). Define for each $x \in A$

$$v_i(x) := \inf_{z_i \in \mathcal{F}_i(x)} f_i(x, z_i).$$

In a similar way as in the proof of Theorem 13 one can show that $\operatorname{diam}(\mathcal{F}_i(x)) \leq -2v_i(x)$ for every $x \in A$ and $i \in I$.

Fix now $x^0 \in A$ and select for each $i \in I$ an element $x_i^1 \in \mathcal{F}_i(x^0)$ such that

$$f_i(x^0, x_i^1) \leq v_i(x^0) + 2^{-1}.$$

Put $x^1 := (x_1^1, \ldots, x_m^1) \in A$ and select for each $i \in I$ an element $x_i^2 \in \mathcal{F}_i(x^1)$ such that

$$f_i(x^1, x_i^2) \leq v_i(x^1) + 2^{-2}.$$

Put $x^2 := (x_1^2, \ldots, x_m^2) \in A$. Continuing this process we define a sequence $\{x^n\}$ in A such that $x_i^{n+1} \in \mathcal{F}_i(x^n)$ for each $i \in I$ and $n \in N$ and

$$f_i(x^n, x_i^{n+1}) \leq v_i(x^n) + 2^{-(n+1)}.$$

Using a same argument as in the proof of Theorem 13 one can show that

$$\operatorname{diam}(\mathcal{F}_i(x^n)) \leq -2v_i(x^n) \leq 2 \cdot 2^{-n} \to 0, \quad n \to \infty,$$

for each $i \in I$.

Now define for each $x \in A$ the sets

$$\mathcal{F}(x) := \mathcal{F}_1(x) \times \cdots \times \mathcal{F}_m(x) \subseteq A.$$

The sets $\mathcal{F}(x)$ are closed and using (c) it is immediate to check that for each $y \in \mathcal{F}(x)$ it follows that $\mathcal{F}(y) \subseteq \mathcal{F}(x)$. Therefore, we also have $\mathcal{F}(x^{n+1}) \subseteq \mathcal{F}(x^n)$ for each $n \in \{0, 1, \ldots\}$. On the other hand, for each $y, z \in \mathcal{F}(x^n)$ we have

$$|||y - z||| = \max_{i \in I} \|y_i - z_i\|_i \leq \max_{i \in I} \operatorname{diam}\mathcal{F}_i(x^n)) \to 0,$$

thus, $\operatorname{diam}(\mathcal{F}(x^n)) \to 0$ as $n \to \infty$. In conclusion we have

$$\bigcap_{n=0}^{\infty} \mathcal{F}(x^n) = \{\bar{x}\}, \ \bar{x} \in A.$$

Since $\bar{x} \in \mathcal{F}(x^0)$, i.e., $\bar{x}_i \in \mathcal{F}_i(x^0)$ $(i \in I)$ we obtain

$$f_i(x^0, \bar{x}_i) + \|x_i^0 - \bar{x}_i\|_i \leq 0 \quad \forall i \in I,$$

and so, (33) holds. Moreover, $\bar{x} \in \mathcal{F}(x^n)$ implies $\mathcal{F}(\bar{x}) \subseteq \mathcal{F}(x^n)$ for all $n = 0, 1, \ldots$, therefore,

$$\mathcal{F}(\bar{x}) = \{\bar{x}\}$$

implying

$$\mathcal{F}_i(\bar{x}) = \{\bar{x}_i\} \quad \forall i \in I.$$

Now for every $x_i \in A_i$ with $x_i \neq \bar{x}_i$ we have by the previous relation that $x_i \notin \mathcal{F}_i(\bar{x})$ and so

$$f_i(\bar{x}, x_i) + \|\bar{x}_i - x_i\|_i > 0.$$

Thus (34) holds too, and this completes the proof. □

4.2 New Existence Results for Equilibria on Compact Sets

As shown by the literature, the existence results of equilibrium problems usually require some convexity (or generalized convexity) assumptions on at least one of the variables of the function involved. In this section, using Theorems 13 and 14, we show the nonemptiness of the solution set of (EP) and (SEP), without any convexity requirement. To this purpose, we recall the definition of approximate equilibrium point, for both cases (see [5, 21]). We start our analysis with (EP).

Definition 5. *Given $f : A \times A \to \mathbb{R}$ and $\varepsilon > 0$, $\overline{x} \in A$ is said to be an ε-equilibrium point of f if*

$$f(\overline{x}, y) \geq -\varepsilon \|\overline{x} - y\| \qquad \forall y \in A \tag{35}$$

The ε-equilibrium point is strict, if in (35) the inequality is strict for all $y \neq \overline{x}$.

Notice that the second relation of (31) gives the existence of a strict ε-equilibrium point, for every $\varepsilon > 0$. Moreover, by (b) and (c) of Theorem 12 it follows by the first relation of (31) that

$$f(\overline{x}, x_0) \geq \varepsilon \|\overline{x} - x_0\|,$$

"localizing," in a certain sense, the position of $\overline{x}$.

Theorem 12 leads to a set of conditions that are sufficient for the nonemptiness of the solution set of (EP).

Proposition 2. (Bianchi et al. [5]) *Let A be a compact (not necessarily convex) subset of an Euclidean space and $f : A \times A \to \mathbb{R}$ be a function satisfying the assumptions:*

(a) *$f(x, \cdot)$ is lower bounded and lower semicontinuous, for every $x \in A$;*
(b) *$f(t, t) = 0$, for every $t \in A$;*
(c) *$f(z, x) \leq f(z, y) + f(y, x)$, for every $x, y, z \in A$;*
(d) *$f(\cdot, y)$ is upper semicontinuous, for every $y \in A$.*

Then, the set of solutions of EP is nonempty.

Proof. For each $n \in \mathbb{N}$, let $x_n \in A$ a $1/n$-equilibrium point (such point exists by Theorem 12), i.e.,

$$f(x_n, y) \geq -\frac{1}{n}\|x_n - y\| \qquad \forall y \in A.$$

Since A is compact, we can choose a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \to \overline{x}$ as $n \to \infty$. Then, by (d),

$$f(\overline{x}, y) \geq \limsup_{k \to \infty} \left(f(x_{n_k}, y) + \frac{1}{n_k}\|x_{n_k} - y\| \right) \qquad \forall y \in A,$$

thereby proving that $\overline{x}$ is a solution of EP. □

Let us now consider the following definition of ε-equilibrium point for systems of equilibrium problems. As before, the index set I consists of the finite set $\{1, 2, \ldots, m\}$.

Definition 6. *Let A_i, $i \in I$ be subsets of certain Euclidean spaces and put $A = \prod_{i \in I} A_i$. Given $f_i : A \times A_i \to \mathbb{R}$, $i \in I$, and $\varepsilon > 0$, the point $\overline{x} \in A$ is said to be an ε-equilibrium point of $\{f_1, f_2, \ldots, f_m\}$ if*

$$f_i(\overline{x}, y_i) \geq -\varepsilon \|\overline{x}_i - y_i\|_i \qquad \forall y_i \in A_i, \ \ \forall i \in I.$$

The following result is an extension of Proposition 2, and it can be proved in a similar way.

Proposition 3. (Bianchi et al. [5]) *Assume that, for every $i \in I$, A_i is compact and $f_i : A \times A_i \to \mathbb{R}$ is a function satisfying the assumptions:*

(a) *$f_i(x, \cdot)$ is lower bounded and lower semicontinuous, for every $x \in A$;*
(b) *$f_i(x, x_i) = 0$, for every $x = (x^i, x_i) \in A$;*
(c) *$f_i(z, x_i) \leq f_i(z, y_i) + f_i(y, x_i)$, for every $x, y, z \in A$, where $y = (y^i, y_i)$;*
(d) *$f_i(\cdot, y_i)$ is upper semicontinuous, for every $y_i \in A_i$.*

Then, the set of solutions of (SEP) is nonempty.

4.3 Equilibria on Noncompact Sets

The study of the existence of solutions of the equilibrium problems on unbounded domains usually involves the same sufficient assumptions as for bounded domains together with a coercivity condition. Bianchi and Pini [7] found coercivity conditions as weak as possible, exploiting the generalized monotonicity properties of the function f defining the equilibrium problem.

Let A be a closed subset of X, not necessarily convex, not necessarily compact, and $f : A \times A \to \mathbb{R}$ be a given function.

Consider the following coercivity condition (see [7]):

$$\exists r > 0 : \quad \forall x \in A \setminus K_r, \quad \exists y \in A, \quad \|y\| < \|x\| : \ f(x, y) \leq 0, \tag{36}$$

where $K_r := \{x \in A : \|x\| \leq r\}$.

We now show that within the framework of Proposition 2 condition (36) guarantees the existence of solutions of (EP) without supposing compactness of A.

Theorem 15. (Bianchi et al. [5]) *Suppose that*

(a) *$f(x, \cdot)$ is lower bounded and lower semicontinuous, for every $x \in A$;*
(b) *$f(t, t) = 0$, for every $t \in A$;*
(c) *$f(z, x) \leq f(z, y) + f(y, x)$, for every $x, y, z \in A$;*
(d) *$f(\cdot, y)$ is upper semicontinuous, for every $y \in A$.*

If (36) holds, then (EP) admits a solution.

Proof. We may suppose without loss of generality that K_r is nonempty. For each $x \in A$ consider the nonempty set

$$S(x) := \{y \in A : \|y\| \leq \|x\| : f(x,y) \leq 0\}.$$

Observe that for every $x, y \in A$, $y \in S(x)$ implies $S(y) \subseteq S(x)$. Indeed, for $z \in S(y)$ we have $\|z\| \leq \|y\| \leq \|x\|$ and by (c) $f(x,z) \leq f(x,y) + f(y,z) \leq 0$. On the other hand, since $K_{\|x\|}$ is compact, by (a) we obtain that $S(x) \subseteq K_{\|x\|}$ is a compact set for every $x \in A$. Furthermore, by Proposition 2, there exists an element $x_r \in K_r$ such that

$$f(x_r, y) \geq 0 \quad \forall y \in K_r. \tag{37}$$

Suppose that there exists $x \in A$ with $f(x_r, x) < 0$ and put

$$a := \min_{y \in S(x)} \|y\|$$

(the minimum is taken since $S(x)$ is nonempty, compact and the norm is continuous). We distinguish two cases.

Case 1: $a \leq r$. Let $y_0 \in S(x)$ such that $\|y_0\| = a \leq r$. Then we have $f(x, y_0) \leq 0$. Since $f(x_r, x) < 0$, it follows by (c) that

$$f(x_r, y_0) \leq f(x_r, x) + f(x, y_0) < 0,$$

contradicting (37).

Case 2: $a > r$. Let again $y_0 \in S(x)$ such that $\|y_0\| = a > r$. Then, by (36) we can choose an element $y_1 \in A$ with $\|y_1\| < \|y_0\| = a$ such that $f(y_0, y_1) \leq 0$. Thus, $y_1 \in S(y_0) \subseteq S(x)$ contradicting

$$\|y_1\| < a = \min_{y \in S(x)} \|y\|.$$

Therefore, there is no $x \in A$ such that $f(x_r, x) < 0$, i.e., x_r is a solution of (EP) (on A). This completes the proof. □

Next we consider (SEP) for noncompact setting. Let us consider the following coercivity condition:

$$\begin{aligned} &\exists r > 0 : \quad \forall x \in A \text{ such that } \|x_i\|_i > r \text{ for some } i \in I, \\ &\quad \exists y_i \in A_i, \quad \|y_i\|_i < \|x_i\|_i \quad \text{and} \quad f_i(x, y_i) \leq 0. \end{aligned} \tag{38}$$

We conclude this section with the following result which guarantees the existence of solutions for (SEP).

Theorem 16. (Bianchi et al. [5]) *Suppose that, for every $i \in I$,*

(a) *$f_i(x, \cdot)$ is lower bounded and lower semicontinuous, for every $x \in A$;*
(b) *$f_i(x, x_i) = 0$, for every $x = (x^i, x_i) \in A$;*

(c) $f_i(z, x_i) \leq f_i(z, y_i) + f_i(y, x_i)$, *for every* $x, y, z \in A$, *where* $y = (y^i, y_i)$;
(d) $f_i(\cdot, y_i)$ *is upper semicontinuous, for every* $y_i \in A_i$.

If (38) holds, then (SEP) admits a solution.

Proof. For each $x \in A$ and every $i \in I$ consider the set

$$S_i(x) := \{y_i \in A_i,\ \|y_i\|_i \leq \|x_i\|_i,\ f_i(x, y_i) \leq 0\}.$$

Observe that, by (c), for every x and $y = (y^i, y_i) \in A$, $y_i \in S_i(x)$ implies $S_i(y) \subseteq S_i(x)$. On the other hand, since the set $\{y_i \in A_i : \|y_i\|_i \leq r\} = K_i(r)$ is a compact subset of A_i, by (a) we obtain that $S_i(x)$ is a nonempty compact set for every $x \in A$. Furthermore, by Proposition 3, there exists an element $x_r \in \prod_i K_i(r)$ (observe, we may suppose that $K_i(r) \neq \emptyset$ for all $i \in I$) such that

$$f_i(x_r, y_i) \geq 0 \quad \forall y_i \in K_i(r), \qquad \forall i \in I. \tag{39}$$

Suppose that x_r is not a solution of (SEP). In this case, there exists $j \in I$ and $z_j \in A_j$ with $f_j(x_r, z_j) < 0$. Let $z^j \in A^j$ be arbitrary and put $z = (z^j, z_j) \in A$. Define

$$a_j := \min_{y_j \in S_j(z)} \|y_j\|_j.$$

We distinguish two cases.

Case 1: $a_j \leq r$. Let $\overline{y}_j(z) \in S_j(z)$ such that $\|\overline{y}_j(z)\|_j = a_j \leq r$. Then we have $f_j(z, \overline{y}_j(z)) \leq 0$. Since $f_j(x_r, z_j) < 0$, it follows by (c) that

$$f_j(x_r, \overline{y}_j(z)) \leq f(x_r, z_j) + f(z, \overline{y}_j(z)) < 0,$$

contradicting (39).

Case 1: $a_j > r$. Let again $\overline{y}_j(z) \in S_j(z)$ such that $\|\overline{y}_j(z)\|_j = a_j > r$. Let $\overline{y}^j \in A^j$ be arbitrary and put $\overline{y}(z) = (\overline{y}^j, \overline{y}_j(z)) \in A$. Then, by (38) we can choose an element $y_j \in A_j$ with $\|y_j\|_j < \|\overline{y}_j(z)\|_j = a_j$ such that $f_j(\overline{y}(z), y_j) \leq 0$. Clearly, $y_j \in S_j(\overline{y}(z)) \subseteq S_j(z)$, a contradiction since $\overline{y}_j(z)$ has minimal norm in $S_j(z)$. This completes the proof. □

5 Conclusions

Finally, let us recall the most important issues discussed in this chapter. As emphasized in Introduction, our purpose was to give an overlook on equilibrium problem (abbreviated (EP)) underlining its importance and usefulness from both theoretical and practical points of view.

In the second section we have presented the most important particular cases of (EP). One of them is the optimization problem (minimization/maximization of a real-valued function over a so-called feasible set). As

well known, optimization problems appear as mathematical models of many problems of practical interest. Another particular case of (EP) presented here is the so-called Kirszbraun's problem, which can be successfully applied in extending nonexpansive functions (these functions are important among others, in fixed point theory). The saddlepoint (or minimax) problems have shown to be also particular instances of (EP). We have pointed out the applicability of these problems in game theory on one hand and in duality theory in optimization, on the other hand. We have concluded the presentation of the particular cases of (EP) with variational inequalities, which constitute models of various problems arising from mechanics and economy.

Section 3 has been devoted to the exposition of some classical and recent results concerning existence of solutions of (EP). We have underlined that in general these results can be deduced in two ways: either using fixed point tools or separation (Hahn–Banach) tools. For the reader's convenience, the most important results of this section have been presented together with their proofs. Moreover, we have tried to keep these proofs as simple as possible.

When dealing with (EP), one frequently encounters the situation when the set of solutions is empty. In these situations it is important to study the existence of approximate solutions in some sense. Since (EP) contains, in particular, optimization problems, and the celebrated Ekeland's variational principle provides the existence of approximate optimal solutions, it comes natural to investigate whether this principle can be extended to (EP). Based on recent results of the author, we have presented in the last section some of these possible extensions both for (EP) and a more general situation: system of equilibrium problems (SEP).

Throughout this chapter we have limited ourselves to the scalar case, i.e., when the functions involved in (EP) or (SEP) are real-valued. In the last decade the vector-valued case has also been studied (see, for instance, [1, 4, 16]). We think that a possible research for the future could be to investigate whether the results presented here for the scalar case can be extended also for the vector case.

References

1. Ansari, Q.H., Schaible, S., Yao, J.C.: System of vector equilibrium problems and its applications. J. Optim. Theory Appl. 107, 547–557 (2000)
2. Aubin, J.P.: Mathematical Methods of Game and Economic Theory, North Holland, Amsterdam (1979)
3. Başar, T., Olsder, G.J.: Dynamic Noncooperative Game Theory (2nd ed.), SIAM, Philadelphia (1999)
4. Bianchi, M., Hadjisavvas, N., Schaible, S.: Vector equilibrium problems with generalized monotone bifunctions. J. Optim. Theory Appl. 92, 527–542 (1997)
5. Bianchi, M., Kassay, G., Pini, R.: Existence of equilibria via Ekeland's principle. J. Math. Anal. Appl. 305, 502–512 (2005)

6. Bianchi, M., Pini, R.: A note on equilibrium problems with properly quasimonotone bifunctions. J. Global Optim. 20, 67–76 (2001)
7. Bianchi, M., Pini, R.: Coercivity conditions for equilibrium problems. J. Optim. Theory Appl. 124, 79–92 (2005)
8. Bianchi, M., Schaible, S.: Generalized monotone bifunctions and equilibrium problems. J. Optim. Theory Appl. 90, 31–43 (1996)
9. Bigi, G., Castellani, M., Kassay, G.: A dual view of equilibrium problems. J. Math. Anal. Appl. 342, 17–26 (2008)
10. Blum, E., Oettli, W.: From optimization and variational inequalities to equilibrium problems. Math. Stud. 63, 123–145 (1994)
11. Brézis, H., Nirenberg, G., Stampacchia, G.: A remark on Ky Fan's minimax principle. Bollettino U.M.I. 6, 293–300 (1972)
12. Chang, S.S., Zhang, Y.: Generalized KKM theorem and variational inequalities. J. Math. Anal. Appl. 159, 208–223 (1991)
13. Ekeland, I.: On the variational principle. J. Math. Anal. Appl. 47, 324–353 (1974)
14. Fan, K.: A generalization of Tychonoff's fixed point theorem. Math. Ann. 142, 305–310 (1961)
15. Fan, K.: A minimax inequality and its application, In: O. Shisha (Ed.), Inequalities (Vol. 3, pp. 103–113), Academic, New York (1972)
16. Finet, C., Quarta, L., Troestler, C.: Vector-valued variational principles. Nonlinear Anal. 52, 197–218 (2003)
17. Iusem, A.N., Kassay, G., Sosa, W.: On certain conditions for the existence of solutions of equilibrium problems. Math. Program. 116, 259–273 (2009) http://dx.doi.org/10.1007/s10107-007-0125-5
18. Iusem, A.N., Sosa, W.: New existence results for equilibrium problems. Nonlinear Anal. 52, 621–635 (2003)
19. Iusem, A.N., Sosa, W.: Iterative algorithms for equilibrium problems. Optimization 52, 301–316 (2003)
20. Jones, A.J.: Game Theory: Mathematical Models of Conflict, Horwood Publishing, Chichester (2000)
21. Kas, P., Kassay, G., Boratas-Sensoy, Z.: On generalized equilibrium points. J. Math. Anal. Appl. 296, 619–633 (2004)
22. Kassay, G.: The Equilibrium Problem and Related Topics, Risoprint, Cluj-Napoca (2000)
23. Kassay, G., Kolumbán, J.: On a generalized sup-inf problem. J. Optim. Theory Appl. 91, 651–670 (1996)
24. Kirszbraun, M.D.: Über die Zusammenziehenden und Lipschitzschen Transformationen. Fund. Math. 22, 7–10 (1934)
25. Knaster, B., Kuratowski, C., Mazurkiewicz, S.: Ein Beweis des Fixpunktsatzes für n-dimensionale Simplexe. Fund. Math. 14, 132–138 (1929)
26. Kuhn, H.W.: Lectures on the Theory of Games, Princeton University Press, Princeton, NJ (2003)
27. von Neumann, J.: Zur theorie der gesellschaftsspiele. Math. Ann. 100, 295–320 (1928)
28. Oettli, W.: A remark on vector-valued equilibria and generalized monotonicity. Acta Math. Vietnam. 22, 215–221 (1997)
29. Rockafellar, R.T.: Convex Analysis, Princeton University Press, Princeton, NJ (1970)

30. Rudin, W.: Principles of Mathematical Analysis, McGraw-Hill, New York, NY (1976)
31. Schauder, J.: Der Fixpunktsatz in Funktionalräumen. Studia Math. 2, 171–180 (1930)
32. Vorob'ev, N.N.: Game Theory: Lectures for Economists and Systems Scientists, Springer, New York, NY (1977)

Scalarly Compactness, $(S)_+$-Type Conditions, Variational Inequalities, and Complementarity Problems in Banach Spaces

George Isac

Department of Mathematics, Royal Military College of Canada, P.O. Box 17000 STN Forces Kingston, Ontario, K7K 7B4, Canada
isac-g@rmc.ca

Summary. We present in this chapter the notion of scalarly compactness which is related to condition $(S)_+$, well known in nonlinear analysis. Some applications to the study of variational inequalities and to complementarity problems are also presented.

Key words: Scalarly compactness, Variational inequalities and complementarity problems

1 Introduction

The main goal of this chapter is to present a topological method applicable to the study of solvability of variational inequalities and of complementarity problems in reflexive Banach spaces.

Our topological method is based on scalarly compactness and on $(S)_+$-type conditions. The notions of *scalarly compact* operator is strongly related to condition $(S)_+$, defined and used by Browder [3–6].

We note that condition $(S)_+$ is an important mathematical tool used in nonlinear analysis. There exists also a topological degree defined for mapping which satisfies condition $(S)_+$ [28].

The notion of scalarly compact operator was defined by Isac [21]. Now we present in this chapter several examples of scalarly compact operators and we will conclude that this is a remarkable class of nonlinear operators. We will use also the notion of scalar asymptotic derivative. Our main results are solvability theorems for variational inequalities and for complementarity problems, considered in reflexive Banach spaces and defined by a difference of two operators. The first operator is supposed to satisfy an $(S)_+$-type condition and the second is supposed to be a scalarly compact operator.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_4,

The variational inequalities have many applications in physics, engineering, and in other domains of applied mathematics [1, 24].

Complementarity problems are generally related to the equilibrium as it is considered in physics, engineering, and economics [12–14, 19, 22, 23]. Also, complementarity theory has interesting applications to optimization.

In Hilbert spaces, variational inequalities and complementarity problems have been studied by *KKM*-type theorems or by the fixed point theory. A variational inequality or a nonlinear complementarity problem in a Hilbert space can be transformed into a fixed point problem using the projection operator onto a closed convex set [12–18]. We note that the fixed point method cannot be used in Banach spaces, and therefore the method presented now in this chapter may be considered as a new direction in the study of variational inequalities and of complementarity problems.

2 Preliminaries

Let $(E, \|\cdot\|)$ be a Banach space and let $\mathbb{K} \subset E$ be a closed convex cone, i.e., $\mathbb{K}$ is a closed set satisfying the following properties:

$k_1)$ $\mathbb{K} + \mathbb{K} \subseteq \mathbb{K}$,
$k_2)$ $\lambda\mathbb{K} \subseteq \mathbb{K}$ for any $\lambda \in \mathbb{R}_+$,
$k_3)$ $\mathbb{K} \cap (-\mathbb{K}) = \{0\}$.

If E^* is the topological dual of E, we denote by $\langle E, E^* \rangle$ a duality (pairing) between E and E^*, where $\langle \cdot, \cdot \rangle$ is the canonical bilinear form of this duality. We denote by $\mathbb{K}^*$ the dual cone of $\mathbb{K}$, that is, $\mathbb{K} = \{y \in E^* \mid \langle x, y, \rangle \geq 0 \text{ for any } x \in \mathbb{K}\}$. Given a mapping $f : E \to E^*$, the *general nonlinear complementarity problem* associated with f and $\mathbb{K}$ is

$$\mathrm{NCP}(f, \mathbb{K}) : \begin{cases} \text{find } x_0 \in \mathbb{K} \text{ such that} \\ f(x_0) \in \mathbb{K}^* \text{ and } \langle x_0, f(x_0) \rangle = 0. \end{cases}$$

If D is a non-empty closed convex subset in E, the variational inequality associated with f and D is

$$\mathrm{VI}(f, D) : \begin{cases} \text{find } x_0 \in D \text{ such that} \\ \langle x - x_0, f(x_0) \rangle \geq 0 \text{ for any } x \in D. \end{cases}$$

We recall that a mapping $f : E \to E^*$ is *completely continuous* if f is continuous and for any bounded set $B \subset E$, $f(B)$ is relatively compact, and we say that f is *demicontinuous* if for any sequence $\{x_n\}_{n \in N} \subset E$, convergent in norm to an element x_* we have that $\{f(x_n)\}_{n \in N}$ is weakly $(*)$-convergent to $f(x_*)$. We say that f is bounded if, for any bounded set B, $f(B)$ is bounded. We say that a Banach space $(E, \|\cdot\|)$ is a *Kadeç space* if for each sequence $\{x_n\}_{n \in N} \subset E$ which converges weakly to x_* with $\lim_{n \to \infty} \|x_n\| = \|x_*\|$ we have that, $\lim_{n \to \infty} \|x_n - x_*\| = 0$. Any space, $L^p(\Omega, \mu)$, $(1 < p < \infty)$, any

uniformly convex space, and any locally uniformly convex Banach space are *Kadeç spaces*.

We recall that a Banach space E is said to be *strictly convex* if for every $x, y \in E$ with $x \neq y$, $\|x\| = \|y\|$ we have that $\|\lambda x + (1-\lambda)y\| < 1$ for every $\lambda \in]0,1[$. Equivalently, a Banach space E is strictly convex if, $x, y \in E$, $\|x\| = \|y\| = 1$ and $x \neq y$ imply $\|x + y\| < 2$ [8, 29].

We say that a Banach space $E, \|\cdot\|$ is *uniformly convex* if for any $\epsilon \in]0,2]$, there exists $\delta > 0$ depending only on $\epsilon > 0$ such that $\|x + y\| \leq 2(1-\delta)$, for any $x, y \in E$ with $\|x\| = \|y\| = 1$ and $\|x - y\| \geq \epsilon$. More general, we say that a Banach space $(E, \|\cdot\|)$ is locally *uniformly convex* if for any $\epsilon > 0$ and any x with $\|x\| = 1$ there exists $\delta(\epsilon, x) > 0$ such that the inequality $\|x - y\| \geq \epsilon$ implies $\|x + y\| \leq 2(1 - \delta(\epsilon, x))$ for any $y \in E$ with $\|y\| = 1$.

Obviously, every uniformly convex Banach space is locally uniformly convex and reflexive. Every locally uniformly convex Banach space is strictly convex. Any Hilbert space is strictly convex and uniformly convex. Finally we recall the following form of the classical Eberlein–Šmulian theorem.

Theorem 1. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, A a bounded subset of E, and x_0 a point in the weak closure of A. Then there exists an infinite sequence $\{x_n\}_{n\in N}$ in A converging weakly to x_0 in E.*

About the form of Theorem 1 the reader is referred to [5].

3 $(S)_+$-Type Conditions

We present in this section some $(S)_+$-type conditions based on the classical conditions $(S), (S)_+$, and $(S)_0$ defined by Browder and used in several papers [3–6]. We note that $(S)_+$ is a fundamental condition used in nonlinear analysis [2, 30]. Generally this condition is used when in some problems related to functional equations the compactness is absent.

In the classical conditions $(S), (S)_+$, and $(S)_0$ the general scheme is the following: If a sequence $\{x_n\}_{n\in N}$ is weakly convergent to an element x_* and some special conditions are satisfied then the sequence $\{x_n\}_{n\in N}$ is convergent in norm to x_*.

In the conditions introduced in this section we will use a conclusion as in the compactness case, that is, the sequence $\{x_n\}_{n\in N}$ has a subsequence convergent in norm to x_*. This modification is useful in some situations.

Let $(E, \|\cdot\|)$ be a Banach space, E^* the topological dual of E, and $\langle E, E^*\rangle$ a duality (pairing) between E and E^*. Let $D \subseteq E$ be a non-empty subset.

Definition 1. *A mapping $f : E \to E^*$ is said to satisfy condition $(S)_+$ with respect to D if any sequence $\{x_n\}_{n\in N} \subset D$ weakly convergent to an element $x_* \in E$ and satisfying the property $\limsup_{n\to\infty} \langle x_n - x_*, f(x_n)\rangle \leq 0$ has a subsequence $\{x_{n_k}\}_{k\in N}$ convergent in norm to x_*.*

Definition 2. *We say that a mapping $f : E \to E^*$ satisfies condition (S) with respect to D if any sequence $\{x_n\}_{n\in N} \subset D$ weakly convergent to an element $x_* \in E$ and such that $\lim_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle = 0$ has a subsequence $\{x_{n_k}\}_k \in N$ convergent in norm to x_*.*

Proposition 1. *If a mapping $f : E \to E^*$ satisfies condition $(S)_+$ with respect to a subset $D \subset E$, then f satisfies condition (S).*

Proof. Let $\{x_n\}_{n\in N} \subset D$ be a sequence weakly convergent to an element $x_* \in E$ and such that, $\lim_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle = 0$. We have

$$\langle x_n - x_*, f(x_n)\rangle = \langle x_n - x_*, f(x_n) - f(x_*)\rangle + \langle x_n - x_*, f(x_*)\rangle,$$

which implies

$$\begin{aligned}\limsup_{n\to\infty}\langle x_n - x_*, f(x_n)\rangle \le &\lim_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle \\ &+ \lim_{n\to\infty}\langle x_n - x_*, f(x_*)\rangle = 0.\end{aligned}$$

Because f satisfies condition $(S)_+$ we obtain that the sequence $\{x_n\}_{n\in N}$ has a subsequence $\{x_{n_k}\}_{k\in N}$ convergent in norm to x_*. Therefore, f satisfies condition (S).

A variant of condition (S) is the condition defined by the following definition.

Definition 3. *We say that a mapping $f : E \to E^*$ satisfies condition (S) with respect to D if any sequence $\{x_n\}_{n\in N} \subset D$ weakly convergent to an element $x_* \in E$ and such that $\lim_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle \le 0$ has a subsequence $\{x_{n_k}\}_{k\in N}$ convergent in norm to x_*.*

We have the following result.

Proposition 2. *If a mapping $f : E \to E^*$ satisfies condition $(S)_+$ with respect to a subset $D \subset E$, then f satisfies condition (S).*

Proof. The proof is similar to the proof of Proposition 1.

Indeed, if $\{x_n\}_{n\in N} \subset D$ is a sequence weakly convergent to an element $x_* \in E$ and $\limsup_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle \le 0$ then we have

$$\begin{aligned}\limsup_{n\to\infty}\langle x_n - x_*, f(x_n)\rangle \le &\limsup_{n\to\infty}\langle x_n - x_*, f(x_n) - f(x_*)\rangle \\ &+ \limsup_{n\to\infty}\langle x_n - x_*, f(x_n)\rangle \le 0.\end{aligned}$$

Because f satisfies condition $(S)_+$ we have that $\{x_n\}_{n\in N}$ has a subsequence convergent in norm to x_*.

The following condition is due to Isac and it was introduced in [20].

Definition 4. *We say that a mapping $f : E \to E^*$ satisfies condition $(S)^1_+$ with respect to D if any sequence $\{x_n\}_{n\in N} \in D$ weakly convergent to an element $x_* \in E$ and such that $\{f(x_n)\}_{n\in N}$ is weakly $(*)$-convergent to an element $u \in E^*$ and $\limsup_{n\to\infty}\langle x_n, f(x_n)\rangle \le \langle x_*, u\rangle$ has a subsequence convergent in norm to x_*.*

Several examples of mappings satisfying condition $(S)^1_+$ are given in [20]. It is known that any mapping which satisfies condition $(S)_+$ satisfies also condition $(S)^1_+$.

Now we recall the notion of *duality* mapping between E and E^*.

We say that a continuous and strictly increasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a weight if $\phi(0) = 0$ and $\lim_{r\to+\infty} \phi(r) = +\infty$.

We recall that, given a weight ϕ a *duality mapping* on E associated with ϕ is a mapping $J : E \to 2^{E^*}$ such that, $J(x) = \{x^* \in E^* \mid \langle x, x^*\rangle = \|x\|\|x^*\|$ *and* $\|x^*\|_* = \phi(\|x\|)\}$. We recall also that a Banach space $(E, \|\cdot\|)$ is *strictly convex* if for two elements $x, y \in E$ which are linearly independent we have $\|x + y\| < \|x\| + \|y\|$ (see [29]).

The following results are known [8, 29]. A duality mapping is a monotone operator and it is strictly monotone if E is strictly convex. If $(E, \|\cdot\|)$ is a reflexive Banach space with $(E^*, \|\cdot\|*)$ strictly convex then a duality mapping associated with a weight function ϕ is a demicontinuous point-to-point mapping.

If $(E, \|\cdot\|)$ is a Banach space which is a Kadeç space such that E^* is strictly convex, then any duality mapping $J : E \to E^*$ associated with a weight ϕ satisfies condition $(S)^1_+$. A proof of this result is in [20].

We note that the class of operators satisfying condition $(S)^1_+$ is invariant under completely continuous perturbations, i.e., if $f_1 : E \to E^*$ satisfies $(S)_+$ and $f_2 : E \to E^*$ is completely continuous then $f_1 + f_2$ satisfies $(S)_+$. When E is a Hilbert space any completely continuous vector field, i.e., a mapping of the form $f = I - g$, where I is the identity mapping and $g : E \to E$ is completely continuous, satisfies condition $(S)_+$.

Also any strongly ρ-monotone mapping $f : E \to E^*$ satisfies condition $(S)_+$ (see [20]). The reader can find other examples of mappings satisfying condition $(S)_+$ in [4–6, 15]. Conditions (S) and $(S)_+$ have many applications in nonlinear analysis [2–6, 15, 28, 30]. We note that there exists a topological degree for mappings of class $(S)_+$ [28]. Condition $(S)^1_+$ has interesting applications to the complementarity theory and to the study of variational inequalities [7, 9–11, 17, 20]. We note that condition $(S)^1_+$ can be defined also for multivalued mappings [10].

Definition 5. *We say that a mapping $f : E \to E^*$ satisfies condition $(S)_0$ with respect to D if any sequence $\{x_n\}_{n\in N} \subset D$ weakly convergent to an element $x_* \in E$ and such that $\{f(x_n)\}_{n\in N}$ is weakly $(*)$-convergent to an element $u \in E^*$ and $\lim_{n\to\infty}\langle x_n, f(x_n)\rangle = \langle x_*, u\langle$ has a subsequence convergent in norm to x_*.*

From Definition 5 we have the following result.

Proposition 3. *If a mapping $f : E \to E^*$ satisfies condition $(S)^1_+$ with respect to $D \subset E$, then f satisfies condition $(S)_0$.*

Definition 6. *We say that a mapping $f : E \to E^*$ satisfies condition (M) with respect to E if any sequence $\{x_n\}_{n\in N}$ weakly convergent to an element x_* such that $\{f(x_n)\}_{n\in N}$ is weakly $(*)$-convergent to an element $u \in E^*$ and $\lim_{n\to\infty} \sup\langle x_n, f(x_n)\rangle \leq \langle x_*, u\rangle$, we have that $f(x_*) = u$.*

We note that condition (M) is very much used in the study of solvability of nonlinear equations [27]. Examples of mappings satisfying condition (M) are given in [27].

It is easy to prove that if f is continuous and satisfies condition $(S)^1_+$ then f satisfies condition (M).

Using Lemma 1 of [3] we can prove that if $f : E \to E^*$ is continuous and satisfies condition $(S)^1_+$, then for any bounded closed set $B \subset E$, we have that $f(B)$ is a closed set in E^*.

Finally, let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $h : H \to H$ a mapping. We recall that h is a ϕ-contraction (in Boyd and Wong's sense) if there is a mapping satisfying

(i) $\|h(x) - h(y)\| \leq \phi(\|x - y\|)$, for any $x, y \in H$,
(ii) $\phi(t) < t$, for any $t \in \mathbb{R}_+ \setminus \{0\}$.

It is known that if h is a ϕ-contraction then the mapping $f = I - h$ satisfies condition $(S)^1_+$ [15].

4 Scalar Asymptotic Derivatives

Let $(E, \|\cdot\|)$ be a Banach space and E^* the topological dual of E. Let $\langle\cdot,\cdot\rangle$ be a duality (pairing) between E and E^*, that is, $\langle\cdot,\cdot\rangle$ is a separable bilinear mapping from $E \times E^*$ into $\mathbb{R}$. Let $\mathcal{L}(E, E^*)$ be the Banach space of linear continuous mappings from E into E^*. Let $\mathbb{K} \subset E$ be an *unbounded closed convex set*. We suppose that $0 \in \mathbb{K}$. The set $\mathbb{K}$ can be in particular a closed convex cone.

Definition 7. *We say that is a scalar asymptotic derivative of a mapping $f : E \to E^*$ along the set $\mathbb{K}$ if*

$$\lim_{\substack{\|x\| \to \infty \\ x \in \mathbb{K}}} \sup \frac{\langle x, f(x) - T(x)\rangle}{\|x\|^2} \leq 0.$$

In this case we denote the linear mapping T by f^∞_s.

The notion of scalar asymptotic derivative is due to Isac [16]. The origin of the name of this kind of derivative is its relation with the notion of scalar derivative due to Németh (see [21]). The following notion is a classical notion due to Krasnoselskii [25, 26], and it is an important tool in nonlinear analysis.

Definition 8. *We say that $T \in \mathcal{L}(E, E^*)$ is an asymptotic derivative of a mapping $f : E \to E^*$ along the set $\mathbb{K}$ if*

$$\lim_{\substack{\|x\| \to \infty \\ x \in \mathbb{K}}} \frac{\|f(x) - T(x)\|}{\|x\|} = 0.$$

About the applications of this notion in nonlinear analysis the reader is referred to [21, 25, 26]. Some methods for computation of asymptotic derivatives are given certainly in [25, 26] and also in [21]. We note that if $\mathbb{K}$ is a closed convex cone and $E = \mathbb{K} - \mathbb{K}$ we have that the asymptotic derivative (when this derivative exists) is unique. It is easy to show that if f has an asymptotic derivative T along $\mathbb{K}$, then T is also a scalar asymptotic derivative of f along the same set $\mathbb{K}$.

5 Scalar Compactness

Let $(E, \|\cdot\|)$ be a Banach space, E^* the topological dual of E, and $\langle E, E^* \rangle$ a pairing between E and E^*. We denote by J the duality mapping between E and E^*, that is, for any $x \in E$, $J(x) = \{f \in E^* \mid \langle x, f\rangle = \|x\|^2 = \|f\|\}$. It is known that if E^* is strictly convex, then $J(x)$ is a singleton for any $x \in E$ [8]. In this case we have $\langle x, J(x)\rangle = \|x\|^2$ for any $x \in E$ and J is a monotone mapping, i.e.,

$$\langle x - y, J(x) - J(y)\rangle \geq 0, \text{ for any } x, y \in E.$$

We can consider more general a duality mapping J associated with a weight ϕ [8].

Proposition 4. *If $\{x_n\}_{n\in N} \subset E$ and $\{y_n\}_{n\in N} \subset E^*$ are two sequences such that $\{x_n\}_{n\in N}$ is weakly convergent to an element $x_* \in E$ and $\{y_n\}_{n\in N}$ is convergent in norm to an element $y_* \in E$, then $\lim_{n\to\infty}\langle x_n, y_n\rangle = \langle x_*, y_*\rangle$.*

Proof. We have

$$\langle x_n, y_n\rangle - \langle x_*, y_*\rangle = \langle x_n - x_*, y_n - y_*\rangle + \langle x_*, y_n\rangle + \langle x_n, y_*\rangle - 2\langle x_*, y_*\rangle,$$

which implies

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x_* - y_*\rangle| \leq |\langle x_n - x_*, y_n - y_*\rangle| + |\langle x_*, y_n\rangle + \langle x_*, y_n\rangle + \langle x_n, y_*\rangle \\ -2\langle x_*, y_*\rangle| \leq \|x_n - x_*\|\|y_n - y_*\| + |\langle x_*, y_n - y_*\rangle + \langle x_n - x_*, y_*\rangle|. \end{aligned}$$

Because the sequence $\{x_n - x_*\}_{n\in N}$ is weakly convergent, there exists $M > 0$ such that $\|x_n - x_*\| \leq M$ and because $\{y_n\}_{n\in N}$ is convergent in norm to y_* it is weakly $(*)$-convergent to y_*. Therefore we have that, $\lim_{n\to\infty} |\langle x_n, y_n\rangle| - \langle x_*, y_*\rangle = 0$, that is, $\lim_{n\to\infty}\langle x_n, y_n\rangle = \langle x_*, y_*\rangle$.

Similarly, we have also the following result.

Proposition 5. *If $\{x_n\}_{n\in N} \subset E$ and $\{y_n\}_{n\in N} \subset E^*$ are two sequences such that $\{x_n\}_{n\in N}$ is convergent in norm to an element $x_* \in E$ and $\{y_n\}_{n\in N}$ is weakly $(*)$-convergent to an element $y_* \in E^*$, then $\lim_{n\to\infty}\langle x_n, y_n\rangle = \langle x_*, y_*\rangle$.*

Proof. As in the proof of Proposition 4 we have

$$\{x_n, y_n\} - \langle x_*, y_*\rangle = \langle x_n - x_*, y_n - y_*\rangle + \langle x_*, y_n\rangle + \langle x_n, y_*\rangle - 2\langle x_*, y_*\rangle,$$

which implies that there exists $M > 0$ such that

$$|\langle x_n, y_n\rangle - \langle x_*, y_*\rangle| \leq M\|x_n - x_*\| + \|y_*\|\|x_n - x_*\| + |\langle x_*, y_n - y_*\rangle|,$$

and computing the limit we obtain that

$$\lim_{n\to\infty}\langle x_n, y_n\rangle = \langle x_*, y_*\rangle.$$

The following definition is inspired by condition $(S)_+$.

In condition $(S)_+$ we have that if $\{x_n\}_{n\in N} \subset E \subseteq E$ is weakly convergent to an element $x_* \in E$ and $\lim_{n\to\infty}\sup\langle x_n - x_*, f(x_n)\rangle \leq 0$, then the sequence is convergent in norm to x_*. Related to this condition a natural question is, Under what conditions about the mapping f we have that if $\{x_n\}_{n\in N}$ is weakly convergent to $x_* \in E$ do we have that $\{x_n\}$ has a subsequence $\{x_{n_k}\}_{k\in K}$ such that $\lim_{n\to\infty}\sup\langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq 0$.?

We introduce the following notion. Let $D \subset E$ be a non-empty subset.

Definition 9. *We say that a mapping $f : D \to E^*$ is scalarly compact if for any sequence $\{x_n\}_{n\in N} \subset D$, weakly convergent to an element $x_* \in D$ there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that, $\lim_{n\to\infty}\sup\langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq 0$.*

In the next propositions we present several examples of scalarly compact operators.

Proposition 6. *If $f : E \to E^*$ is completely continuous, then f is scalarly compact.*

Proof. Let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$. Then $\{x_{n_k}\}_{k\in N}$ is bounded. Because f is completely continuous there exists a subsequence $\{x_{n_k}\}_{kinN}$ of the sequence $\{x_n\}_{n\in N}$ such that $\{f(x_{n_k})\}_{k\in N}$ is convergent in norm in E^* to an element $y_* \in E^*$. By Proposition 4 we have $\lim_{n\to\infty}\langle x_{n_k} - x_*, f(x_{n_k})\rangle = 0$, and hence $\lim_{k\to\infty}\sup\langle x_{n_k} - x_*, f(x_{n_k})\rangle = 0$.

In the next results we will see that there exist mappings which are scalarly compact but not completely continuous.

Proposition 7. *If $f : E \to E^*$ has a decomposition of the form $f = h - g$, where $h : E \to E^*$ is completely continuous and $g : E \to E^*$ is monotone, then f is scalarly compact.*

Proof. Let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$. Then $\{x_n\}_{n\in N}$ is bounded and because h is completely continuous, there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that $\{h(x_{n_k})\}_{k\in N}$ is convergent in norm to an element $y_* \in E^*$. We have

$$\begin{aligned}
\langle x_{n_k} - x_*, f(x_{n_k})\rangle &= \langle x_{n_k} - x_*, h(x_{n_k}) - g(x_{n_k})\rangle \\
&= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \langle x_{n_k} - x_*, g(x_{n_k})\rangle \\
&= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - [\langle x_{n_k} - x_*, g(x_{n_k}) - g(x_*)\rangle + \langle x_{n_k} - x_*, g(x_*)\rangle] \\
&= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \langle x_{n_k} - x_*, g(x_{n_k} - g(x_*))\rangle + \langle x_{n_k} - x_*, g(x_*)\rangle \\
&\leq \langle x_{n_k} - x_*, h(x_{n_k})\rangle + \langle x_{n_k} - x_*, g(x_*)\rangle.
\end{aligned}$$

Consider Proposition 4 and computing lim sup we obtain

$$\lim_{k\to\infty} \sup \langle x_{n_k} - x_*, h(x_{n_k}) - g(x_{n_k})\rangle \leq 0.$$

Corollary 1. *If $(E, \|\cdot\|)$ is a Banach space such that a duality mapping $J : E \to E^*$ is at any $x \in E$ a singleton and $h : E \to E^*$ is a completely continuous mapping, then the mapping $f = h - J$ is a scalarly compact mapping.*

Proof. Because the mapping J is monotone, we apply Proposition 7.

Corollary 2. *If $(H, \langle\cdot,\cdot\rangle)$ is a Hilbert space and $h : H \to H$ is a completely continuous mapping, then the mapping $f(x) = h(x) - x$, for any $x \in H$, is scalarly compact, but not completely continuous.*

We say that a mapping $f : E \to E^*$ is *antimonotone* if for any $x, y \in E$ we have $\langle x - y, f(x) - f(y)\rangle \leq 0$.

We recall that a mapping $h : E \to E^*$ is *strongly monotone* if $\langle x-y, h(x)-h(y)\rangle \geq \rho\|x - y\|^2$ for any $x, y \in E$, where $\rho > 0$.

Let E be a Hilbert space. If it is strongly monotone then the mapping $f(x) = \rho x - h(x)$ is antimonotone. If $(E, \|\cdot\|)$ is a Banach space, $J : E \to E^*$ is a duality mapping and $f : E \to E^*$ is such that $\langle x - y, f(x) - f(y)\rangle \geq \langle x-y, J(x)-J(y)\rangle$ for any $x, y \in E$, then the mapping $J - f$ is antimonotone.

More general if $f_1, f_2 : E \to E^*$ are two mappings such that

$$\langle x - y, f_1(x) - f_1(y)\rangle \geq \langle x - y, f_2(x) - f_2(y)\rangle \text{ for any } x, y \in E,$$

then the mapping $f_2 - f_1$ is antimonotone.

Proposition 8. *If $f : E \to E^*$ is an antimonotone mapping then f is scalarly compact.*

Proof. Let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$. Then we have

$$\begin{aligned}\langle x_n - x_*, f(x_n)\rangle &= \langle x_n - x_*, f(x_n) - f(x_*) + f(x_*)\rangle \\ &= \langle x_n - x_*, f(x_n) - f(x_*)\rangle + \langle x_n - x_*, f(x_*)\rangle \leq \langle x_n - x_*, f(x_*)\rangle.\end{aligned}$$

Computing the limit sup we have

$$\lim_{n\to\infty} \sup\langle x_n - x_*, f(x_n)\rangle \leq 0.$$

The scalarly compactness is a property which is invariant with respect to compact perturbations. In this sense we have the following result.

Proposition 9. *If $h : E \to E^*$ is a scalarly compact mapping and $g : E \to E^*$ is a completely continuous mapping then the mapping $f = h + g$ is scalarly compact.*

Proof. Let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$. Because h is scalarly compact and g is completely continuous, we can select a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that $\lim_{k\to\infty} \sup\langle x_{n_k} - x_*, h(x_{n_k})\rangle \leq 0$ and $\{g(x_{n_k})\}_{k\in N}$ is convergent in norm to an element $w \in E$. Using also Proposition 4, we have

$$\begin{aligned}&\lim_{k\to\infty} \sup\langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq \lim_{k\to\infty} \sup\langle x_{n_k} - x_*, h(x_{n_k})\rangle \\ &+ \lim_{k\to\infty} \sup\langle x_{n_k} - x_*, g(x_{n_k})\rangle \leq 0 + \langle 0, w\rangle = 0.\end{aligned}$$

Proposition 10. *If h and g are scalarly compact mappings from E into E^*, then for any positive real numbers a and b the mapping $f = ah + bg$ is scalarly compact.*

Proof. The proposition is a consequence of definition of scalar compactness.

Corollary 1. *If $f : E \to E^*$ is scalarly compact and $g : E \to E^*$ is anti-monotone then the mapping $f = h + g$ is scalarly compact.*

Proposition 11. *Let $D \subseteq E$ be a closed convex non-empty subset and $f, g : D \to E^*$ two mappings. If there exist two real numbers α, β such that $\langle x - y, f(x) - f(y)\rangle \leq \alpha$ and $\langle x - y, g(x) - g(y)\rangle \leq \beta$ for any x, $y \in D$ and $\alpha + \beta \leq 0$, then the mapping $f + g$ is scalarly compact.*

Proof. Indeed, if $\{x_n\} \subset D$ is a sequence weakly convergent to an element $x_* \in D$, then we have

$$\begin{aligned}&\limsup_{n\to\infty} \langle x_n - x_*, f(x_n) + g(x_n)\rangle \leq \limsup_{n\to\infty} \langle x_n - x_*, f(x_*)\rangle + \\ &+ \limsup_{n\to\infty} \langle x_n - x_*, f(x_n) - f(x_*)\rangle + \limsup_{n\to\infty} \langle x_n - x_*, g(x_n)\rangle + \\ &+ \limsup_{n\to\infty} \langle x_n - x_*, g(x_n) - g(x_*)\rangle \leq \alpha + \beta \leq 0.\end{aligned}$$

Proposition 12. *Let $D \subseteq E$ be a closed convex, non-empty subset. Let $h : D \to E^*$ be a scalarly compact mapping, $g : D \to E^*$ a monotone mapping, and $\rho : D \to \mathbb{R}_+$ a sequentially weakly continuous mapping. Then the mapping $f(x) = h(x) - \rho(x)g(x)$, for any $x \in D$ is scalarly compact.*

Proof. Let $\{x_n\}_{n\in N} \subset D$ be a sequence, weakly convergent to an element $x_* \in D$. Because h is scalarly compact, there exists a subsequence $\{x_{n_k}\}_k \in N$ of the sequence $\{x_n\}_{n\in N}$ such that $\limsup_{n\to\infty} \langle x_{n_k} - x_*, h(x_{n_k})\rangle \leq 0$. We have

$$\begin{aligned}
\langle x_{n_k} - x_*, f(x_{n_k})\rangle = \langle x_{n_k} - x_*, h(x_{n_k}) - \rho(x_{n_k})g(x_{n_k})\rangle = \\
= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \langle x_{n_k} - x_*, \rho(x_{n_k})g(x_{n_k})\rangle = \\
= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \rho(x_{n_k}) \langle x_{n_k} - x_*, g(x_{n_k})\rangle = \\
= \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \rho(x_{n_k}) \langle x_{n_k} - x_*, g(x_{n_k}) - g(x_*)\rangle - \\
-\rho(x_{n_k}) \langle x_{n_k} - x_*, g(x_*)\rangle \leq \\
\leq \langle x_{n_k} - x_*, h(x_{n_k})\rangle - \rho(x_{n_k}) \langle x_{n_k} - x_*, g(x_*)\rangle ,
\end{aligned}$$

which implies

$$\limsup_{n\to\infty} \langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq 0.$$

Remark 1. In *Proposition* 12 the set D can be a closed convex cone $\mathbb{K} \subset E$ and $\rho \in K^*$.

Proposition 13. *Let $D \subset E$ be a closed convex, non-empty subset and $f : D \to E^*$ a mapping. If there exists a completely continuous mapping $h : D \to E^*$ such that $\langle y, f(x)\rangle \leq \langle y, h(x)\rangle$ for any x, $y \in D$, then f is scalarly compact.*

Proof. Indeed, let $\{x_n\}_{n\in N} \subset D$ be a sequence, weakly convergent to an element $x_* \in D$. Because h is completely continuous there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that $\{h(x_{n_k})\}_{k\in N}$ is convergent in norm to an element $w \in E^*$. We have

$$\langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq \| \langle x_{n_k} - x_*, h(x_{n_k})\rangle \|$$

which implies,

$$\limsup_{k\to\infty} \langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq 0.$$

The following definition is inspired by condition $(S)'$.

We say that a mapping $f : E \to E^*$ is lim sup-antimonotone if for any sequence $\{x_n\}_{n\in N} \subset E$, weakly convergent to an element x_*, there exists a subsequence $\{x_{n_k}\}_{k\in N}$ such that

$$\limsup_{k\to\infty} \langle x_{n_k} - x_*, f(x_{n_k}) - f(x_*)\rangle \leq 0.$$

Proposition 14. *Any limsup-antimonotone mapping $f : E \to E^*$ is scalarly compact.*

Proof. Indeed, let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$. From our assumption, there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that

$$\langle x_{n_k} - x_*, f(x_{n_k})\rangle = \langle x_{n_k} - x_*, f(x_{n_k}) - f(x_*)\rangle + \langle x_{n_k} - x_*, f(x_*)\rangle\,,$$

which implies

$$\begin{aligned}\limsup_{k\to\infty} \langle x_{n_k} - x_*, f(x_{n_k})\rangle \leq\ & \limsup_{k\to\infty} \langle x_{n_k} - x_*, f(x_{n_k}) - f(x_*)\rangle + \\ & + \limsup_{k\to\infty} \langle x_{n_k} - x_*, f(x_*)\rangle\end{aligned}$$

The following results are consequences of the scalarly compactness.

Proposition 15. *If $f_1,\ f_2 : E \to E^*$ are two mappings such that*

(i) f_1 satisfies condition $(S)_+$,
(ii) f_2 is scalarly compact,

then $f_1 - f_2$ satisfies condition $(S)_+$.

Proof. Let $\{x_n\}_{n\in N} \subset E$ be a sequence weakly convergent to an element $x_* \in E$, such that $\limsup_{k\to\infty} \langle x_n - x_*, f_1(x_n) - f_2(x_n)\rangle \leq 0$. Because f_2 is scalarly compact there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence $\{x_n\}_{n\in N}$ such that $\limsup_{k\to\infty} \langle x_{n_k} - x_*, f_2(x_{n_k})\rangle \leq 0$. We can show that

$$\limsup_{k\to\infty} \langle x_{n_k} - x_*, f_1(x_{n_k}) - f_2(x_{n_k})\rangle \leq 0.$$

We have

$$\begin{aligned}\limsup_{k\to\infty} \langle x_{n_k} - x_*, f_1(x_{n_k})\rangle \leq\ & \limsup_{k\to\infty} \langle x_{n_k} - x_*, f_1(x_{n_k}) - f_2(x_{n_k})\rangle \\ & + \limsup_{k\to\infty} \langle x_{n_k} - x_*, f_2(x_{n_k})\rangle \leq 0.\end{aligned}$$

Using the fact that f_1 is scalarly compact we obtain that the subsequence $\{x_{n_k}\}_{k\in N}$ has another subsequence $\{x_{n_j}\}_{j\in N}$ convergent in norm to x_*. Therefore $f_1 - f_2$ satisfies condition $(S)_+$.

Corollary 3. *If $f_1,\ f_2 : E \to E^*$ are two mappings such that:*

(i) f_1 satisfies condition $(S)_+$,
(ii) $-f_2$ is scalarly compact,

then $f_1 + f_2$ satisfies condition $(S)_+$.

From Corollary 3 we also deduce the following interesting result.

Corollary 4. *If f_1, $f_2 : E \to E^*$ are two mappings such that*

(i) f_1 satisfies condition $(S)_+$,
(ii) f_2 is monotone,

then $f_1 + f_2$ satisfies condition $(S)_+$.

Proof. Because f_2 is monotone, we have that $-f_2$ is antimonotone, and hence $-f_2$ is scalarly compact and we can apply Corollary 3.

We recall the following definition due to Browder ([6], Definition 2). We say that a mapping $f : E \to E^*$ is pseudo-monotone if for any sequence $\{x_n\}_{n\in N} \subset E$ weakly convergent to an element $x_* \in E$ and satisfying the property

$$\limsup_{n\to\infty} \langle x_n - x_*, f(x_n)\rangle \leq 0$$

we have that $\limsup_{n\to\infty} \langle x_n - x_*, f(x_n)\rangle = 0$ is weakly convergent to $f(x*)$. From this definition we deduce immediately the following result.

Proposition 16. *Let $(E, \|\cdot\|)$ be a reflexive Banach space. If $f : E \to E^*$ is pseudo-monotone (in Browder's sense) and scalarly compact, then f has the following property: for any bounded sequence $\{x_n\}_{n\in N} \subset E$, there exists a subsequence $\{x_{n_k}\}_{k\in N}$ of the sequence, such that $\{f(x_{n_k})\}$ is weakly convergent, i.e., f is sequentially weakly compact.*

6 Existence Theorems for Variational Inequalities and Complementarity Problems

We present in this section some existence theorems for variational inequalities and complementarity problems.

Theorem 2. *Let $(E, \|\cdot\|)$ be a reflexive Banach space and T_1, $T_2 : E \to E^*$ two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition (S^1_+),*
2. *T_2 is scalarly compact,*

then for every non-empty bounded convex set $D \subset E$ the problem $\mathrm{VI}(T_1 - T_2, D)$ has a solution.

Proof. Let Λ be the family of all finite dimensional subspaces F of E such that $F \cap D$ is non-empty. Denote by

$$h(x) = T_1(x) - T_2(x) \text{ for all } x \in D \text{ and } D(F) = D \cap F \text{ for each } F \in \Lambda.$$

For each $F \in \Lambda$ we set

$$A_F = \{y \in D \mid \langle x - y, h(y)\rangle \geq 0 \text{ for all } x \in D(F)\}.$$

For each $F \in \Lambda$ the set A_F is non-empty.

Indeed, the solution set of the problem VI$(h, D(F))$ is a subset of A_F. The solution set of the problem VI$(h, D(F))$ is non-empty because of the following reason. Let $j : F \to E$ denote the inclusion and $j^* : E^* \to F^*$ the adjoint of j^*.

By our assumptions we have that $j^* \circ h \circ j : D(F) \to F^*$ is continuous and

$$\langle x - y, (j^* \circ h \circ j)(y)\rangle = \langle j(x - y), (h \circ j)(y)\rangle = \langle (x - y), h(y)\rangle$$

for all $x, y \in D(F)$. Applying the classical Hartman–Stampacchia theorem [13] to the mapping $j^* \circ h \circ j$ and the set $D(F)$ we obtain that the problem VI$(h, D(F))$ has a solution. Denote by $\bar{A}_F^\sigma$ the weak closure of A_F. We have that $\cap_{F\in\Lambda}\bar{A}_F^\sigma$ is non-empty. Indeed, let $\bar{A}_{F_1}^\sigma, \bar{A}_{F_2}^\sigma, \ldots, \bar{A}_{F_n}^\sigma$ be a finite subfamily of the family $\{\bar{A}_F^\sigma\}_{F\in\Lambda}$. Let F_0 be the finite dimensional subspace of E generated by $F_1, F_2, \ldots, F_n$. Because $F_k \subset F_0$ for all $k = 1, 2, \ldots, n$, we have that $D(F_k) \subseteq D(F_0)$ for all $k = 1, 2, \ldots, n$. We have $A_{F_0} \subseteq A_{F_k}$, which implies $\bar{A}_{F_0}^\sigma \subseteq \bar{A}_{F_k}^\sigma$ for $k = 1, 2, \ldots, n$, and finally we have, that is, $\cap_{k=1}^n \bar{A}_F^\sigma$ non-empty. Since D is a weakly compact set, we conclude that $\cap_{F\in\Lambda}\bar{A}_F^\sigma$ is non-empty. Let $y_* \in \cap_{F\in\Lambda}\bar{A}_F^\sigma$, i.e., for every $F \in \Lambda$ we have $y_* \in \bar{A}_F^\sigma$. Let $x \in D$ be an arbitrary element. There exists some $F \in \Lambda$ such that $x, y_* \in F$. Since $y_* \in \bar{A}_F^\sigma$, by Eberlein–Smulian theorem there exists a sequence $\{y_n\}_{n\in N} \subset A_F$ weakly convergent to y_*. We have

$$\begin{cases} \langle y_* - y_n, h(y_n)\rangle \geq 0 \\ \text{and} \\ \langle x - y_n, h(y_n)\rangle \geq 0 \end{cases}$$

or

$$\langle y_* - y_n, T_1(y_n)\rangle \leq \langle y_* - y_n, T_2(y_n)\rangle \tag{1}$$

and

$$\langle y_* - y_n, T_1(y_n)\rangle \leq \langle y_* - y_n, T_2(y_n)\rangle . \tag{2}$$

Using (1) and the fact that T_2 is scalarly compact we have that $\{y_n\}_{n\in N}$ has a subsequence, denoted again by $\{y_n\}_{n\in N}$ such that

$$\limsup_{n\to\infty} \langle y_n - y_*, T_1(y_n)\rangle \leq 0. \tag{3}$$

Because T_1 is bounded, we can suppose (taking eventually a subsequence of $\{y_n\}_{n\in N}$) that $\{T_1(y_n)\}_{n\in N}$ is weakly $(*)$-convergent to an element $v_0 \in E*$. Because

$$\langle y_n, T_1(y_n)\rangle = \langle y_n - y_* + y_*, T_1(y_n)\rangle = \langle y_n - y_*, T_1(y_n)\rangle + \langle y_*, T_1(y_n)\rangle$$

and considering formula (3) we obtain

$$\limsup_{n\to\infty} \langle y_n, T_1(y_n)\rangle \leq \langle y_*, \nu_0\rangle .$$

Hence, by condition $(S)^1_+$ (we obtain that the sequence $\{y_n\}_{n\in N}$ has a subsequence denoted again by $\{y_n\}_{n\in N}$ convergent in norm to y_*. Then the sequence $\{x-y_n\}_{n\in N}$ is convergent in norm to $x-y_*$. Considering Proposition 5 and formula (2) we obtain

$$\langle x-y_*, T_1(y_*)-T_2(y_*)\rangle \geq 0$$

for any $x \in D$ and the proof is complete. □

From Theorem 2 we deduce the following result.

Corollary 5. *Let $(E, \|\cdot\|)$ be a reflexive Banach space and $T_1, T_2 : E \to E^*$ two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is monotone,*

then for every non-empty bounded closed convex set $D \subset E$ the problem VI (T_1+T_2, D) *has a solution.*

Remark 2. If $(E, \|\cdot\|)$ is a reflexive Banach space, then for any mapping $f : E \to E^*$ which is a perturbation of demicontinuous monotone mapping by a bounded demicontinuous mapping which satisfies condition $(S)^1_+$ (*i.e.*, $f = T_1 - T_2$, where T_1 and T_2 are as in Corollary 5) then the problem VI (f, D) has a solution.

As application of Theorem 2 we have the following existence theorem for variational inequalities.

Theorem 3. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, $\mathbb{K} \subset E$ an unbounded closed convex set such that $0 \in K$. Let $T_1, T_2 : E \to E^*$ be two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is scalarly compact,*
3. *there exist $r > 0$ and $c > 0$ such that $c\|x\| \leq \langle x, T_1(x)\rangle$ for all $x \in \mathbb{K}$, with $\|x\| > r$,*
4. *T_2 has a scalar asymptotic derivative $T^\infty_{2,s}$, such that $\|T^\infty_{2,s}\| < c$,*

then the problem VI $(T_1 - T_2, \mathbb{K})$ *has a solution.*

Proof. For every $n \in N$ we denote by

$$\mathbb{K} = \{x \in \mathbb{K} \,|\, \|x\| \leq n\}.$$

Obviously, $\mathbb{K} = \cup_{n=1}^\infty \mathbb{K}_n$ and we observe that for each $n \in N$, $\mathbb{K}_n$ is a bounded closed convex set. By Theorem 2 the problem VI$(T_1 - T_2, \mathbb{K}_n)$ has a solution $y_n \in \mathbb{K}_n$ for every $n \in N$. Hence, we have

$$\langle x-y_n, (T_1-T_2)(y_n)\rangle \geq 0 \quad \text{for all} x \in \mathbb{K}_n. \tag{4}$$

If in (4) we put $x = 0$ we obtain

$$\langle y_n, T_1(y_n)\rangle \leq \langle y_n, T_2(y_n)\rangle . \tag{5}$$

The sequence $\{y_n\}_{n\in N}$ is bounded. Indeed, if we suppose that $\|y_n\| \to +\infty$ as $n \to \infty$, then by assumptions (3) and (4) we have (supposing that $\|y_n\| \neq 0$, for all $n \in N$) that $\|y_n\| > r$, for all $n \in N$, and

$$c \leq \frac{\langle y_n, T_1(y_n)\rangle}{\|y_n\|^2} \leq \frac{\langle y_n, T_1(y_n)\rangle}{\|y_n\|^2} = \frac{\langle y_n, T_2(y_n) - T^\infty_{2,s}(y_n)\rangle}{\|y_n\|^2} + \frac{\langle y_n, T^\infty_{2,s}(y_n)\rangle}{\|y_n\|^2},$$

which implies

$$\limsup_{n\to\infty} \frac{\langle y_n, T_1(y_n)\rangle}{\|y_n\|^2} \leq \limsup_{n\to\infty} \frac{\langle y_n, T^\infty_{2,s}(y_n)\rangle}{\|y_n\|^2} \leq T^\infty_{2,s}(y_n) < c,$$

which is a contradiction. Hence the sequence $\{y_n\}_{n\in N}$ is bounded. By the reflexivity of E, by the fact that $\mathbb{K}$ is a weakly closed set and using the *Eberlein–Šmulian* theorem, there exists a subsequence of the sequence $\{y_n\}_{n\in N}$, denoted again by $\{y_n\}_{n\in N}$ weakly convergent to an element $y_* \in \mathbb{K}$. Since T_1 is bounded and considering eventually again a subsequence (and Eberlein–Šmulian theorem) we can suppose that $\{T_1(y_n)\}_{n\in N}$ is weakly $(*)$-convergent in E^* to an element $u \in E^*$. Let $x \in \mathbb{K}$ be an arbitrary element. There exists $n_0 \in N$ such that $\{y_*, x\} \subset \mathbb{K}_{n_0}$ and obviously $\{y_*, x\} \subset \mathbb{K}_n$, for any $n \geq n_0$. Considering formula (4) we deduce

$$\langle y_* - y_n, (T_1 - T_2)(y_n)\rangle \geq 0$$

or

$$\langle y_n - y_*, (T_1 - T_2)(y_n)\rangle \leq 0 \tag{6}$$

and

$$\langle x - y_n, (T_1 - T_2)(y_n)\rangle \leq 0. \tag{7}$$

From (6) we deduce

$$\langle y_n - y_*, T_1(y_n)\rangle \leq \langle y_n - y_*, T_2(y_n)\rangle$$

and using the fact that T_2 is scalarly compact, we deduce that there exists a subsequence $\{y_{n_k}\}_{k\in N}$ of $\{y_n\}_{n\in N}$ such that

$$\limsup_{k\to\infty} \langle y_{n_k} - y_*, T_1(y_{n_k})\rangle \leq 0.$$

From this inequality and the following equality

$$\langle y_{n_k}, T_1(y_{n_k})\rangle = \langle y_{n_k} - y_*, T_1(y_{n_k})\rangle + \langle y_*, T_1(y_{n_k})\rangle$$

we deduce

$$\limsup_{k\to\infty} \langle y_{n_k}, T_1(y_{n_k})\rangle \leq \langle y_*, u\rangle .$$

Using the fact that T_1 satisfies condition $(S)^1_+$ we obtain that $\{y_{n_k}\}_{k\in N}$ contains a subsequence denoted again by $\{y_{n_k}\}_{k\in N}$ convergent in norm to an element which must be y_*. Now computing the limit in (7) (using the demicontinuity of T_1 and T_2 and Proposition 5) we obtain

$$\langle x - y_*, (T_1 - T_2)(y_*)\rangle \geq 0$$

for all $x \in \mathbb{K}$ and the proof is complete.

Corollary 2. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, $\mathbb{K} \subset E$ an unbounded closed convex set such that $0 \subset \mathbb{K}$. Let $T_1, T_2 : E \to E^*$ be two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is monotone,*
3. *there exist $r > 0$ and $c > 0$ such that, $c\|x\|^2 \leq \langle x, T_1(x)\rangle$ for all $x \in \mathbb{K}$ with $\|x\| > r$,*
4. *$-T_2$ has a scalar asymptotic derivative $(-T_2)^\infty_s$ such that $\|(-T_2)^\infty_s\| < c$,*

then the problem has a solution.

Corollary 3. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, $\mathbb{K} \subset E$ a closed convex cone. Let T_1, $T_2 : E \to E^*$ be two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is scalarly compact,*
3. *there exist $r > 0$ and $c > 0$ such that with $c\|x\|^2 \leq \langle x, T_1(x)\rangle$ for all $x \in \mathbb{K}$ with $\|x\| > r$,*
4. *T_2 has a scalar asymptotic derivative $T^\infty_{2,s}$ such that, $\|T^\infty_{2,s}\| < c$,*

then the problem $NCP(T_1 - T_2, \mathbb{K})$ has a solution.

Definition 10. *We say that the mapping $T_2 : E \to E^*$ satisfies Altmans condition on the set $K \subset E$ with respect to the mapping $T_1 : E \to E^*$ if there exists $r > 0$ such that $\langle x, T_2(x)\rangle \leq \langle x, T_1(x)\rangle$ for any $x \in \mathbb{K}$ with $\|x\| = r$. (The set $\mathbb{K}$ is as in Theorem 2.)*

We have the following result.

Theorem 4. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, $\mathbb{K} \subset E$ a closed convex cone, and T_1, $T_2 : E \to E^*$ be two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is scalarly compact,*
3. *T_2 satisfies Altmans condition with respect to T_1 for some $r > 0$, then the problem $NCP(T_1 - T_2, \mathbb{K})$ has a solution.*

Proof. Consider the set $\mathbb{K}_r = \{x \in \mathbb{K} | \|x\| \leq r\}$, where r is given by assumption (3). Obviously, $\mathbb{K}_r$ is a bounded closed convex set in E. By Theorem 2 we obtain an element $x_* \in \mathbb{K}_r$ such that

$$\langle x - x_*, (T_1 - T_2)(x_*)\rangle \geq 0 \quad \text{for all } x \in \mathbb{K}_r. \tag{8}$$

Taking in (8) $x = 0$ we have

$$\langle x_*, T_1(x_*)\rangle \leq \langle x_*, T_2(x_*)\rangle . \tag{9}$$

Now, we prove the following inequality:

$$\langle x_*, T_1(x_*)\rangle \geq \langle x_*, T_2(x_*)\rangle . \tag{10}$$

We have only two possibilities:

(I) $\|x_*\| = r$. In this case (10) is true by assumption (3) (Altmans condition).

(II) $\|x_*\| < r$. In this case there exists $\lambda_* > 1$ such that $x = \lambda_* x_* \in \mathbb{K}_r$. Taking $x = \lambda_* x_*$ in (8) we obtain that (10) is true. Hence, let $x \in \mathbb{K}$ be an arbitrary element. There exists $\lambda > 0$ such that $\lambda x \in \mathbb{K}_r$ and from (8) we have

$$\langle \lambda x - x_*, (T_1 - T_2)(x_*)\rangle \geq 0,$$

which implies

$$\begin{aligned} 0 \leq \langle \lambda x - x_*, (T_1 - T_2)(x_*)\rangle &= \langle \lambda x - [\lambda x_* + (1-\lambda)x_*], (T_1 - T_2)(x_*)\rangle \\ &= \lambda \langle x - x_*, (T_1 - T_2)(x_*)\rangle . \end{aligned}$$

Therefore, for all $x \in \mathbb{K}$ we have

$$\langle x - x_*, (T_1 - T_2)(x_*)\rangle \geq 0 \quad \text{for all } x \in \mathbb{K},$$

which implies that the problem NCP$(T_1 - T_2, \mathbb{K})$ has a solution.

As an application of Theorem 4 we consider the problem NCP $(T_1 - \lambda T_2, \mathbb{K})$, where λ is a positive real number. This is a complementarity problem with eigenvalues. To study this problem we need to introduce the following condition.

Definition 11. *We say that the mapping T_1, T_2 satisfy condition (C) if there exists $r > 0$ such that*

$$\inf\{\langle x, T_1(x)\rangle \,|x \in \mathbb{K} \text{ and } \|x\| = r\} = \rho_1 > 0$$

and

$$\sup\{\langle x, T_1(x)\rangle \,|x \in \mathbb{K} \text{ and } \|x\| = r\} = \rho_2 > 0.$$

We have the following result.

Theorem 5. *Let $(E, \|\cdot\|)$ be a reflexive Banach space, $\mathbb{K} \subset E$ a closed pointed convex cone, and T_1, $T_2 : E \to E^*$ two demicontinuous mappings. If the following assumptions are satisfied:*

1. *T_1 is bounded and satisfies condition $(S)^1_+$,*
2. *T_2 is scalarly compact,*
3. *T_1, T_2 satisfy condition (C),*

then for any λ such that $0 < \lambda < \rho_1/\rho_2$ the problem NCP $(T_1 - \lambda T_2, \mathbb{K})$ *has a solution which is not the trivial solution if $T_1(0) - T_2(0) \notin \mathbb{K}^*$.*

Proof. We observe that the assumptions of Theorem 4 are satisfied.

7 Comments

We presented in this chapter the notion of *scalarly compact mapping*. We introduced this notion analyzing the condition $(S)_+$, well known in nonlinear analysis. The main results presented in this chapter are strongly based on the notion of scalarly compact mapping. New developments of the results presented in this chapter are possible.

References

1. Baiocchi, C., Capelo, A.: Variational and Quasivariational Inequalities. Applications to Free-Boundary Problems, Wiley, New York, NY (1984)
2. Brezis, H.: Équations et inéquations non linéaires dans les espaces vectoriels en dualité. Ann. Inst. Fourier 18, 115–175 (1968)
3. Browder, F.E.: Nonlinear eigenvalues problems and Galerkin approximations. Bull. Amer. Math. Soc. 74, 651–656 (1968)
4. Browder, F.E.: Existence theorems for nonlinear partial differential equations. Proc. Sympos. Pure Math., Am. Math. Soc, Providence RI 16, 1–60 (1970)
5. Browder, F.E.: Nonlinear operators and nonlinear equations of evolution in Banach spaces. Proc. Symp. Pure Math., Am. Math. Soc, Providence RI 18(2), 269–286 (1976)
6. Browder, F.E.: Fixed point theory and nonlinear problems. Bull. Am. Math. Soc. 9, 1–39 (1983)
7. Chiang, Y.: The $(S)^1_+$ condition for generalized vector variational inequalities. J. Optim. Theory Appl. 124(3), 581–594 (2005)
8. Ciorănescu, I.: Geometry of Banach Spaces. Duality Mappings and Nonlinear Problems, Mathematics and Its Applications 62, Kluwer Academic Publishers, Dordrecht, Netherlands (1990)
9. Cubiotti, P.: General nonlinear variational inequalities with $(S)^1_+$ operators. Applied Mathematics Letters, 10(2), 11–15 (1997)
10. Cubiotti, P., Yao, J.C.: Multivalued $(S)^1_+$ operators and generalized variational inequalities. Comput. Math. Appl. 29(12), 40–56 (1995)

11. Guo, J.S., Yao, J.C.: Variational inequalities with nonmonotone operators. J. Optim. Theory Appl. 80(1), 63–74 (1994)
12. Hyers, D.H., Isac, G., Rassias, Th. M.: Topics in Nonlinear Analysis and Applications, World Scientific Publishing Company, Singapore, New Jersey, London (1997)
13. Isac, G.: Complementarity Problems. Number 1528 in Lecture Notes in Mathematics. Springer Verlag, Berlin, New York, NY (1992)
14. Isac, G.: Nonlinear complementarity problem and Galerkin method. J. Math. Anal. Appl. 108, 563–574 (1995)
15. Isac, G.: On an Altman type fixed-point theorem on convex cones. Rocky Mountain J. Math. 25(2), 701–714 (1995)
16. Isac, G.: The scalar asymptotic derivative and the fixed-point theory on cones. Nonlinear Anal. Related Topics 2, 92–97 (1999)
17. Isac, G.: Topological Methods in Complementarity Theory, Springer-Verlag New York, Inc., Secaucus, NJ (2000)
18. Isac, G.: Leray-Schauder Type Alternatives, Complementarity Problems and Variational Inequalities, Kluwer Academic Publishers, Netherlands (2006)
19. Isac, G., Bulavsky, V.V., Kalashnikov, V.V.: Complementarity, Equilibrium, Efficiency and Economics, Kluwer, Dordrecht (2002)
20. Isac, G., Gowda, M.S.: Operators of class $(S)^1_+$, Altmans conditions and the complementarity problem. J. Fac. Sci. The Univ. Tokyo Sec. IA 40(1), 1–16 (1993)
21. Isac, G., Németh, S.: Scalar Derivative and Scalar Asymptotic Derivatives, Theory and Applications, Springer, New York, NY (2008)
22. Isac, G., Théra, M.: A Variational Principle Application to the Nonlinear Complementarity Problem. Nonlinear and Convex Analysis (Eds. B. L. Lin and S. Simons), Marcel Dekker, Inc., New York, NY, pages 127–145, (1987)
23. Isac, G., Théra, M.: Complementarity problem and the existence of the post critical equilibrium state of a thin elastic plate. J. Optim. Theory Appl. 58, 241–257 (1988)
24. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and their Applications, Academic Press, New York, NY (1980)
25. Krasnoselskii, M.A.: Topological Methods in the Theory of Nonlinear Integral Equations, Gostekhizdat, Moscow (1956)
26. Krasnoselskii, M.A.: Positive Solutions of Operators Equations, Noordhoff, Groningen (1964)
27. Showalter, R.E.: Monotone Operators in Banach Spaces and Nonlinear Partial Differential Equations, Mathematical Surveys and Monographs (Vol. 49), American Mathematical Society, Providence, RI, (1997)
28. Skrypnik, I.V.: Methods for Analysis of Nonlinear Elliptic Boundary Value Problems, Translations of Mathematical Monographs (Vol. 139), American Mathematical Society, Providence, RI (1994)
29. Takahashi, W.: Nonlinear Functional Analysis (Fixed Point Theory and its Applications). Yokohama Publishers, Yokohama, Japan (2000)
30. Zeidler, E.: Nonlinear Functional Analysis and its Applications, volume II B of Nonlinear Monotone Operators. Springer, New York, NY (1990)

Quasi-equilibrium Inclusion Problems of the Blum–Oettli-Type and Related Problems

Nguyen Xuan Tan[1] and Lai-Jiu Lin[2]

[1] Institute of Mathematics, Hanoi, Vietnam nxtan@math.ac.vn
[2] Department of Mathematics, National Changhua University of Education, Changhua, Taiwan maljlin@math.ncue.edu.tw

Summary. The quasi-equilibrium inclusion problems of Blum–Oettli type are formulated and sufficient conditions on the existence of solutions are shown. As special cases, we obtain several results on the existence of solutions of general vector ideal (resp. proper, Pareto, weak) quasi-optimization problems, of quasivariational inequalities, and of quasivariational inclusion problems.

Key words: upper and lower quasivariational inclusions, inclusions, α-quasi-optimization problems, vector optimization problem, quasi-equilibrium problems, upper and lower C-quasiconvex multivalued mappings, upper and lower C-continuous multivalued mappings

1 Introduction

Let Y be a topological vector space and let $C \subset Y$ be a cone. We put $l(C) = C \cap (-C)$. If $l(C) = \{0\}$, then C is said to be a pointed cone. For a given subset $A \subset Y$, one can define efficient points of A with respect to C in different senses as ideal, Pareto, proper, weak, etc. (see [6]). The set of these efficient points is denoted by $\alpha\mathrm{Min}(A/C)$ with $\alpha = I, \alpha = P, \alpha = \mathrm{Pr}, \alpha = \mathrm{w}$, etc., for the case of ideal, Pareto, proper, weak efficient points, respectively. Let D be a subset of another topological vector space X. By 2^D we denote the family of all subsets in D. For a given multivalued mapping $f : D \to 2^Y$, we consider the problem of finding $\bar{x} \in D$ such that

$$f(\bar{x}) \cap \alpha\mathrm{Min}(f(D)/C) \neq \emptyset. \qquad (GVOP)_\alpha$$

This is called a general vector α optimization problem corresponding to D, f, and C. The set of such points $\bar{x}$ is said to be a solution set of $(GVOP)_\alpha$. The elements of $\alpha\mathrm{Min}(f(D)/C)$ are called α optimal values of $(GVOP)_\alpha$.

Now, let X, Y, and Z be topological vector spaces; let $D \subset X, K \subset Z$ be nonempty subsets; and let $C \subset Y$ be a cone. Given the following multivalued mappings

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_5,

$$S : D \to 2^D,$$
$$P : D \to 2^K, T : D \times D \to 2^K,$$
$$F : K \times D \times D \to 2^Y,$$

we are interested in the problem of finding $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, \bar{x}) \cap \alpha\text{Min}(F(y, \bar{x}, S(\bar{x}))/C) \neq \emptyset \quad \text{for all } y \in P(\bar{x}).$$

This is called a general vector α quasi-optimization problem depending on a parameter (α is, respectively, one of qualifications: ideal, Pareto, proper, weak). Such a point $\bar{x}$ is said to be a solution of $(GVQOP)_\alpha$. The above multivalued mappings S, P, and F are said to be, respectively, a constraint, a parameter potential, and an utility mapping. These problems also play a central role in the vector optimization theory concerning multivalued mappings and have many relations to the following problems:

(UIQEP), upper ideal quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \subset C \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(LIQEP), lower ideal quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \cap C \neq \emptyset \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(UPQEP), upper Pareto quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \not\subset -(C \setminus l(C)) \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(LPQEP), lower Pareto quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \cap -(C \setminus l(C)) = \emptyset \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(UWQEP), upper weakly quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \not\subset -\text{int}(C) \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(LWQEP), lower weakly quasi-equilibrium problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \cap \text{-int}(C) = \emptyset \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

In general, we call the above problems by γ quasi-equilibrium problems involving D, K, S, T, F with respect to C, where γ is one of the following qualifications: upper ideal, lower ideal, upper Pareto, lower Pareto, upper weakly, lower weakly. These problems generalize many well-known problems in the optimization theory as quasi-equilibrium problems, quasivariational inequalities, fixed point problems, complementarity problems, saddle point problems, minimax problems, as well as different others which have been studied by many authors, for example, Park [11], Chan and Pang [2], Parida and Sen [10], Gurraggio and Tan [4] for quasi-equilibrium problems and quasivariational problems, Blum and Oettli [1], Lin, Yu, and Kassay [5], Tan [12], Minh and Tan [8], Fan [3] for equilibrium and variational inequality problems and by some others in the references therein. One can easily see that the above problems also have many relations with the following quasivariational inclusion problems which have been considered in Tan [12], Luc and Tan [7], and Minh and Tan [8].

(UQVIP), upper quasivariational inclusion problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, x) \subset F(y, \bar{x}, \bar{x}) + C \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

(LQVIP), lower quasivariational inclusion problem. Find $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$F(y, \bar{x}, \bar{x}) \subset F(y, \bar{x}, x) - C \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

The purpose of this chapter is to give some sufficient conditions on the existence of solutions to the above γ quasi-equilibrium problems involving D, K, S, T, F with respect to $(-C)$, where F is of the form $F(y, x, x') = G(y, x', x) - H(y, x, x')$ with $G, H : K \times D \times D \to 2^Y$ being two different multivalued mappings. We also call them quasi-equilibrium problems of the Blum–Oettli type.

2 Preliminaries and definitions

Throughout this chapter, we denote by X, Y, and Z real Hausdorff topological vector spaces. The space of real numbers is denoted by R. Given a subset $D \subset X$, we consider a multivalued mapping $F : D \to 2^Y$. The effective domain of F is denoted by

$$\mathrm{dom}F = \{x \in D / F(x) \neq \emptyset\}.$$

Further, let Y be a topological vector space with a cone C. We introduce new definitions of C-continuities.

Definition 1. *Let $F : D \to 2^Y$ be a multivalued mapping.*

(i) F is said to be upper (resp. lower) C-continuous at $\bar{x} \in$ dom F if for any neighborhood V of the origin in Y there is a neighborhood U of $\bar{x}$ such that

$$F(x) \subset F(\bar{x}) + V + C$$
$$(F(\bar{x}) \subset F(x) + V - C, \text{respectively})$$

holds for all $x \in U \cap \text{dom}F$.

(ii) If F is simultaneously upper C-continuous and lower C-continuous at $\bar{x}$, then we say that it is C-continuous at $\bar{x}$.

(iii) If F is upper, lower,..., C-continuous at any point of domF, we say that it is upper, lower,..., C-continuous on D.

(iv) In the case $C = \{0\}$ in Y, we shall only say F is upper, lower continuous instead of upper, lower 0-continuous. The mapping F is continuous if it is simultaneously upper and lower continuous.

Definition 2. *Let $F : D \times D \to 2^Y$ be a multivalued mapping with nonempty values. We say that*

(i) F is upper C-monotone if

$$F(x, y) \subset -F(y, x) - C$$

holds for all $x, y \in D$.

(ii) F is lower C-monotone if for any $x, y \in D$ we have

$$(F(x, y) + F(y, x)) \cap (-C) \neq \emptyset.$$

Definition 3. *Let $F : K \times D \times D \to 2^Y, T : D \times D \to 2^K$ be multivalued mappings with nonempty values. We say that*

(i) F is diagonally upper (T, C)-quasiconvex in the third variable on D if for any finite $x_i \in D, t_i \in [0, 1], i = 1, ..., n, \sum_{i=1}^n t_i = 1, x_t = \sum_{i=1}^n t_i x_i$, there exists j=1,2,...,n such that

$$F(y, x_t, x_j) \subset F(y, x_t, x_t) + C \quad \text{for all} \quad y \in T(x_t, x_j).$$

(ii) F is diagonally lower (T, C)-quasiconvex in the third variable on D if for any finite $x_i \in D, t_i \in [0, 1], i = 1, ..., n, \sum_{i=1}^n t_i = 1, x_t = \sum_{i=1}^n t_i x_i$, there exists j=1,2,...,n such that

$$F(y, x_t, x_t) \subset F(y, x_t, x_j) - C \quad \text{for all} \quad y \in T(x_t, x_j).$$

To prove the main results we shall need the following theorem:

Theorem 1. *Let D be a nonempty convex compact subset of X and $F : D \to 2^D$ be a multivalued mapping satisfying the following conditions:*

1. *for all $x \in D, x \notin F(x)$ and $F(x)$ is convex;*
2. *for all $y \in D, F^{-1}(y)$ is open in D.*

Then there exists $\bar{x} \in D$ such that $F(\bar{x}) = \emptyset$.

3 Main Results

Let $D \subset X$, $K \subset Z$ be nonempty convex compact subsets, $C \subset Y$ be a convex closed pointed cone. We assume implicitly that multivalued mappings S, T and G, H are as in Introduction. In the sequel, we always suppose that the multivalued mapping S has nonempty convex values and $S^{-1}(x)$ is open for any $x \in D$. We have

Theorem 2. *Assume that*

1. *for any $x' \in D$, the set*

$$A_1(x') = \{x \in D| \quad (G(y,x,x') - H(y,x',x)) \not\subset -C \text{ for some } y \in T(x,x')\}$$

   is open in D;
2. *the multivalued mapping $G + H$ is diagonally upper (T,C)-quasiconvex in the third variable;*
3. *for any fixed $y \in K$, the multivalued mapping $G(y,.,.) : D \times D \to 2^Y$ is upper C-monotone;*
4. *$(G(y,x,x) + H(y,x,x)) \subset C$ for all $(y,x) \in K \times D$.*

Then there exists $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$(G(y,x,\bar{x}) - H(y,\bar{x},x)) \subset -C \quad \text{for all} \quad x \in S(\bar{x}), \quad y \in T(\bar{x},x).$$

Proof. We define the multivalued mapping $M_1 : D \to 2^D$ by

$$M_1(x) = \{x' \in D| \quad (G(y,x',x) - H(y,x,x')) \not\subset -C \quad \text{for some} \quad y \in T(x,x')\}.$$

Observe that if for some $\bar{x} \in D, \bar{x} \in S(\bar{x})$, one has $M_1(\bar{x}) \cap S(\bar{x}) = \emptyset$, then

$$(G(y,x,\bar{x}) - H(y,\bar{x},x)) \subset -C \quad \text{for all} \quad x \in S(\bar{x}), \quad y \in T(\bar{x},x)$$

and hence the proof is completed. Thus, our aim is to show the existence of such a point $\bar{x}$. Consider the multivalued mapping Q from D to itself defined by

$$Q(x) = \begin{cases} \text{co}M_1(x) \cap S(x) & \text{if } x \in S(x), \\ S(x) & \text{otherwise,} \end{cases}$$

where the multivalued mapping $\text{co}M_1 : D \to 2^D$ is defined by $\text{co}M_1(x) = \text{co}(M_1(x))$ with $\text{co}(B)$ denoting the convex hull of the set B. We now show that Q satisfies all conditions in step 4 of Theorem 2. It is easy to see that for any $x \in D, Q(x)$ is convex and

$$\begin{aligned} Q^{-1}(x) &= [(\text{co}M_1)^{-1}(x) \cap S^{-1}(x)] \cup [S^{-1}(x) \setminus \{x\}] \\ &= [\text{co}A_1(x) \cap S^{-1}(x)] \cup [S^{-1}(x) \setminus \{x\}] \end{aligned}$$

is open in D.

Further, we claim that $x \notin Q(x)$ for all $x \in D$. Indeed, suppose to the contrary that there exists a point $\bar{x} \in D$ such that $\bar{x} \in Q(\bar{x}) = \mathrm{co}M_1(\bar{x}) \cap S(\bar{x})$. In particular, $\bar{x} \in \mathrm{co}M_1(\bar{x})$, we then conclude that there exist $x_1, ..., x_n \in M_1(\bar{x})$ such that $\bar{x} = \sum_{i=1}^{n} t_i x_i, x_i \in M_1(\bar{x}), t_i \geq 0, \sum_{i=1}^{n} t_i = 1$. By the definition of M_1 we can see that

$$(G(y_i, x_i, \bar{x}) - H(y_i, \bar{x}, x_i)) \not\subset -C \text{ for some } y_i \in T(\bar{x}, x_i) \text{ and for all } i = 1, \ldots, n. \tag{1}$$

Since the multivalued mapping $G + H$ is diagonally upper (T, C)-quasiconvex in the third variable, there exists $j \in \{1, ..., n\}$ such that

$$G(y, \bar{x}, x_j) + H(y, \bar{x}, x_j) \subset C + G(y, \bar{x}, \bar{x}) + H(y, \bar{x}, \bar{x}) \subset C \text{ for all } y \in T(\bar{x}, x_j). \tag{2}$$

Since G is upper C-monotone, we deduce

$$G(y, x_j, \bar{x}) \subset (-C - G(y, \bar{x}, x_j)) \quad \text{for } y \in T(\bar{x}, x_j). \tag{3}$$

A combination of (2) and (3) gives

$$\begin{aligned}(G(y, x_j, \bar{x}) - H(y, \bar{x}, x_j)) &\subset (-C - \{G(y, \bar{x}, x_j) + H(y, \bar{x}, x_j)\}) \\ &\subset -C - C = -C \quad \text{for all } y \in T(\bar{x}, x_j).\end{aligned}$$

This contradicts (1). Applying step 4 of Theorem 2, we conclude that there exists a point $\bar{x} \in D$ with $Q(\bar{x}) = \emptyset$. If $\bar{x} \notin S(\bar{x})$, then $Q(\bar{x}) = S(\bar{x}) = \emptyset$, which is impossible. Therefore, we deduce $\bar{x} \in S(\bar{x})$ and $Q(\bar{x}) = \mathrm{co}M_1(\bar{x})) \cap S(\bar{x}) = \emptyset$. This implies $M_1(\bar{x}) \cap S(\bar{x}) = \emptyset$ and hence

$$\begin{aligned}&\bar{x} \in S(\bar{x}), \\ &(G(y, x, \bar{x}) - H(y, \bar{x}, x)) \subset -C \quad \text{for all } x \in S(\bar{x}), \quad y \in T(\bar{x}, x).\end{aligned}$$

The proof is complete.

Theorem 3. *Assume that*

1. *for any $x' \in D$, the set*

$$\begin{aligned}A_2(x') = \{x \in D |\ &(G(y, x, x') - H(y, x', x)) \cap (-C) \neq \emptyset \\ &for\ some\ y \in T(x, x')\}\end{aligned}$$

   is open in D;
2. *the multivalued mapping $G + H$ is diagonally lower (T, C)-quasiconvex in the third variable;*
3. *for any fixed $y \in K$, the multivalued mapping $G(y, \cdot, \cdot) : D \times D \to 2^Y$ is upper C-monotone;*
4. *$(G(y, x, x) + H(y, x, x)) \subset C$ for all $(y, x) \in K \times D$.*

Then there exists $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$(G(y,x,\bar{x}) - H(y,\bar{x},x)) \cap (-C) \neq \emptyset \quad \text{for all} \quad x \in S(\bar{x}), \quad y \in T(\bar{x},x).$$

Proof. The proof proceeds exactly as the one in step 1 of Theorem 3 with M_1 replaced by

$$M_2(x) = \{x' \in D |\ (G(y,x',x) - H(y,x,x')) \cap (-C) = \emptyset \text{ for some } y \in T(x,x')\}.$$

Similarly, as in (1) we obtain

$$(G(y_i,x_i,\bar{x}) - H(y_i,\bar{x},x_i)) \cap (-C) = \emptyset \quad \text{for} \quad i = 1,...,n, \quad y_i \in T(\bar{x},x_i). \quad (4)$$

Since the multivalued mapping $G + H$ is diagonally lower (T,C)-quasiconvex in the third variable, there exists $j \in \{1,...,n\}$ such that

$$(G(y,\bar{x},x_j) + H(y,\bar{x},x_j)) \cap C \neq \emptyset \quad \text{for all } y \in T(\bar{x},x_j).$$

Since G is upper C-monotone, we deduce

$$(G(y,x_j,\bar{x})) \subset (-C - G(y,\bar{x},x_j)) \quad \text{for } y \in T(\bar{x},x_j).$$

Therefore, we have

$$G(y,\bar{x},x_j) + H(y,\bar{x},x_j) \subset (-C - \{G(y,x_j,\bar{x}) - H(y,\bar{x},x_j)\})$$

and then

$$\emptyset \neq (G(y,\bar{x},x_j) + H(y,\bar{x},x_j)) \cap C \subset C \cap (-C - \{(G(y,x_j,\bar{x}) - H(y,\bar{x},x_j))\}).$$

This implies

$$(G(y,x_j,\bar{x}) - H(y,\bar{x},x_j)) \cap (-C) \neq \emptyset \quad \text{for all } y \in T(\bar{x},x_j).$$

This contradicts (4). Further, we can argue as in the proof as in step 1 of Theorem 3.

Theorem 4. *Assume that*

1. *for any $x' \in D$, the set*

$$A_3(x') = \{x \in D |\ (G(y,x,x') - H(y,x',x)) \subset (C \setminus \{0\}) \text{ for some } y \in T(x,x')\}$$

   is open in D;
2. *the multivalued mapping $G + H$ is diagonally lower (T,C)-quasiconvex in the third variable;*
3. *for any fixed $y \in K$, the multivalued mapping $G(y,\cdot,\cdot) : D \times D \to 2^Y$ is upper C-monotone;*
4. *$(G(y,x,x) + H(y,x,x)) \cap (-C \setminus \{0\}) = \emptyset$ for all $(y,x) \in K \times D$.*

Then there exists $\bar{x} \in D$ *such that*

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$(G(y,x,\bar{x}) - H(y,\bar{x},x)) \not\subset (C \setminus \{0\}) \quad \text{for all} \quad x \in S(\bar{x}), \quad y \in T(\bar{x},x).$$

Proof. The proof proceeds exactly as the one in step 1 of Theorem 3 with M_1 replaced by

$$M_3(x) = \{x' \in D |\ (G(y,x',x) - H(y,x,x')) \subset C \setminus \{0\} \text{ for some } y \in T(x,x')\}.$$

Similarly, as in (1) we obtain

$$(G(y_i,x_i,\bar{x}) - H(y_i,\bar{x},x_i)) \subset C \setminus \{0\} \quad \text{for} \quad i = 1,...,n, \quad y_i \in T(\bar{x},x_i). \quad (5)$$

Since the multivalued mapping $G + H$ is diagonally lower (T,C)-quasiconvex in the third variable, there exists $j \in \{1,...,n\}$ such that

$$(G(y,\bar{x},x_j) + H(y,\bar{x},x_j)) \cap (C + G(y,\bar{x},\bar{x}) + H(y,\bar{x},\bar{x})) \neq \emptyset \quad \text{for all } y \in T(\bar{x},x_j).$$

Since G is upper C-monotone, we then have

$$(G(y,\bar{x},x_j) + H(y,\bar{x},x_j)) \subset (-C - \{G(y,x_j,\bar{x}) - H(y,\bar{x},x_j)\})$$
$$\text{for all } y \in T(\bar{x},x_j).$$

This implies

$$(C + G(y,\bar{x},\bar{x}) + H(y,\bar{x},\bar{x})) \cap (-C - \{G(y,x_j,\bar{x}) - H(y,\bar{x},x_j)\}) \neq \emptyset$$
$$\text{for all } y \in T(\bar{x},x_j).$$

Together with (5) we get

$$(G(y_j,\bar{x},\bar{x}) + H(y_j,\bar{x},\bar{x})) \cap -(C \setminus \{0\}) \neq \emptyset,$$

which is impossible by Assumption 4.

The rest of the proof can be done as in proving step 1 of Theorem 3.

Theorem 5. *Assume that*

1. *for any* $x' \in D$, *the set*

$$A_4(x') = \{x \in D |\ (G(y,x,x') - H(y,x',x)) \cap (C \setminus \{0\}) \neq \emptyset$$
$$\textit{for some} \quad y \in T(x,x')\}$$

   is open in D;
2. *the multivalued mapping* $G + H$ *is diagonally upper* (T,C)*-quasiconvex in the third variable with* $G(y,x,x) + H(y,x,x) \subset C$, *for any* $(y,x) \in D \times K$;
3. *for any fixed* $y \in K$, *the multivalued mapping* $G(y,\cdot,\cdot) : D \times D \to 2^Y$ *is upper* C*-monotone;*

4. $(G(y,x,x)+H(y,x,x))\cap(-C\setminus\{0\})=\emptyset$ *for all* $(y,x)\in K\times D$.

Then there exists $\bar{x}\in D$ *such that*

$$\bar{x}\in S(\bar{x}) \text{ and}$$
$$(G(y,x,\bar{x})-H(y,\bar{x},x))\cap(C\setminus\{0\})=\emptyset \quad \text{for all} \quad x\in S(\bar{x}),\quad y\in T(\bar{x},x).$$

Proof. The proof proceeds exactly as the one in step 1 of Theorem 3 with M_1 replaced by

$$M_4(x)=\{x'\in D|\quad (G(y,x',x)-H(y,x,x'))\cap(C\setminus\{0\})\neq\emptyset \quad \text{for some} \quad y\in T(x,x')\}.$$

Similarly, as in (1) we obtain

$$(G(y_i,x_i,\bar{x})-H(y_i,\bar{x},x_i))\cap(C\setminus\{0\})\neq\emptyset \quad \text{for} \quad i=1,...,n,\quad y_i\in T(\bar{x},x_i). \tag{6}$$

Since the multivalued mapping $G+H$ is diagonally upper (T,C)-quasiconvex in the third variable, there exists $j\in\{1,...,n\}$ such that

$$G(y,\bar{x},x_j)+H(y,\bar{x},x_j)\subset C+G(y,\bar{x},\bar{x})+H(y,\bar{x},\bar{x}) \quad \text{for all } y\in T(\bar{x},x_j). \tag{7}$$

Since G is upper C-monotone,

$$(G(y,x_j,\bar{x})-H(y,\bar{x},x_j))\subset(-C-\{G(y,\bar{x},x_j)+H(y,\bar{x},x_j)\}) \quad \text{for all } y\in T(\bar{x},x_j)$$

and then together with (7), we deduce

$$(G(y,x_j,\bar{x})-H(y,\bar{x},x_j))\subset(-C-\{G(y,\bar{x},\bar{x})+H(y,\bar{x},\bar{x})\}) \quad \text{for all } y\in T(\bar{x},x_j). \tag{8}$$

A combination of (6) and (8) gives

$$(C\setminus\{0\})\cap(-C-\{G(y_j,\bar{x},\bar{x})+H(y_j,\bar{x},\bar{x})\})\neq\emptyset.$$

It follows that

$$(G(y_j,\bar{x},\bar{x})+H(y_j,\bar{x},\bar{x}))\cap-(C\setminus\{0\})\neq\emptyset.$$

This is impossible by Assumption 4.
Further, we continue the proof as in step 1 of Theorem 3.

Theorem 6. *Assume that*

1. for any $x'\in D$, *the set*

$$A_5(x') = \{x \in D| \quad (G(y,x,x') - H(y,x',x)) \subset \text{int } C \\ for\ some \quad y \in T(x,x')\}$$

is open in D;

2. *the multivalued mapping $G+H$ is diagonally lower (T,C)-quasiconvex in the third variable with $G(y,x,x)+H(y,x,x) \subset C$, for any $(y,x) \in D \times K$;*
3. *for any fixed $y \in K$, the multivalued mapping $G(y,\cdot,\cdot) : D \times D \to 2^Y$ is upper C-monotone.*
4. *$(G(y,x,x) + H(y,x,x)) \cap -\text{int } C = \emptyset$ for all $(y,x) \in K \times D$.*

Then there exists $\bar{x} \in D$ such that

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$(G(y,x,\bar{x}) - H(y,\bar{x},x)) \not\subset \text{int } C \quad for\ all\ x \in S(\bar{x}), \quad y \in T(\bar{x},x).$$

Proof. The proof proceeds exactly as the one in step 1 of Theorem 3 with M_1 replaced by

$$M_5(x) = \{x' \in D| \quad (G(y,x',x) - H(y,x,x')) \subset \text{int } C \\ \text{for some} \quad y \in T(x,x')\}.$$

Similarly, as in (1) we obtain

$$(G(y_i,x_i,\bar{x}) - H(y_i,\bar{x},x_i)) \subset \text{int } C \quad \text{for } i = 1,...,n, \quad y_i \in T(\bar{x},x_i). \tag{9}$$

Since the multivalued mapping $G+H$ is diagonally lower (T,C)-quasiconvex in the third variable, there exists $j \in \{1,...,n\}$ such that

$$G(y,\bar{x},x_j) + H(y,\bar{x},x_j) \cap (C + G(y,\bar{x},\bar{x}) + H(y,\bar{x},\bar{x})) \neq \emptyset \\ \text{for all } y \in T(\bar{x},x_j).$$

Since G is upper C-monotone, we then have

$$G(y,\bar{x},x_j) + H(y,\bar{x},x_j) \subset (-C - \{G(y,x_j,\bar{x}) - H(y,\bar{x},x_j)\}) \\ \subset (-C - \text{int } C) = -\text{int } C \quad \text{for all } y \in T(\bar{x},x_j).$$

Together with (9), we conclude

$$(C + G(y_i,\bar{x},\bar{x}) + H(y_i,\bar{x},\bar{x})) \cap -\text{int } C \neq \emptyset.$$

It is impossible by Assumption 4.

Further, we continue the proof as in step 1 of Theorem 3.

Theorem 7. *Assume that*

1. *for any $x' \in D$, the set*

$$A_6(x') = \{x \in D| \quad (G(y,x,x') - H(y,x',x)) \cap \text{int } C \neq \emptyset \\ for\ some\ y \in T(x,x')\}$$

is open in D;

2. *the multivalued mapping* $G + H$ *is diagonally upper* (T, C)*-quasiconvex in the third variable;*
3. *for any fixed* $y \in K$*, the multivalued mapping* $G(y, \cdot, \cdot) : D \times D \to 2^Y$ *is upper* C*-monotone.*
4. $(G(y, x, x) + H(y, x, x)) \subset C$ *for all* $(y, x) \in K \times D$.

Then there exists $\bar{x} \in D$ *such that*

$$\bar{x} \in S(\bar{x}) \text{ and}$$
$$(G(y, x, \bar{x}) - H(y, \bar{x}, x)) \cap \text{int } C = \emptyset \quad \text{for all } x \in S(\bar{x}), \quad y \in T(\bar{x}, x).$$

Proof. The proof proceeds exactly as the one in step 1 of Theorem 3 with M_1 replaced by

$$M_6(x) = \{x' \in D| \quad (G(y, x', x) - H(y, x, x')) \cap \text{int } C \neq \emptyset \quad \text{for some } y \in T(x, x')\}.$$

Similarly, as in (1) we obtain

$$(G(y_i, x_i, \bar{x}) - H(y_i, \bar{x}, x_i)) \cap \text{int } C \neq \emptyset \quad \text{for } i = 1, ..., n, \quad y_i \in T(\bar{x}, x_i). \quad (10)$$

Since the multivalued mapping $G + H$ is diagonally upper (T, C)-quasiconvex in the third variable, there exists $j \in \{1, ..., n\}$ such that

$$(G(y, \bar{x}, x_j) + H(y, \bar{x}, x_j)) \subset (C + G(y, \bar{x}, \bar{x}) + H(y, \bar{x}, \bar{x})) \quad \text{for all } y \in T(\bar{x}, x_j).$$

Remarking that

$$(G(y, \bar{x}, \bar{x}) + H(y, \bar{x}, \bar{x})) \subset C,$$

we obtain

$$(G(y, \bar{x}, x_j) + H(y, \bar{x}, x_j)) \subset C \quad \text{for all } y \in T(\bar{x}, x_j). \quad (11)$$

Since G is upper C-monotone, we then have

$$(G\ (y, x_j, \bar{x}) - H(y, \bar{x}, x_j)) \subset (-C - \{G(y, \bar{x}, x_j) + H(y, \bar{x}, x_j)\})$$
$$\text{for all } y \in T(\bar{x}, x_j).$$

Taking account of (11), we conclude that

$$(G(y, x_j, \bar{x}) - H(y, \bar{x}, x_j)) \subset -C \quad \text{for all } y \in T(\bar{x}, x_j).$$

A combination of (10) and (11) gives

$$\text{int } C \cap (-C) \neq \emptyset.$$

It is impossible, since C is a pointed cone.
Further, we continue the proof as in step 1 of Theorem 3.

Remark 1.

1. In the case $G(y,x,x') = \{0\}$ (resp. $H(y,x,x') = \{0\}$) for all $(y,x,x') \in K \times D \times D$, the above theorems show the existence of solutions of quasi-equilibrium inclusion problems of the Ky Fan (of the Browder–Minty, respectively) type. These also generalize the results obtained by Luc and Tan [7], Minh and Tan [8, 9], and many other well-known results for vector optimization problems, variational inequalities, equilibrium, quasi-equilibrium problems concerning scalar and vector functions optimization, etc.
2. If G and H are single-valued mappings, then we can see that step 1 of Theorem 3 coincides with step 2 of Theorem 3, step 3 of Theorem 3 with step 4 of Theorem 3, and step 5 of Theorem 3 with step 6 of Theorem 3.

Further, the following propositions give sufficient conditions putting on the multivalued mappings T and F such that conditions 1 of the earlier theorems are satisfied.

Theorem 8. *Let $F : K \times D \to 2^Y$ be a lower C-continuous multivalued mapping with nonempty values and $T : D \to 2^K$ be a lower continuous multivalued mapping with nonempty values. Then the set*

$$A_1 = \{x \in D | \quad F(T(x), x) \not\subset -C\}$$

is open in D.

Proof. Let $\bar{x} \in A_1$ be arbitrary. We have $F(T(\bar{x}), \bar{x}) \not\subset -C$. Therefore, there exists $\bar{y} \in T(\bar{x})$ such that $F(\bar{y}, \bar{x}) \not\subset -C$. Since F is lower C-continuous at $(\bar{y}, \bar{x}) \in K \times D$, then for any neighborhood V of the origin in Y one can find neighborhoods U of $\bar{x}, W$ of $\bar{y}$ such that

$$F(\bar{y}, \bar{x}) \subset F(y,x) + V - C \quad \text{for all } (y,x) \in W \times U.$$

Since T is lower continuous at $\bar{x}$, one can find a neighborhood $U_0 \subset U$ of $\bar{x}$ such that

$$T(x) \cap W \neq \emptyset \quad \text{for all } x \in U_0 \cap D.$$

Hence, for any $x \in U_0 \cap D$ there is $y \in T(x) \cap W$, such that

$$F(\bar{y}, \bar{x}) \subset F(y,x) + V - C.$$

If there is some $x \in U_0 \cap D, y \in T(x), F(y,x) \subset -C$, then we have $F(\bar{y}, \bar{x}) \subset V - C$ for any V. It then follows that $F(\bar{y}, \bar{x}) \subset -C$ and we have a contradiction. So, we have shown that

$$F(T(x), x) \not\subset -C \quad \text{for all } x \in U_0 \cap D.$$

This means that $U_0 \cap D \subset A_1$ and then A_1 is open in D.

Theorem 9. *Let $F : K \times D \to 2^Y$ be an upper C-continuous multivalued mapping with nonempty values and $T : D \to 2^K$ be a lower continuous multivalued mapping with nonempty closed values. Then the set*

$$A_2 = \{x \in D| \quad F(y,x) \cap (-C) = \emptyset \quad for\ some \quad y \in T(x)\}$$

is open in D.

Proof. Let $\bar{x} \in A_2$ be arbitrary, $F(\bar{y},\bar{x}) \cap (-C) = \emptyset$, for some $\bar{y} \in T(\bar{x})$. Since F is upper C-continuous at $(\bar{y},\bar{x}) \in K \times D$, then for any neighborhood V of the origin in Y one can find neighborhoods U of $\bar{x}, W$of $\bar{y}$ such that

$$F(y,x) \subset F(\bar{y},\bar{x}) + V + C \quad \text{for all } (y,x) \in W \times U.$$

Since T is lower continuous at $\bar{x}$, one can find a neighborhood U_0 of $\bar{x}$ such that

$$T(x) \cap W \neq \emptyset \quad \text{for all } x \in U_0 \cap D.$$

Therefore, for any $x \in U_0 \cap D$ there is $y \in T(x) \cap W$, we have

$$F(y,x) \subset F(\bar{y},\bar{x}_0) + V + C.$$

If there is some $x \in U_0 \cap D, y \in T(x), F(y,x) \cap (-C) \neq \emptyset$, then we have $(F(\bar{y},\bar{x}) + V + C) \cap (-C) \neq \emptyset$ for any V. It then follows that $F(\bar{y},\bar{x}) \cap (-C) \neq \emptyset$ and we have a contradiction. So, we have shown that

$$F(T(x),x) \cap (-C) = \emptyset \quad \text{for all } x \in U_0 \cap D.$$

This means that $U_0 \cap D \subset A_2$ and then A_2 is open in D.

Analogously, we can prove the following propositions.

Proposition 1. *Let $F : K \times D \to 2^Y$ be an upper C-continuous multivalued mapping with nonempty values and $T : D \to 2^K$ be a lower continuous multivalued mapping with nonempty values. Then the set*

$$A_3 = \{x \in D| \quad F(y,x) \subset \text{int } C \quad for\ some \quad y \in T(x)\}$$

is open in D.

Proposition 2. *Let $F : K \times D \to 2^Y$ be a lower C-continuous multivalued mapping with nonempty values and $T : D \to 2^K$ be a lower continuous multivalued mapping with nonempty values. Then the set*

$$A_4 = \{x \in D| \quad F(y,x) \cap \text{int } C \neq \emptyset \quad for\ some \quad y \in T(x)\}$$

is open in D.

Remark 2.

1. Assume that the multivalued mappings T, G, and H are given as in step 1–6 of Theorem 3 with nonempty values. In addition, suppose that T is a lower continuous multivalued mapping. For any fixed $x \in D$ if the multivalued mapping $F : K \times D \to D$ defined by

$$F(y, x') = G(y, x, x') - H(y, x', x), \quad (y, x') \in K \times D,$$

   is lower, upper, upper, and lower C-continuous, then condition 1 of steps 1,2,5, and 6 of Theorem 3 is satisfied, respectively.
2. Assume that there exists a cone $\tilde{C} \subset Y$ such that $\tilde{C}$ is not whole space Y and $(C \setminus \{0\}) \subset \operatorname{int} \tilde{C}$ and the mapping T is lower continuous, the mapping F defined as above is upper (resp. lower) C-continuous, then step 3 of Theorem 3 (resp. step 4 of Theorem 3) is also true without condition 1 (apply steps 5 and 6 of Theorem 3 with C replaced by $\tilde{C}$).

To conclude this section, we consider the simple case when G and H are real functions. We can see that steps 1–6 of Theorem 3–are extensions of a result by Blum and Oettli to vector and multivalued problems. We have

Theorem 10. *Let D, K, S, T be as above with T lower continuous. Let $G, H : K \times D \times D \to R$ be real functions satisfying the following conditions:*

1. *for any fixed $(y, x) \in K \times D$ the function $F : D \to R$ defined by $F(x') = G(y, x, x') - H(y, x', x)$ is lower semi-continuous in the usual sense. For any fixed, $y \in K, x_1, x_2 \in D$, the function $g : [0, 1] \to R$ defined by $g(t) = G(y, tx_1 + (1 - t)x_2, x_2)$ is upper semi-continuous in the usual sense.*
2. *for any fixed $(y, x) \in K \times D, G(y, x, \cdot), H(y, x, \cdot)$ are convex functions.*
3. *for any fixed $y \in K$ the function $G(y, \cdot, \cdot)$ is monotone (i.e., $G(y, x, x') + G(y, x'x) \leq 0$ for all $x, x' \in D$).*
4. *$G(y, x, x) = H(y, x, x) = 0$ for all $(y, x) \in K \times D$.*

Then there exists $\bar{x} \in D$ such that $\bar{x} \in S(\bar{x})$ and

$$G(y, \bar{x}, x) + H(y, \bar{x}, x) \geq 0 \quad \text{for all } x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

Proof. Take $Y = R, C = R_+$, we can see that all assumptions in steps 1–6 of Theorem 3 are satisfied. Applying any of the theorems, we conclude that there exists $\bar{x} \in D$ with $\bar{x} \in S(\bar{x})$ such that

$$G(y, x, \bar{x}) - H(y, \bar{x}, x) \leq 0 \quad \text{for all} \quad x \in S(\bar{x}),\ \ y \in T(\bar{x}, x).$$

This is equivalent to

$$G(y, \bar{x}, x) + H(y, \bar{x}, x) \geq 0 \quad \text{for all} \quad x \in S(\bar{x}),\ \ y \in T(\bar{x}, x)$$

(see the proof in [1]).

References

1. Blum, E., Oettli, W.: From optimization and student. 64, 1–23 (1993)
2. Chan, D., Pang, J.S.: The generalized quasi-variational inequality problem. Math. Oper. Res. 7, 211–222 (1982)
3. Fan, K.: A Minimax Inequality and Application. In: O. Shisha (Ed.) Inequalities III, Academic, New York, NY (pp. 33), (1972)
4. Gurraggio, A., Tan, N. X.: On general vector quasi-optimization problems. Math. Meth. Oper. Res. 55, 347–358 (2002)
5. Lin, L.J., Yu, Z. T., Kassay, G.: Existence of equilibria for monotone multivalued mappings and its applications to vectorial equilibria. J. Optim. Theory Appl. 114, 189–208 (2002)
6. Luc, D.T.: Theory of vector optimization. Lect. Notes Eco. Math. Syst. 319, Springer-Verlag, (1989)
7. Luc, D.T., Tan, N.X.: Existence conditions in variational inclusions with constraints. Optimization 53 (5–6), 505–515 (2004)
8. Minh, N.B., Tan, N.X.: Some sufficient conditions for the existence of equilibrium points concerning multivalued mappings. Vietnam. J. Math. 28, 295–310, (2000)
9. Minh, N.B., Tan, N.X.: On the existence of solutions of quasivariational inclusion problems of Stampacchia type. Adv. Nonlinear Var. Inequal. 8, 1–16 (2005)
10. Parida, J., Sen, A.: A Variational-like inequality for multifunctions with applications. J. Math. Anal. Appl. 124, 73–81 (1987)
11. Park, S.: Fixed points and quasi-equilibrium problems. Nonlinear Oper. Theory. Math. Com. Model. 32, 1297–1304 (2000)
12. Tan, N.X.: On the existence of solutions of quasi-variational inclusion problems. J. Optim. Theory Appl. 123, 619–638 (2004)

General Quadratic Programming and Its Applications in Response Surface Analysis

Rentsen Enkhbat[1] and Yadam Bazarsad[2]

[1] Department of Mathematics, School of Economics Studies, National University of Mongolia, Ulaanbaatar, Mongolia
renkhbat46@ses.edu.mn

[2] Department of Econometrics, School of Computer Science and Management, Mongolian University of Science and Technology
ya_zardas4@yahoo.com

Summary. In this chapter, we consider the response surface problems that are formulated as the general quadratic programming. The general quadratic programming is split into convex quadratic maximization, convex quadratic minimization, and indefinite quadratic programming. Based on optimality conditions, we propose finite algorithms for solving those problems. As application, some real practical problems arising in the response surface, one of the main part of design of experiment, have been solved numerically by the algorithms.

Key words: concave programming, quadratic programming, global optimization, response surface problems

1 Introduction

The mathematical theory of experimental design is divided into two parts: design of extremal experiments and response surface problems. The main principle of extremal experiment is to obtain the maximum information about investigated process for a given number of experiments and reduce the number of experiments for a given precision for the model expressed by nonlinear regression functions. Meanwhile, the response surface deals with optimization problems defined over a given criteria of experiment and experimental region. In general, in design of experiment there are three types of optimization problems. First type requires to choose some design of experiments, in other words ways of construction of experimental data related to the model of the process, satisfying certain properties called optimality criteria. For example, there are A, D, E, and G optimality criteria. Such optimization problems arising in design of extremal experiments are usually deterministic and multiextremum. The second type of optimization is to solve identification problems or to find

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_6,

unknown parameters of the regression models for a fixed design of experiment and data.

The last is the response surface optimization problem. It is assumed that the experimenter is concerned with a technological process involving some response f which depends on the input variables $x_1, x_2, \ldots, x_n$ from a given experimental region. The standard assumptions on f are that f is twice differentiable on the experimental region and the independent variables $x_1, x_2, \ldots, x_n$ are controlled in the experimental process and measured with negligible error. The experimental region of variables can be even nonconvex set but in most cases, for simplicity, the experimenter restrict usually himself to spherical or box type of regions.

As an example, consider a situation where a chemist or chemical engineer is interested in the yield (output), f, of chemical reaction. The output depends on the reaction temperature (x_1), reaction pressure (x_2), concentration of one of the reactants (x_2), etc. In general, one has $f = f(x_1, x_2, x_3)$. The success of the response surface analysis depends on the approximation of f in its experimental region of variables by some polynomial, the so-called response surface model. For example, if the approximating function is linear, then we write $f = d_0 + d_1x_1 + \cdots + d_nx_n$. The response surface analysis for this linear model was studied in [2]. It is also natural that for the experimenter to consider second-order model as a generalized model of the linear model if he does not feel the latter good in terms of adequacy. On the other hand, the experimenter knows that the second-order model applied in the response surface adequately represents many scientific phenomena. Assume that the experimenter has the following second-order response surface model which is an adequate representation of the experimental data:

$$f = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j + \sum_{j=1}^{n} d_jx_j + q,$$

where coefficients $a_{ij}, d_j, q,\ i = 1, 2, \ldots, n$ are assumed to be found by solving an identification problem for a chosen design of experiment, for example, orthogonal central composite design [17]. Note that x_ix_j mean interactions between two factors x_i and x_j.

Then response surface optimization problem is to find global extremum of response surface models over an experimental region. In other words, to obtain maximum (minimum) output on the experimental region is a goal of the experimenter in the response surface problem. Main methods for solving the response surface problems in the literature [1, 2, 4–6, 8, 13–18, 20, 23] are local search algorithms based on descent methods.

This chapter is mainly motivated by the response surface analysis which requires to solve the general quadratic programming problem globally. This chapter is organized as follows. In Section 2, we describe the response surface methodology. In Section 3, we consider the quadratic maximization problem and propose an algorithm for its solution. In Sections 4 and 5, we consider

the quadratic minimization and the indefinite quadratic programming, respectively. In the last section, we deal with the response surface problems taken from real engineering applications and give their numerical solutions obtained by the proposed algorithms.

2 Response Surface Methodology

The actual plan of experimental levels in the x's is called the experimental design. Experimental designs for fitting a second-order response surface must involve at least three levels of each variable so that the coefficients in the model can be estimated. Obviously, the design that is automatically suggested by the model requirements is the 3^n factorial, a factorial experiment with each factor at three levels. In general case, the coefficients of the second-order model can be estimated by the least-squares method based on the following experimental observation data:

x_1	x_2	$\ldots$	x_n	f
x_{11}	x_{12}	$\ldots$	x_{1n}	f_1
x_{21}	x_{22}	$\ldots$	x_{2n}	f_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_{m1}	x_{m2}	$\ldots$	x_{mn}	f_m

Then according to the least-squares method, in order to find coefficients $A = \{a_{ij}\}$ and d_j, $i, j = 1, 2, \ldots, n$, we need to solve the following unconstrained minimization problem:

$$F(A, d) = \sum_{k=1}^{m} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_{ki} x_{kj} + \sum_{j=1}^{n} d_j x_{kj} + q - f_k \right)^2 \to \min .$$

Now a model which was assumed by the experimenter can be written as

$$f_k = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_{ki} x_{kj} + \sum_{j=1}^{n} d_j x_{kj} + q + \epsilon_k, \quad k = 1, 2, \ldots, m,$$

where ϵ_k is a random variable with zero mean and variance σ^2.

In practice it is convenient to use the orthogonal central composite design [20] which provides easy computations of the coefficients for the second-order model. On the other hand, the most useful and versatile class of experimental designs for fitting second-order models is the central composite design. This design serves as a natural alternative to the 3^n factorial design due to its requirement of fewer experimental observations and its flexibility. The central composite design is the 2^n factorial or fractional factorial (the levels of each variable coded to the usual -1, $+1$) augmented by the following.

$$\begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_n \\ 0 & 0 & 0 & \dots & 0 \\ -\alpha & 0 & 0 & \dots & 0 \\ \alpha & 0 & 0 & \dots & 0 \\ 0 & -\alpha & 0 & \dots & 0 \\ 0 & \alpha & 0 & \dots & 0 \\ 0 & 0 & -\alpha & \dots & 0 \\ 0 & 0 & \alpha & \dots & 0 \\ \vdots & \dots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\alpha \\ 0 & 0 & 0 & \dots & \alpha \end{pmatrix}.$$

One can construct the central composite design by choosing the appropriate value for α, the quantity which specifies the axial points. That is why, the central composite design with respect to 3^n-factorial provides a certain flexibility for the experimenter. Values for α for an orthogonal central composite design are given in the following table [20]:

n	2	3	4	5	6	7	8
α	1.00	1.216	1.414	1.596	1.761	1.910	2.045

The designs that are considered in the table contain a single center point. In general case, α and the regression coefficients are computed by the formulas

$$\alpha = \sqrt{2^{(n/2-1)}(\sqrt{m} - 2^{n/2})},\ q = \frac{\sum_{k=1}^{m} f_k}{m},\ d_j = \frac{\sum_{j=1}^{m} x_{ji} f_j}{2^n + 2\alpha^2},$$
$$a_{ik} = \frac{\sum_{j=1}^{m} x_{ji} x_{jk} f_j}{2^n},\ i,k = 1,2,\dots,n,\ i \neq k,\ a_{ii} = \frac{\sum_{j=1}^{m} \tilde{x}_{ji}^2 f_j}{d},$$
$$d = 2^n - \frac{(2^n + 2\alpha^2)^2}{m} + 2\alpha^4,\ \tilde{x}_i^2 = x_i^2 - \bar{x}_i^2,\ \bar{x}_i^2 = \frac{1}{m}\sum_{j=1}^{m} x_{ji}^2.$$

When we deal with first- and second-order models based on the complete factorial experiment 2^n, it is convenient to "code" the independent variables, with (-1) representing the low level of a variable and $(+1)$ the high level. This, of course, corresponds to the transformation

$$x_i = 2\left(\frac{y_i - \bar{y}_i}{x_i^{\max} - x_i^{\min}}\right),\ i = 1,2,\dots,n,$$

where y_i is the actual reading in the original units and $\bar{y}_i = \frac{(x_i^{\max} + x_i^{\min})}{2}$. The scalars $x_i^{\min}$ and $x_i^{\max}$ are the corresponding low and high levels of the variable y_i.

Assuming that the experimental region is box type, we can reformulate the response surface problem as the general quadratic programming with box constraints:

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j + \sum_{j=1}^{n} d_j x_j + q \;\longrightarrow \min(\max),\; x \in D, \tag{1}$$
$$D = \{x \in R^n \mid a_i \le x_i \le b_i,\; i = 1, 2, \ldots, n\}.$$

Note that maximization or minimization of the response f depends on the goal of the experimenter. For example, a chemical manufacturer is interested in their products with maximum concentration of primary component. And metallurgy researcher might be interested in the percentage of certain alloys which result in minimum corrosion.

Now we can treat problem (1) as the quadratic convex maximization, the quadratic convex minimization, and the indefinite quadratic programming problems, respectively, depending on the matrix $A = (a_{ij}),\; i, j = 1, 2, \ldots, n$.

3 Quadratic Convex Maximization Problem

Consider the quadratic maximization problem:

$$f(x) = \langle Cx, x\rangle + \langle d, x\rangle + q \;\longrightarrow\; \max,\; x \in D, \tag{2}$$

where C is a positive semidefinite $(n \times n)$ matrix and $D \subset R^n$ is a nonempty arbitrary subset of R^n. A vector $d \in R^n$ and a number $q \in R$ are given. Then the optimality conditions for problem (2) are stated as follows.

Theorem 1 (Enkhbat [10]). *Let $z \in D$ be such that $f'(z) \neq 0$. Then z is a solution of problem (2) if and only if*

$$\langle f'(y), x - y\rangle \le 0 \text{ for all } y \in E_{f(z)}(f) \text{ and } x \in D, \tag{3}$$

where $E_c(f) = \{y \in R^n \mid f(y) = c\}$.

Now introduce the definitions.

Definition 1. *The set $E_{f(z)}(f)$ defined by*

$$E_{f(z)}(f) = \{y \in R^n \mid f(y) = f(z)\}$$

is called the level set of f at z.

Definition 2. *The set A_z^m defined by*

$$A_z^m = \{y^1, y^2, \ldots, y^m \mid y^i \in E_{f(z)}(f),\; i = 1, 2, \ldots, m\} \tag{4}$$

is called the approximation set to the level set $E_{f(z)}(f)$ at the point z.

For further purpose, consider the following quadratic maximization problem over a box constraint:

$$f(x) = \langle Cx, x\rangle + \langle d, x\rangle + q \longrightarrow \max,\ x \in D \subset R^n, \tag{5}$$
$$D = \{x \in R^n \mid a_i \le x_i \le b_i,\ i = 1, 2, \ldots, n\},$$

where C is a symmetric positive semidefinite $n \times n$ matrix and the vectors $a, b, d \in R^n$ and a number $q \in R$ are given.

Let $z = (z_1, z_2, \ldots, z_n)$ be a local maximizer of problem (5). Then due to [21], $z_i = a_i \vee b_i,\ i = 1, 2, \ldots, n$. In order to construct an approximation set A_z^m take the following steps.
Define points $v^1, v^2, \ldots, v^{n+1}$ by formulas

$$v_i^k = \begin{cases} z_i \ \text{if}\ i \neq k, \\ a_k \ \text{if}\ z_k = b_k, \\ b_k \ \text{if}\ z_k = a_k,\ i, k = 1, 2, \ldots, n \end{cases} \tag{6}$$

and

$$v_i^{n+1} = \begin{cases} a_i \ \text{if}\ z_i = b_i, \\ b_i \ \text{if}\ z_i = a_i,\ i = 1, 2, \ldots, n. \end{cases} \tag{7}$$

Clearly,

$$\|v^{n+1} - z\| > \|v^k - z\|,\ k = 1, 2, \ldots, n,$$
$$\sum_{i=1}^{n} (a_i - b_i)^2 = \|v^{n+1} - z\|^2.$$

Denote by h^i vectors $h^i = v^i - z,\ i = 1, 2, \ldots, n+1$. Note that $\langle h^k, h^j\rangle = 0,\ k \neq j,\ k, j = 1, 2, \ldots, n$. Define the approximation set A_z^{n+1} by

$$A_z^{n+1} = \{y^1, y^2, \ldots, y^{n+1} \mid y^i \in E_{f(z)}(f),\ y^i = z - \alpha_i h^i,\ i = 1, \ldots, n+1\}, \tag{8}$$

where $\alpha_i = \frac{\langle 2Cz + d, h^i\rangle}{\langle Ch^i, h^i\rangle},\ i = 1, 2, \ldots, n+1$.

Then, an algorithm for solving (5) is described in the following.

Algorithm 1

Input: A convex quadratic function f and a box set D.
Output: An approximate solution x to problem (5); i.e., an approximate global maximizer of f over D.
Step 1. Choose a point $x^0 \in D$. Set $k := 0$.
Step 2. Find a local maximizer $z^k \in D$ by the projected gradient method starting with an initial approximation point x^k.
Step 3. Construct an approximation set $A_{z^k}^{n+1}$ at the point z^k by formulas (6), (7), and (8).
Step 4. For each $y^i \in A_{z^k}^{n+1},\ i = 1, 2, \ldots, n+1$ solve the problems

$$\langle f'(y^i), x\rangle \longrightarrow \max,\ x \in D,$$

which have analytical solutions $u^i,\ i = 1, 2, \ldots, n+1$ found as

$$u_s^i = \begin{cases} a_s \text{ if } (2Cy^i + d)_s \le 0, \\ b_s \text{ if } (2Cy^i + d)_s > 0, \end{cases}$$

where $i = 1, 2, \ldots, n+1$ and $s = 1, 2, \ldots, n$.
Step 5. Find a number $j \in \{1, 2, \ldots, n+1\}$ such that

$$\theta_{n+1}^k = \langle f'(y^j), u^j - y^j\rangle = \max_{i=1,2,\ldots,n+1} \langle f'(y^i), u^i - y^i\rangle.$$

Step 6. If $\theta_{n+1}^k > 0$ then $x^{k+1} := u^j,\ k := k+1$ and go to step 1.
Step 7. Find $y \in E_{f(z^k)}(f)$ such that

$$y = z^k - \frac{\langle 2Cz^k + d, u^j - z^k\rangle}{\langle C(u^j - z^k), u^j - z^k\rangle}(u^j - z^k).$$

Step 8. Solve the problem $\langle f'(y), x\rangle \longrightarrow \max,\ x \in D$.
Let v be the solution, i.e., $\langle f'(y), v\rangle = \max_{x\in D}\langle f'(y), x\rangle$. Compute $\theta^k = \langle f'(y), v - y\rangle$.
Step 9. If $\theta^k > 0$ then $x^{k+1} := v,\ k := k+1$ and go to step 1. Otherwise, z^k is an approximate maximizer and terminate.

4 Indefinite Quadratic Programming

Consider the general quadratic programming problem of the form.

$$f(x) = \langle Cx, x\rangle + \langle d, x\rangle + q \ \longrightarrow\ \min,\ x \in D, \tag{9}$$

where $D = \{x \in R^n \mid a_i \le x_i \le b_i,\ i = 1, 2, \ldots, n\}$ is a box set, C is a symmetric indefinite $n \times n$ matrix, and $a, b, d \in R^n,\ q \in R$.

We use the fact that a symmetric quadratic matrix can be presented as a difference of a two positive semidefinite matrices [19]. Let C' and C'' be positive semidefinite matrices such that $C = C' - C'$.

Define the convex functions $\varphi(x)$ and $\psi(x)$ as follows:

$$\varphi(x) = \langle C'x, x\rangle + \langle d, x\rangle + q,$$
$$\psi(x) = \langle C''x, x\rangle.$$

Then, problem (9) reduces to its equivalent so-called a d.c. programming problem:

$$f(x) = \varphi(x) - \psi(x) \ \longrightarrow\ \min,\ x \in D. \tag{10}$$

Moreover, it can be shown that the latter is equivalent to the following convex maximization problem:

$$g(x, x_{n+1}) = \psi(x) - x_{n+1} \longrightarrow \max, \tag{11}$$

subject to $\varphi(x) - x_{n+1} \leq 0,\ x \in D$.

Clearly, if (z, z_{n+1}) is a solution to problem (11) then z is a solution to (10) with $\varphi(z) = z_{n+1}$. Denote by $\bar{D}$ and $\bar{E}_{g(z,z_{n+1})}(g)$ the following sets:

$$\bar{D} = \{(x, x_{n+1}) \in D \times R \mid \varphi(x) - x_{n+1} \leq 0\},$$
$$\bar{E}_{g(z,z_{n+1})}(g) = \{(y, y_{n+1}) \in R^{n+1} \mid g(y, y_{n+1}) = g(z, z_{n+1})\}.$$

Then optimality conditions for problem (11) are given by the following theorem.

Theorem 2. *A point $(z, z_{n+1}) \in \bar{D}$ is a solution of problem (11) if and only if*

$$\langle \psi'(y), x - y\rangle - x_{n+1} + y_{n+1} \leq 0 \tag{12}$$

hold for all $(y, y_{n+1}) \in \bar{E}_{g(z,z_{n+1})}(g)$ and $(x, x_{n+1}) \in \bar{D}$.

The proof is immediate from Theorem 1.

An algorithm for solving problem (11) is constructed similar to Algorithm 1, but we need to specify a way of constructing approximation set to the level set $\bar{E}$ and choose an appropriate method for solving the problem

$$\langle g'(y^k, y^k_{n+1}), (x, x_{n+1})\rangle \longrightarrow \max,\ (x, x_{n+1}) \in \bar{D} \tag{13}$$

at kth iteration.

Let z^k, z^k_{n+1} be a stationary point or a local maximizer of problem (11) found by applying one of the gradient methods. Suppose that we have an approximation set $A^m_{z^k}$ to the level set $E_{\psi(z^k)}(\psi)$ of function $\psi(x)$ at a point z^k. Define points $y^i_{n+1},\ i = 1, 2, \ldots, m$ as $y^i_{n+1} = -g(z^k, z^k_{n+1}) + \psi(y^i)$. Then an approximation set to the level set $\bar{E}$ is as follows:

$$\bar{A}^m_{(z^k, z^k_{n+1})} = \{(y^1, y^1_{n+1}), \ldots, (y^m, y^m_{n+1}) \mid (y^i, y^i_{n+1}) \in \bar{E}_{g(z^k, z^k_{n+1})}(g),$$
$$i = 1, 2, \ldots, m\}. \tag{14}$$

Note that problem (13) can be written in the form

$$\langle \psi'(y^k), x\rangle - x_{n+1} \longrightarrow \max,\ \varphi(x) - x_{n+1} \leq 0,\ x \in D.$$

We can easily reduce this to its equivalent problem

$$\varphi(x) - \langle \psi'(y^k), x\rangle \longrightarrow \min,\ x \in D,$$

which is convex (quadratic) minimization problem due to the positive semidefinite matrix C'. If x^k is a solution of the latter then (x^k, x^k_{n+1}) is a solution to problem (13) with $x^k_{n+1} = \varphi(x^k)$.

Now based on the above results and Algorithm 1, we can describe Algorithm 2 for solving problem (9) or its equivalent convex maximization problem (11).

Algorithm 2
Input: A quadratic function f and a box set D. The positive semidefinite matrices C' and C'' be such that $C = C' - C''$.
Output: An approximate solution x to problem (9); i.e., an approximate global minimizer of f over D.
Step 1. Choose a point $(x^0, x^0_{n+1}) \in \bar{D} = \{(x, x_{n+1}) \in R^n \times R \mid \langle C'x, x\rangle + \langle d, x\rangle + q - x_{n+1} \le 0,\ x \in D\}$. Set $k := 0$.
Step 2. Let z^k, z^k_{n+1} be a stationary point or a local maximizer of the problem

$$g(x, x_{n+1}) = \langle C''x, x\rangle - x_{n+1} \longrightarrow \max,\ x \in \bar{D}$$

found by one of the gradient methods starting with the initial approximation point (x^k, x^k_{n+1}).
Step 3. Construct an approximation set $A^{n+1}_{z^k}$ at the point z^k by formulas (6), (7), and (8). Then construct $\bar{A}^{n+1}_{(z^k, z^k_{n+1})}$ by (14) with $y^i_{n+1} = -g(z^k, z^k_{n+1}) + \langle C''y^i, y^i\rangle,\ i = 1, 2, \ldots, n+1$.
Step 4. For each $y^i \in A^{n+1}_{z^k},\ i = 1, 2, \ldots, n+1$ solve the problems

$$\langle C'x, x\rangle + \langle d - 2C''y^i, x\rangle + q \longrightarrow \min,\ x \in D,$$

by the projection gradient method. Let $u^i,\ i = 1, 2, \ldots, n+1$ be solutions. Then set $u^i_{n+1} = \langle C'u^i, u^i\rangle + \langle d, u^i\rangle + q,\ i = 1, 2, \ldots, n+1$.
Step 5. Find a number $j \in \{1, 2, \ldots, n+1\}$ such that

$$\begin{aligned}\theta^k_{n+1} &= \langle 2C''y^j, u^j - y^j\rangle - u^j_{n+1} + y^j_{n+1}\\ &= \max_{i=1,2,\ldots,n+1}(\langle 2C''y^i, u^i - y^i\rangle - u^i_{n+1} + y^i_{n+1}).\end{aligned}$$

Step 6. If $\theta^k_{n+1} > 0$ then $x^{k+1} := u^j,\ x^{k+1}_{n+1} = u^j_{n+1},\ k := k+1$ and go to step 1.
Step 7. Find a $(y, y_{n+1}) \in \bar{E}_{g(z^k, z^k_{n+1})}(g)$ such that

$$\begin{aligned}y &= z^k - \frac{\langle 2C''z^k, u^j - z^k\rangle}{\langle C''(u^j - z^k), u^j - z^k\rangle}(u^j - z^k),\\ y_{n+1} &= -g(z^k, z^k_{n+1}) + \langle C''y, y\rangle.\end{aligned}$$

Step 8. Solve the minimization problem

$$\langle C'x, x\rangle + \langle d - 2C''y, x\rangle + q \longrightarrow \min,\ x \in D,$$

by the projection gradient method. Let v be solution of this problem. Then set $v_{n+1} = \langle C'v, v\rangle + \langle d, v\rangle + q$.
Step 9. If $\langle 2C''y, v - y\rangle - v_{n+1} + y_{n+1} > 0$ then $x^{k+1} := v,\ x^{k+1}_{n+1} = v_{n+1}, k := k+1$ and go to step 1. Otherwise, z^k is an approximate minimizer of problem (9) and terminate.

5 Quadratic Convex Minimization Problem

Consider the quadratic minimization problem

$$f(x) = \langle Cx, x\rangle + \langle d, x\rangle + q \;\longrightarrow\; \min,\; x \in D, \tag{15}$$

where $D = \{x \in R^n \mid a_i \le x_i \le b_i,\; i = 1, 2, \ldots, n\}$ is a box set, C is a symmetric positive semidefinite $n \times n$ matrix, and $a, b, d \in R^n,\; q \in R$.

Then the well-known optimality condition for problem (15) is in Rockafellar [21]:

Theorem 3. *Let $z \in D$. Then z is a solution of problem (15) if and only if*

$$\langle f'(z), x - z\rangle \ge 0 \quad for\ all x \in D. \tag{16}$$

Introduce the index set

$$I(x) = \{\; i \mid x_i = a_i \vee b_i,\; i = 1, 2, \ldots, n\}$$

at a point $x \in D$.

Then the optimality condition (16) for problem (15) is transformed into the following condition in terms of the index set.

Theorem 4. *$z \in D$ is a solution to problem (15) if and only if*

$$\begin{cases} 2(Cz)_j + d_j = 0 \; if j \notin I(z), \\ 2(Cz)_j + d_j \ge 0 \; if j \in I(z) : z_j = a_j, \\ 2(Cz)_j + d_j \le 0 \; if j \in I(z) : z_j = b_j,\; i = 1, 2, \ldots, n. \end{cases} \tag{17}$$

An algorithm for solving problem (15) based on Theorem 4 is given in [12]. Before describing this algorithm denote by $P_D(y)$ a projection of a point $y \in R^n$ on the box set D which is a solution to the following quadratic programming problem:

$$\|x - y\|^2 \;\longrightarrow\; \min,\; x \in D.$$

We can solve this problem analytically to obtain its solution as follows [25]:

$$(P_D(y))_i = \begin{cases} a_i \text{ if } y_i \le a_i, \\ y_i \text{ if } a_i < y_i < b_i, \\ b_i \text{ if } y_i \ge b_i,\; i = 1, 2, \ldots, n. \end{cases} \tag{18}$$

Algorithm 3
Input: A quadratic convex function f and a box set D.
Output: A solution x to problem (15).
Step 1. Choose a parameter $\gamma \in (0, 1)$, a point $y \in R^n$ and find a $x^0 = P_D(y)$ by (18). Set $k := 0$ and $m := 0$.
Step 2. Construct the index set

$$I(x^k) = \{\ i \mid x_i^k = a_i \vee b_i,\ i = 1, 2, \ldots, n\}$$

at a point $x^k \in D$.
Step 3. Let u^k be a solution of the problem

$$f(x) = \langle Cx, x\rangle + \langle d, x\rangle + q \longrightarrow \min,\ x_i = a_i \vee b_i,\ i \in I(x^k)$$

solved by the conjugate gradient method.
Step 4. If $u^k \notin D$ then construct $x^{k+1} = x^k + \lambda_k(u^k - x^k)$, where

$$\lambda_k = \min\left\{\min_{j\in J_k} \frac{b_j - x_j^k}{u_j^k - x_j^k};\ \min_{j\notin J_k} \frac{x_j^k - a_j}{x_j^k - u_j^k}\right\}$$
$$J_k = \{i \mid u_i^k - x_j^k > 0,\ i \notin I(x^k),\ 1 \le i \le n\},$$

and set $k := k + 1$, and return to step 2.
Step 5. Check optimality condition (17) at the point u^k by

$$\begin{cases} 2(Cu^k)_j + d_j = 0 \text{ if } j \notin I(u^k), \\ 2(Cu^k)_j + d_j \ge 0 \text{ if } j \in I(u^k) : u_j^k = a_j, \\ 2(Cu^k)_j + d_j \le 0 \text{ if } j \in I(u^k) : u_j^k = b_j,\ i = 1, 2, \ldots, n. \end{cases}$$

If this condition hold then u^k is a solution and terminate. Otherwise go to the next step.
Step 6. Construct the point $v = P_D(u^k - \alpha f'(u^k)$ with $\alpha := \gamma^m$.
Step 7. If $f(v) < f(u^k)$ then go to step 2 with $x^k := v$. Otherwise set $m := m + 1$ and go to step 6.

The convergence of this algorithm is given by the following theorem in [12].

Theorem 5. *The sequence* $\{u^k\}, (k = 0, 1, \ldots)$ *generated by Algorithm 3 converges to the solution of problem (15) in a finite number of steps.*

5.1 Response Surface Practical Problems

The following problems, from [3, 7, 9, 11, 20, 22, 23], arisen in actual industrial technological process have been considered. First those problems have been classified into convex and nonconvex and then solved numerically by Algorithms 1–3 on IBM PC/586 in Pascal. For the sake of simplicity, we omitted units of measure of all variables (factors) in some problems. Consider the list of these problems. Note that some of these problems were considered in their coded variables in the interval $[-1, 1]$.

Problem 1 (Myers [20]). Consider a chemical process in which 1,2-prepranditial is being converted to 2,5-dimethylpiperazine. The object is to examine the effect of several factors on the course of the reaction and to determine the conditions which give rise to maximum conversion. The following four variables were studied:

NH_3: amount of ammonia, grams
T: temperature, °C
H_2O: amount of water, grams
P: hydrogen pressure, psi
The variables are coded in the following way:

$$x_1 = \frac{NH_3 - 102}{51},\ x_2 = \frac{T - 250}{20},\ x_3 = \frac{H_2O - 300}{200},\ x_4 = \frac{P - 850}{350}.$$

Using the central composite design, a second-order model was obtained as follows:

$$\begin{aligned} f = {} & 40.198 - 1.511x_1 + 1.284x_2 - 8.739x_3 + 4.955x_4 - 6.332x_1^2 - 4.292x_2^2 \\ & +0.020x_3^2 - 2.506x_4^2 + 2.194x_1x_2 - 0.144x_1x_3 + 1.581x_1x_4 + 8.006x_2x_3 \\ & +2.806x_2x_4 + 0.294x_3x_4. \end{aligned}$$

This problem is an indefinite program and solved by Algorithm 2 providing solutions:
$x_1 = -0.403, x_2 = -0.9899, x_3 = -0.995$ with the improved result of $f = 47.9742$ against $f = 43.53$ in [20, p. 86].

Problem 2 (Myers [20]). It is of interest to know the relationship between the yield of mercaptobenzothiazole (MBT) and the independent variables, time and temperature. A fitted second-order response surface was found to be

$$f = 82.17 - 1.01x_1 - 8.61x_2 + 1.40x_1^2 - 8.76x_2^2 - 7.20x_1x_2,$$

where

$$x_1 = \frac{\text{time (h)} - 12}{8},\ x_2 = \frac{\text{temp. (°C)} - 250}{30}.$$

The experimenter is interested in maximum yield. The problem is an indefinite programming problem with the solutions $x_1 = 0.996, x_2 = -0.8999$ and the improved result $f = 89.66$ against $f = 85.602$ in [20, p. 105].

Problem 3 (Myers [20]). Data are presented in Table 1 from an experiment designed for estimating optimum conditions for storing bovine semen to retain maximum survival. The variables under study are the % sodium citrate (x_1), % glycerol (x_2), and the equilibration time in hours (x_3). The important response measured was % survival of motile spermatozoa (f). Table 1 gives the experimental data for a three-dimensional central composite design with $\alpha = 2.0$.

The coded factor levels are given by

	-2	-1	0	1	2
x_1	1.6	2.3	3.0	3.7	4.4
x_2	2.0	5.0	8.0	11.0	14.0
x_3	4.0	10.0	16.0	22.0	28.0

Table 1. Treatment combination and survival

Treatment combination	% Sodium citrate	% Glycerol	Equilibration time, h	% Survival
1	−1	−1	−1	57
2	1	−1	−1	40
3	−1	1	−1	19
4	1	1	−1	40
5	−1	−1	1	54
6	1	−1	1	41
7	−1	1	1	21
8	1	1	1	43
9	0	0	0	63
10	−2	0	0	28
11	2	0	0	11
12	0	−2	0	2
13	0	2	0	18
14	0	0	−2	56
p15	0	0	2	46

The response function was estimated by the usual techniques and found to be

$$f = 66.3889 - 1.4400x_1 - 2.2812x_2 - 1.0950x_3 - 11.3561x_1^2 - 13.6798x_2^2 \\ -3.4972x_3^2 + 9.1000x_1x_2 + 0.6075x_1x_3 + 0.8125x_2x_3$$

in terms of the coded independent variables. This problem is a convex minimization (concave maximization) and solution was found by Algorithm 3. The result is given by

$$x_1 = -0.1198,\ x_2 = -0.1286,\ x_3 = -0.1819.$$

These values correspond to the uncoded x levels of 2.9% sodium citrate, 7.6% glycerol, and 14.9 h equilibration time. The estimated response at the maximum point is $f = 66.72\%$ survival.

Problem 4 (Myers [20]). In a process designed to purify an antibiotic product (Lind, Goldin, and Hickman), it was decided that a response surface study should be employed in the solvent extraction operation in the process. The yield of the product at this stage of the process and the cost of the operation represent very critical responses. The operation involved extracting the antibiotic into an organic solvent. Certain chemicals, called reagents A and B, were added to form material which is soluble in the solvent. Concentration of the two reagents and the pH in the extraction environment were chosen as the independent variables to be studied. These variables are coded as follows:

$$x_1 = \frac{\%A - 0.5}{0.5},\ x_2 = \frac{\%B - 0.5}{0.5},\ x_3 = \frac{\text{pH} - 5.0}{0.5}$$

and then the second-order model was

$$f = 65.05 + 1.63x_1 + 3.28x_2 + 0.93x_3 - 2.93x_1^2 - 2.02x_2^2 \\ -1.07x_3^2 - 0.53x_1x_2 - 0.68x_1x_3 - 1.44x_2x_3.$$

This problem is a convex minimization (or concave maximization) problem and its solution is $x_1 = 0.2256, x_2 = 0.8589, x_3 = -0.2150$. These correspond to $\%A = 0.6128, \%B = 0.9294, \text{pH} = 4.8925$.

Problem 5 (Sebostyanov and Sebostyanov [23]). $f(x) = -0.31x_1^2 - 0.125x_2^2 + 0.09x_1x_2 + 187x_1 + 9x_2 - 29700 \longrightarrow \max,\ x \in D$,
$D = \{x \in R^2 \mid 316 \leq x_1 \leq 334, 130 \leq x_2 \leq 170\}$, where f is the strength against washing, x_2 the tension of strings, x_1 some angle.

This is a quadratic convex minimization problem and its solution is $x_1 = 323.74, x_2 = 152.436$.

Problem 6 (Anderson and Mclean [3]). $f(x) = 8.300x_1x_2 + 8.076x_1x_3 - 6.625x_1x_4 + 3.213x_2x_3 - 16.998x_2x_4 - 17.127x_3x_4 - 1.558x_1 - 2.351x_2 - 2.426x_3 + 14.372x_4 \longrightarrow \max,\ x \in D$, $D = \{x \in R^4 \mid 0.40 \leq x_1 \leq 0.60, 0.10 \leq x_2 \leq 0.50, 0.10 \leq x_3 \leq 0.50, 0.03 \leq x_4 \leq 0.08\}$, where f is the amount of illumination (measured in 1000 candles), x_1 magnesium, x_2 sodium nitrate, x_3 strontium nitrate, x_4 binder.

This is an indefinite program and its solution is $x_1 = 0.5233, x_2 = 0.2299, x_3 = 0.1668, x_4 = 0.080$.

Problem 7 (Sebostyanov and Sebostyanov [23]). $f(x) = 0.438x_1^2 + 0.423x_2^2 + 0.313x_3^2 - 0.145x_1x_2 + 0.385x_1x_3 - 0.08x_2x_3 + 0.687x_1 + 0.193x_2 + 0.736x_3 + 15.39 \longrightarrow \max,\ x \in D$, $D = \{x \in R^3 \mid -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1, -1 \leq x_3 \leq 1\}$, where f is the thickness of string, x_1 the tension of strings, x_2 some angle, x_3 the density of strings.

This is a quadratic convex maximization problem and its solution in coded variables is $x_1 = 1, x_2 = -1, x_3 = 1$.

Problem 8 (Sebostyanov and Sebostyanov [23]). $f(x) = 0.52x_1^2 - 0.25x_2^2 - 0.93x_3^2 - 0.23x_1x_2 - 0.22x_1x_3 - 0.02x_2x_3 + 2.02x_1 + 0.7x_2 - 0.52x_3 + 17.8 \longrightarrow \min,\ x \in D$, $D = \{x \in R^3 \mid -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1, -1 \leq x_3 \leq 1\}$, where f is the pressure on one string, x_1 the tension of strings, x_2 the some angle, x_3 the thickness.

This is an indefinite programming problem and its solution in coded variables is $x_1 = -1, x_2 = 1, x_3 = 1$.

Problem 9 (Bazarsad, Enkhtuya and Enkhbat [7]). $f(x) = 0.4x_1^2 - 0.16x_2^2 + 0.11x_3^2 - 0.26x_1x_2 - 0.14x_1x_3 + 0.01x_2x_3 + 0.38x_1 + 1.02x_2 + 0.49x_3 + 37.3 \longrightarrow \max,\ x \in D$,
$D = \{x \in R^3 \mid -1.682 \leq x_1 \leq 1.682, -1.682 \leq x_2 \leq 1.682, -1.682 \leq x_3 \leq 1.682\}$, where f is the average diameter of wool strings, x_1 the distance between two spinning wheels, x_2 speed, x_3 moisture.

This is an indefinite programming problem and its solution in coded variables is $x_1 = -1.682, x_2 = 1.682, x_3 = 1.682$.

Problem 10 (Bazarsad, Enkhtuya, and Enkhbat [7]). $f(x) = 0.66x_1^2 + 2.14x_2^2 + 0.87x_3^2 - 0.2x_1x_3 + 0.28x_2x_3 + 1.39x_1 - 3.21x_2 - 1.5x_3 + 23.16 \longrightarrow \min,\ x \in D$,
$D = \{x \in R^3 \mid -1.682 \le x_1 \le 1.682, -1.682 \le x_2 \le 1.682, -1.682 \le x_3 \le 1.682\}$, where f is the quantity of defective wool strings, x_1 the distance between two spinning wheels, x_2 speed, x_3 moisture.

This is a convex minimization problem and its solution in coded variables is $x_1 = -0.95, x_2 = -0.70, x_3 = 0.63$.

Problem 11 (Ruvinshtein and Bolkova [22]). $f(x) = 5.92x_4^2 - 17.71x_5^2 + 3.323x_1x_2 + 1.42x_1x_3 + 2.433x_1x_4 + 2.793x_1x_5 + 1.55x_1x_6 + 1.916x_2x_5 - 3.356x_3x_4 - 2.159x_3x_6 - 1.713x_4x_5 - 1.906x_4x_6 - 2.489x_1 + 1.759x_3 + 1.626x_4 + 1.139x_6 + 72.496 \longrightarrow \max,\ x \in D$,
$D = \{x \in R^6 \mid -1 \le x_i \le 1,\ i = 1, 2, \ldots, 6\}$, where f is the efficiency against dustiness, x_1 the moisture of coal, x_2 the elasticity of coal, x_3 the quantity of air, x_4 amplitude, x_5 frequency, x_6 first category of 1 mm.

This is an indefinite programming problem and its solution in coded variables is $x_1 = 1, x_2 = 1, x_3 = -1, x_4 = 1, x_5 = 0.08, x_6 = 1$.

Problem 12 (Enkhbat and Chuluunhuyag [11]). $f(x) = -0.34x_1^2 + 7.64x_2^2 - 0.061x_3^2 - 12.7x_1x_2 - 1.5x_1x_3 - 1.04x_2x_3 \longrightarrow \min,\ x \in D$,
$D = \{x \in R^3 \mid 1.25 \le x_1 \le 2, 0.8 \le x_2 \le 1.2, 5 \le x_3 \le 14\}$, where f is the refinement of water, x_1 the diameter of filtration, x_2 the height of filtration, x_3 the speed of filtration.

This is an indefinite programming problem and its solution is $x_1 = 2, x_2 = 1.2, x_3 = 14$.

Problem 13 (Chimedochir and Enkhbat [9]). $f(x) = 7.33x_1^2 - 5.451x_2^2 - 0.621x_3^2 + 7.454x_1x_2 - 5.573x_1x_3 + 0.807x_2x_3 + 64.366x_1 + 5.593x_2 + 4.296x_3 + 23.16 \longrightarrow \max,\ x \in D$,
$D = \{x \in R^3 \mid 4 \le x_1 \le 21, 0.001 \le x_2 \le 0.005, 35 \le x_3 \le 55\}$, where f is the quantity of carotenoid in the fruit "chazargan," x_1 the frequency of apparat, x_2 the amplitude of apparat, x_3 the temperature of diffusion process.

This is an indefinite programming problem and its solution is $x_1 = 4, x_2 = 0.005, x_3 = 35$.

Problem 14 (Tsetgee [24]). $f(x) = -2.35x_1^2 - 0.56x_2^2 - 6.84x_3^2 + 0.29x_1x_2 + 0.16x_1x_3 + 0.15x_2x_3 - 1.45x_1 - 0.43x_2 - 1.66x_3 + 23.24 \longrightarrow \min,\ x \in D$,
$D = \{x \in R^3 \mid 135 \le x_1 \le 175, 5 \le x_2 \le 15, 24 \le x_3 \le 30\}$, where f is the absorbency of oil in cookies, x_1 the frying temperature, x_2 the frying time, x_3-moisture.

This is formulated as a convex maximization problem and its solution is $x_1 = 175, x_2 = 5, x_3 = 30$.

Problem 15 (Ruvinshtein and Balkova [22]). $f(x) = -0.7x_1^2 - 0.17x_2^2 - 1.38x_3^2 + 0.38x_4^2 + 0.25x_5^2 - 0.26x_1x_2 + 0.57x_1x_3 + 0.19x_1x_4 + 1.95x_1x_5 + 0.16x_2x_3 + 0.89x_2x_4 + 0.86x_2x_5 + 1.3x_3x_4 + 1.4x_3x_5 + 0.55x_4x_5 - 0.61x_1 + 0.03x_2 + 0.23x_3 - 0.85x_4 - 0.48x_5 + 86.17 \longrightarrow \max,\ x \in D,$
$D = \{x \in R^5 \mid -2 \le x_i \le 2,\ i = 1, 2, \ldots, 5\}$, where f is the concentration, x_1 the duration of powdering process, x_2 the amount of butylic xanthogenate, x_3 the duration of flotation, x_4 the amount of sodium sulfite, x_5 pH pulti.

This is an indefinite programming problem and its solution in coded variables is $x_1 = 2, x_2 = 2, x_3 = 2, x_4 = 2, x_5 = 2$.

6 Conclusion

We carried out the analysis of the response surface problems using the general quadratic programming. The proposed algorithms converge to a global solution in a finite number of steps and were numerically tested on real response surface engineering problems.

References

1. Adler, U.P., Markova, E.V., Granovskii, U.V.: Design of Experiments for Search of Optimality Conditions, Nauka, Moscow (1976)
2. Ahnarazov, S.L., Kafarov, V.V.: Optimization Methods of Experiments in Chemical Technology, Bishaya Shkola, Moscow (1985)
3. Anderson, V.L., McLean, R.A.: Design of Experiments, Marcel Dekker, New-York, Ny (1974)
4. Asaturyan, B.I.: Theory of Design of Experiments, Radio and Net, Moscow (1983)
5. Atkinson, A.C., Bogacka, B., Zhigljavsky, A., (Eds.): Optimum Design 2000, Kluwer, Dordrecht, Boston MA (2001)
6. Atkinson, A.C., Donev, A.N.: Optimum Experimental Designs, Oxford University Press, Oxford (1992)
7. Bazarsad, Y., Enkhtuya, D., Enkhbat, R.: Optimization of a technological process of wool strings. Sci. J. Mongolian Tech. Univ. 1(16), 54–60 (1994)
8. Boer, E.P.J., Hendrih, E.M.T.: Global optimization problems in optimal design of experiments in regression models. J. Global Optim. 18, 385–398 (2000)
9. Chimedochir, S., Enkhbat, R.: Optimization of a technological process of production of fruits's (chazargan) oil. Sci. J. Mongolian Tech. Univ. 1(23), 91–98 (1996)
10. Enkhbat, R.: An algorithm for maximizing a convex function over a simple set. J. Global Optim. 8, 379–391 (1996)
11. Enkhbat, R., Chuluunhuyag, S.: Mathematical model of a technological process of reducing of amount of iron in water. Sci. J. Mongolian Inst. Waterpolicy 1, 68–76 (1996)
12. Enkhbat, R., Kamada, M., Bazarsad, Y.: A finite method for quadratic programming. J. Mongolian Math. Soc. 2, 12–30 (1998)

13. Ermakov, S.M., Zhigljavsky, A.A.: Mathematical Theory of Optimal Experiments, Nauka, Moscow (1987)
14. Fedorov, V.V.: Theory of Optimal Experiments, Academic, New York, NY (1972)
15. Fisher, R.A.: Design of Experiments, Hafner, New York, NY (1966)
16. Fukelsheim, F.: Optimal Design of Experiments, Wiley, New York, NY (1993)
17. Gorskii, B.G., Adler, U.P., Talalai, A.M.: Design of Experiments in Industries, Metallurgy, Moscow (1978)
18. Hill, W.J., Hunter, W.G.: Response Surface Methodology, Technical Report No.62, University of Wisconsin, Madison, WI (1966)
19. Horst, R., Tuy, H.: Global Optimization, Springer, New York, NY (1990)
20. Myers, R.H.: Response Surface Methodology, Allyn and Bacon, Boston, MA (1971)
21. Rockafellar, R.T.: Convex Analysis, Princeton University Press, Princeton, NJ (1970)
22. Ruvinshtein, U.B., Bolkova, L.A.: Mathematical Methods for Extraction of Treasures of the Soil, Nedra, Moscow (1987)
23. Sebostyanov, A.G., Sebostyanov, P.A.: Optimization of Mechanical and Technological Processes of Textile Industries, Nedra, Moscow (1991)
24. Tsetsgee, D.: Optimization of Frying Process of National Cookies. PhD Thesis, Mongolian Technical University, Mongolia (1997)
25. Vasiliev, O.V.: Optimization Methods, World Federation Publishers, Atlanta, GA (1996)

Canonical Dual Solutions for Fixed Cost Quadratic Programs

David Yang Gao[1], Ning Ruan[2], and Hanif D. Sherali[2]

[1] Graduate School of Information Technology and Mathematical Sciences, University of Ballarat, Mt Helen, Victoria 3353, Australia
d.gao@ballarat.edu.au
[2] Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA
hanifs@vt.edu

Summary. This chapter presents a canonical dual approach for solving a mixed-integer quadratic minimization problem with fixed cost terms. We show that this well-known NP-hard problem in $\mathbb{R}^{2n}$ can be transformed into a continuous concave maximization dual problem over a convex feasible subset of $\mathbb{R}^n$ with zero duality gap. The resulting canonical dual problem can be solved easily, under certain conditions, by traditional convex programming methods. Both existence and uniqueness of global optimal solutions are discussed. Application to a decoupled mixed-integer problem is illustrated and analytic solutions for both a global minimizer and a global maximizer are obtained. Examples for both decoupled and general nonconvex problems are presented. Furthermore, we discuss connections between the proposed canonical duality theory approach and the classical Lagrangian duality approach. An open problem is proposed for future study.

Key words: canonical duality, Lagrangian duality, global optimization, mixed-integer programming, fixed-charge objective function

1 Primal Problem and Motivation

In this chapter, we address the following quadratic, mixed-integer fixed-charge problem:

$$(\mathcal{P}_\flat): \quad \min\left\{P(\mathbf{x},\mathbf{v}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} + \mathbf{c}^T\mathbf{x} - \mathbf{f}^T\mathbf{v} \;\mid\; (\mathbf{x},\mathbf{v}) \in \mathcal{X}_v\right\}, \qquad (1)$$

where $A = A^T \in \mathbb{R}^{n\times n}$ is a given (generally indefinite) matrix, $\mathbf{c},\ \mathbf{f} \in \mathbb{R}^n$ are given vectors, the binary variable vector $\mathbf{v} \in \{0,1\}^n$ represents fixed cost variables, and the feasible space $\mathcal{X}_v$ is defined by

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_7,

$$\mathcal{X}_v = \{(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^n \times \{0,1\}^n \mid -\mathbf{v} \leq \mathbf{x} \leq \mathbf{v}\}. \tag{2}$$

Problem $(\mathcal{P}_\flat)$ arises in mathematical economics, facility location, and lot-sizing application contexts [1, 5, 32], where the constraints of the form $\mathbf{x} \in [-\mathbf{v}, \mathbf{v}]$ with $\mathbf{v} \in \{0,1\}^n$ are referred to as *fixed-charge constraints* [39]. These types of constraints have received a great deal of attention in the integer programming literature and many different types of valid inequalities have been developed to deal with this structure (see, for instance, [4, 34, 39]). Since we do not assume that the matrix A is positive semidefinite, the problem remains NP-hard, even with all the fixed cost variables v_i $(i = 1, \ldots, n)$ fixed to one [36, 38, 40, 41]. In order to numerically solve the latter continuous, box-constrained quadratic program, many effective methods have been developed [2, 3, 6, 9–12, 18, 22, 33, 35, 42–44]. Naturally, the problem becomes even more challenging with the addition of the fixed-charge feature.

Canonical duality theory, as developed in [15–17], is a potentially powerful tool for solving general continuous and discrete problems in nonconvex and global optimization. This theory is also called the *pure complementary variational principle* in continuum mechanics and physics [37], where it was originally proposed by Gao and Strang for nonlinear variational/boundary value problems in 1989 [30]. Recently, by using this theory, *perfect dual problems* (with zero duality gap) have been formulated for a class of nonconvex polynomial minimization problems with box and integer constraints [7, 19, 21, 27]. These results exhibit how such nonconvex and discrete minimization problems can be converted into continuous concave maximization dual problems. Under certain conditions, these canonical dual problems can be solved easily to obtain global minimizers of the underlying primal problems.

The main purpose of this chapter is to present a canonical duality approach for solving the fixed-charged problem (1). The chapter is organized as follows. In Section 2, a canonical dual problem is presented, which is equivalent to the primal problem in the sense that they have the same set of KKT points, where these KKT points for the discrete problem are defined with respect to a derived equivalent continuous problem. Connections of the derived dual with the Lagrangian dual under similar transformations are also discussed. The extremality conditions of these KKT solutions are explicitly specified in Section 3. Both existence and uniqueness of solutions are discussed in Section 4 and an illustrative example is presented in Section 5. Finally, certain concluding remarks and open problems are given in Section 6.

2 Canonical Dual Problem

In order to formulate a canonical dual problem for (1) that exhibits a zero duality gap, the key step is to rewrite the box constraints $-\mathbf{v} \leq \mathbf{x} \leq \mathbf{v}$, $\mathbf{v} \in \{0,1\}^n$ in the (relaxed) quadratic form [7, 21]:

$$\mathbf{x} \circ \mathbf{x} \leq \mathbf{v}, \ \mathbf{v} \circ (\mathbf{v} - \mathbf{e}) \leq 0, \tag{3}$$

where $\mathbf{e} = \{1\}^n$ is an n-vector of all ones and the notation $\mathbf{x} \circ \mathbf{v} := (x_1 v_1, x_2 v_2, \ldots, x_n v_n)$ denotes the Hadamard product for any two vectors $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$. Accordingly, consider the following (relaxed) reformulation of the primal problem $(\mathcal{P}_\flat)$:

$$(\mathcal{P}_\mathrm{r}): \quad \min \left\{ P(\mathbf{x}, \mathbf{v}) = \frac{1}{2}\mathbf{x}^T A \mathbf{x} + \mathbf{c}^T \mathbf{x} - \mathbf{f}^T \mathbf{v} \ : \ \mathbf{x} \circ \mathbf{x} \leq \mathbf{v}, \mathbf{v} \circ (\mathbf{v} - \mathbf{e}) \leq 0 \right\}. \tag{4}$$

Introducing a nonlinear transformation (i.e., the so-called *geometrical mapping*):

$$\mathbf{y} = \Lambda(\mathbf{x}, \mathbf{v}) = \begin{pmatrix} \boldsymbol{\epsilon}(\mathbf{x}) \\ \boldsymbol{\xi}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{x} \circ \mathbf{x} - \mathbf{v} \\ \mathbf{v} \circ \mathbf{v} - \mathbf{v} \end{pmatrix} \in \mathbb{R}^{2n},$$

the constraints (3) can be replaced identically by $\Lambda(\mathbf{x}, \mathbf{v}) \leq 0$. Let

$$V(\mathbf{y}) = \begin{cases} 0 & \text{if } \ \mathbf{y} \leq 0 \in \mathbb{R}^{2n} \\ +\infty & \text{otherwise} \end{cases}$$

and let $\mathbf{y}^* = \begin{pmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\tau} \end{pmatrix} \in \mathbb{R}^{2n}$ be the vector of dual variables associated with the corresponding restrictions $\mathbf{y} \leq 0$. The sup-Fenchel conjugate of $V(\mathbf{y})$ can be defined by

$$\begin{aligned} V^\sharp(\mathbf{y}^*) &= \sup_{\mathbf{y} \in \mathbb{R}^{2n}} \{\langle \mathbf{y}, \mathbf{y}^* \rangle - V(\mathbf{y})\} \\ &= \sup_{\boldsymbol{\epsilon} \in \mathbb{R}^n} \sup_{\boldsymbol{\xi} \in \mathbb{R}^n} \{\boldsymbol{\epsilon}^T \boldsymbol{\sigma} + \boldsymbol{\xi}^T \boldsymbol{\tau} - V(\mathbf{y})\} \\ &= \begin{cases} 0 & \text{if } \ \boldsymbol{\sigma} \geq 0 \in \mathbb{R}^n, \quad \boldsymbol{\tau} \geq 0 \in \mathbb{R}^n, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned}$$

By the theory of convex analysis, the following extended canonical duality relations holds:

$$\mathbf{y}^* \in \partial V(\mathbf{y}) \quad \Leftrightarrow \quad \mathbf{y} \in \partial V^\sharp(\mathbf{y}^*) \quad \Leftrightarrow \quad V(\mathbf{y}) + V^\sharp(\mathbf{y}^*) = \mathbf{y}^T \mathbf{y}^*, \tag{5}$$

or equivalently

$$\boldsymbol{\epsilon} \leq 0 \quad \Leftrightarrow \quad \boldsymbol{\sigma} \geq 0 \quad \Leftrightarrow \quad \boldsymbol{\epsilon}^T \boldsymbol{\sigma} = 0, \tag{6}$$

$$\boldsymbol{\xi} \leq 0 \quad \Leftrightarrow \quad \boldsymbol{\tau} \geq 0 \quad \Leftrightarrow \quad \boldsymbol{\xi}^T \boldsymbol{\tau} = 0. \tag{7}$$

Observe that the complementarity condition $\boldsymbol{\xi}^T \boldsymbol{\tau} = \boldsymbol{\tau}^T (\mathbf{v} \circ \mathbf{v} - \mathbf{v}) = 0$ in (7) leads to the integrality condition $\mathbf{v} \circ \mathbf{v} - \mathbf{v} = 0 \ \ \forall \boldsymbol{\tau} > 0$.

Letting $U(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T A \mathbf{x} - \mathbf{c}^T \mathbf{x} + \mathbf{f}^T \mathbf{v}$, the relaxed primal problem $(\mathcal{P}_\mathrm{r})$ can be written in the following unconstrained canonical form [17]:

$$(\mathcal{P}_\mathrm{c}): \ \min \{\Pi(\mathbf{x}, \mathbf{v}) = V(\Lambda(\mathbf{x}, \mathbf{v})) - U(\mathbf{x}, \mathbf{v}) \mid \ \mathbf{x} \in \mathbb{R}^n, \ \ \mathbf{v} \in \mathbb{R}^n\}. \tag{8}$$

Following the original idea of Gao and Strang [30], we replace $V(\Lambda(\mathbf{x},\mathbf{v}))$ in (8) by the Fenchel–Young equality $V(\Lambda(\mathbf{x},\mathbf{v})) = \Lambda(\mathbf{x},\mathbf{v})^T\mathbf{y}^* - V^\sharp(\mathbf{y}^*)$. Then the *total complementary function* $\Xi(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau}) : \mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}\cup\{-\infty\}$ associated with the problem $(\mathcal{P}_{\rm c})$ can be defined as below:

$$\begin{aligned}\Xi(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau}) &= \Lambda(\mathbf{x},\mathbf{v})^T\mathbf{y}^* - V^\sharp(\mathbf{y}^*) - U(\mathbf{x},\mathbf{v})\\ &= \frac{1}{2}\mathbf{x}^T G(\boldsymbol{\sigma})\mathbf{x} + \mathbf{c}^T\mathbf{x} + \mathbf{v}^T\mathrm{Diag}\,(\boldsymbol{\tau})\mathbf{v} - (\mathbf{f}+\boldsymbol{\sigma}+\boldsymbol{\tau})^T\mathbf{v}\\ &\quad -V^\sharp(\mathbf{y}^*),\end{aligned} \tag{9}$$

where

$$G(\boldsymbol{\sigma}) = A + 2\mathrm{Diag}(\boldsymbol{\sigma}), \tag{10}$$

and where the notation $\mathrm{Diag}(\boldsymbol{\sigma})$ stands for a diagonal matrix with σ_i, $i = 1,...,n$, being its diagonal elements. By this complementary function, the canonical dual function $\Pi^{\rm d} : \mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}\cup\{-\infty\}$ can be obtained by

$$\Pi^{\rm d}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \mathrm{sta}\{\Xi(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau}) \mid \mathbf{x}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^n\} = U^\Lambda(\boldsymbol{\sigma},\boldsymbol{\tau}) - V^\sharp(\boldsymbol{\sigma},\boldsymbol{\tau}), \tag{11}$$

where $U^\Lambda(\boldsymbol{\sigma},\boldsymbol{\tau})$ is the Λ-conjugate transformation defined by

$$U^\Lambda(\boldsymbol{\sigma},\boldsymbol{\tau}) = \mathrm{sta}\{\Lambda(\mathbf{x},\mathbf{v})^T\mathbf{y}^* - U(\mathbf{x},\mathbf{v})\mid \mathbf{x}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^n\}. \tag{12}$$

Accordingly, introducing a dual feasible space

$$\mathcal{S}_\sharp = \{(\boldsymbol{\sigma},\boldsymbol{\tau})\in\mathbb{R}^n\times\mathbb{R}^n \mid \ \boldsymbol{\sigma}\geq 0,\ \ \boldsymbol{\tau} > 0,\ \ \mathbf{c}\in\mathcal{C}_{\rm ol}(G(\boldsymbol{\sigma}))\}, \tag{13}$$

where $\mathcal{C}_{\rm ol}(G)$ denotes the column space of G (i.e., a vector space spanned by the columns of the matrix G), the canonical dual function can be formulated as

$$\Pi^{\rm d}(\boldsymbol{\sigma},\boldsymbol{\tau}) = U^\Lambda(\boldsymbol{\sigma},\boldsymbol{\tau}) = -\frac{1}{2}\mathbf{c}^T G^+(\boldsymbol{\sigma})\mathbf{c} - \frac{1}{4}\sum_{i=1}^n \frac{1}{\tau_i}(f_i+\sigma_i+\tau_i)^2 \ \ \forall(\boldsymbol{\sigma},\boldsymbol{\tau})\in\mathcal{S}_\sharp, \tag{14}$$

where G^+ denotes the Moore–Penrose generalized inverse of G. Denoting

$$P^{\rm d}(\boldsymbol{\sigma},\boldsymbol{\tau}) = -\frac{1}{2}\mathbf{c}^T G^+(\boldsymbol{\sigma})\mathbf{c} - \frac{1}{4}\sum_{i=1}^n \frac{1}{\tau_i}(f_i+\sigma_i+\tau_i)^2 \ : \ \mathcal{S}_\sharp\to\mathbb{R}, \tag{15}$$

the dual to $(\mathcal{P}_\flat)$ can then be stated as the following:

$$(\mathcal{P}^\sharp): \quad \max\left\{P^{\rm d}(\boldsymbol{\sigma},\boldsymbol{\tau}) \ = -\frac{1}{2}\mathbf{c}^T G^+(\boldsymbol{\sigma})\mathbf{c} - \frac{1}{4}\sum_{i=1}^n \frac{1}{\tau_i}(f_i+\sigma_i+\tau_i)^2 \mid \right.$$
$$\left.(\boldsymbol{\sigma},\boldsymbol{\tau})\in\mathcal{S}_\sharp\right\}. \tag{16}$$

For any given n-vectors $\mathbf{t} = \{t_i\}^n$ and $\mathbf{s} = \{s_i\}^n$, we denote $\mathbf{t}\oslash\mathbf{s} = \{t_i/s_i\}^n$.

Theorem 1 (Complementary Dual Principle). *Problem* $(\mathcal{P}^\sharp)$ *is canonically (i.e., perfectly) dual to the primal problem* $(\mathcal{P}_\flat)$ *in the sense that if* $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_\sharp$ *is a KKT point of* $(\mathcal{P}^\sharp)$*, then the vector* $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ *defined by*

$$\bar{\mathbf{x}} = -G^+(\bar{\boldsymbol{\sigma}})\mathbf{c}, \tag{17}$$

$$\bar{\mathbf{v}} = \frac{1}{2}(\mathbf{f} + \bar{\boldsymbol{\sigma}} + \bar{\boldsymbol{\tau}}) \oslash \bar{\boldsymbol{\tau}} \tag{18}$$

is feasible to the primal problem $(\mathcal{P}_\flat)$*, and*

$$P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \Xi(\bar{\mathbf{x}}, \bar{\mathbf{v}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = P^d(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}). \tag{19}$$

Proof. By introducing Lagrange multipliers $(\boldsymbol{\epsilon}, \boldsymbol{\xi}) \in \mathbb{R}^n_- \times \mathbb{R}^n_-$ (where $\mathbb{R}^n_-$ is the nonpositive orthant of $\mathbb{R}^n$) associated with the respective inequalities in (13), the Lagrangian $\Theta : \mathcal{S}_\sharp \times \mathbb{R}^n_- \times \mathbb{R}^n_- \to \mathbb{R}$ for problem $(\mathcal{P}^\sharp)$ is given by

$$\Theta(\boldsymbol{\sigma}, \boldsymbol{\tau}, \boldsymbol{\epsilon}, \boldsymbol{\xi}) = P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) - \boldsymbol{\epsilon}^T\boldsymbol{\sigma} - \boldsymbol{\xi}^T\boldsymbol{\tau}. \tag{20}$$

It is easy to prove that the criticality conditions

$$\nabla_{\boldsymbol{\sigma}}\Theta(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}, \boldsymbol{\epsilon}, \boldsymbol{\xi}) = 0, \;\; \nabla_{\boldsymbol{\tau}}\Theta(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}, \boldsymbol{\epsilon}, \boldsymbol{\xi}) = 0$$

lead to

$$\boldsymbol{\epsilon} = \nabla_{\boldsymbol{\sigma}} P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \bar{\mathbf{x}}(\bar{\boldsymbol{\sigma}}) \circ \bar{\mathbf{x}}(\bar{\boldsymbol{\sigma}}) - \bar{\mathbf{v}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}), \tag{21}$$

$$\boldsymbol{\xi} = \nabla_{\boldsymbol{\tau}} P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \bar{\mathbf{v}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \circ \bar{\mathbf{v}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) - \bar{\mathbf{v}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}), \tag{22}$$

and the accompanying KKT conditions include

$$0 \le \bar{\boldsymbol{\sigma}} \;\perp\; \boldsymbol{\epsilon} \le 0, \tag{23}$$

$$0 < \bar{\boldsymbol{\tau}} \;\perp\; \boldsymbol{\xi} \le 0, \tag{24}$$

where $\bar{\mathbf{x}}(\bar{\boldsymbol{\sigma}}) = -G^+(\bar{\boldsymbol{\sigma}})\mathbf{c}$ and $\bar{\mathbf{v}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = \frac{1}{2}(\mathbf{f} + \bar{\boldsymbol{\sigma}} + \bar{\boldsymbol{\tau}}) \oslash \bar{\boldsymbol{\tau}}$. By the strictly inequality condition $\bar{\boldsymbol{\tau}} > 0$, the complementarity condition $\bar{\boldsymbol{\tau}}^T(\bar{\mathbf{v}} \circ \bar{\mathbf{v}} - \bar{\mathbf{v}}) = 0$ in (24) leads to the integrality condition $(\bar{\mathbf{v}} \circ \bar{\mathbf{v}} - \bar{\mathbf{v}}) = 0$. This shows that if $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ is a KKT point of the problem $(\mathcal{P}^\sharp)$, then $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ is feasible to the discrete primal problem $(\mathcal{P}_\flat)$, and moreover, by

Using (17) and (18), we have

$$\begin{aligned} P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) &= \frac{1}{2}\mathbf{c}^T G^+(\bar{\boldsymbol{\sigma}})\mathbf{c} - \mathbf{c}^T G^+(\bar{\boldsymbol{\sigma}})\mathbf{c} - 2\bar{\mathbf{v}}^T \mathrm{Diag}\,(\bar{\boldsymbol{\tau}})\bar{\mathbf{v}} + \bar{\mathbf{v}}^T \mathrm{Diag}\,(\bar{\boldsymbol{\tau}})\bar{\mathbf{v}} \\ &= \frac{1}{2}\bar{\mathbf{x}}^T A\bar{\mathbf{x}} + \bar{\mathbf{x}}^T \mathrm{Diag}\,(\bar{\boldsymbol{\sigma}})\bar{\mathbf{x}} + \mathbf{c}^T\bar{\mathbf{x}} - \bar{\mathbf{v}}^T(\bar{\boldsymbol{\sigma}} + \bar{\boldsymbol{\tau}} + \mathbf{f}) + \bar{\boldsymbol{\tau}}^T(\bar{\mathbf{v}} \circ \bar{\mathbf{v}}) \\ &= \Xi(\bar{\mathbf{x}}, \bar{\mathbf{v}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) + \bar{\boldsymbol{\sigma}}^T(\bar{\mathbf{x}} \circ \bar{\mathbf{x}} - \bar{\mathbf{v}}) + \bar{\boldsymbol{\tau}}^T(\bar{\mathbf{v}} \circ \bar{\mathbf{v}} - \bar{\mathbf{v}}) \\ &= P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) \end{aligned}$$

due to the complementarity conditions (23) and (24). This proves the theorem. □

In order to understand the canonical duality theory and its relation to the nonlinear Lagrangian duality theory, we have the following remark.

Remark 1 (Connections with Lagrangian duality). Note that by replacing the linear inequality constraints and the integer constraints in $\mathcal{X}_v$ with the quadratic forms in (3), where the second set of inequalities is written as equality restrictions, the primal problem $(\mathcal{P}_\flat)$ can be equivalently reformulated as the continuous programming problem:

$$(\mathcal{P}_b): \quad \min\left\{P(\mathbf{x},\mathbf{v})=\frac{1}{2}\mathbf{x}^TA\mathbf{x}+\mathbf{c}^T\mathbf{x}-\mathbf{f}^T\mathbf{v}:\ \mathbf{x}\circ\mathbf{x}\le\mathbf{v},\ \mathbf{v}\circ(\mathbf{v}-\mathbf{e})=0\right\}. \tag{25}$$

Introducing the Lagrange multipliers $\boldsymbol{\sigma}\ge 0$ and $\boldsymbol{\tau}\in\mathbb{R}^n$ to relax the inequality constraint $\mathbf{v}-\mathbf{x}\circ\mathbf{x}\ge 0$ and the equality constraint $\mathbf{v}\circ\mathbf{e}-\mathbf{v}\circ\mathbf{v}=0$ in (25), respectively, the Lagrangian associated with the reformulated problem (25) can be defined as follows:

$$L(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau})=P(\mathbf{x},\mathbf{v})+\sum_{i=1}^{n}[\sigma_i(x_i^2-v_i)+\tau_i(v_i^2-v_i)]. \tag{26}$$

The corresponding Lagrangian dual function is given by

$$P^*(\boldsymbol{\sigma},\boldsymbol{\tau})=\inf\left\{L(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau}):\ \ (\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}\right\}. \tag{27}$$

Now, observe that when $\boldsymbol{\tau}_i\le 0$ for any $i=\{1,\dots,n\}$, the separable minimization over $\mathbf{v}$ in (27) leads to $P^*(\boldsymbol{\sigma},\boldsymbol{\tau})=-\infty$. Hence, since we wish to maximize $P^*(\boldsymbol{\sigma},\boldsymbol{\tau})$ in the Lagrangian dual problem, we can restrict $\boldsymbol{\tau}>0$. Consequently, we obtain the following criticality conditions for (27):

$$\nabla_{\mathbf{x}}L(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau})=A\mathbf{x}+\mathbf{c}+2[\text{Diag }(\boldsymbol{\sigma})]\mathbf{x}=0, \tag{28}$$

$$\nabla_{\mathbf{v}}L(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau})=-\mathbf{f}-\boldsymbol{\sigma}-\boldsymbol{\tau}+2[\text{Diag }(\boldsymbol{\tau})]\mathbf{v}=0. \tag{29}$$

Defining $G(\boldsymbol{\sigma})$ as in (10), we have from (28) that so long as $\mathbf{c}\in\mathcal{C}_{\text{ol}}(G(\boldsymbol{\sigma}))$, we get

$$\mathbf{x}(\boldsymbol{\sigma})=-G^+(\boldsymbol{\sigma})\mathbf{c}. \tag{30}$$

Furthermore, under the foregoing restriction $\boldsymbol{\tau}>0$, we get from (29) that

$$\mathbf{v}(\boldsymbol{\sigma})=\frac{1}{2}(\mathbf{f}+\boldsymbol{\sigma}+\boldsymbol{\tau})\oslash\boldsymbol{\tau}. \tag{31}$$

Therefore, defining $\mathcal{S}_\sharp$ as in (13), and substituting (30) and (31) into (26), the Lagrangian dual function (27) can be reformulated as follows, where $P^{\text{d}}(\boldsymbol{\sigma},\boldsymbol{\tau})$ is given by (15), identically as for the the canonical dual derivation:

$$P^*(\boldsymbol{\sigma},\boldsymbol{\tau})=\inf_{(\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}}L(\mathbf{x},\mathbf{v},\boldsymbol{\sigma},\boldsymbol{\tau})=\begin{cases}P^{\text{d}}(\boldsymbol{\sigma},\boldsymbol{\tau}) & \text{if } (\boldsymbol{\sigma},\boldsymbol{\tau})\in\mathcal{S}_\sharp,\\ -\infty & \text{otherwise.}\end{cases}$$

Therefore, the reformulated Lagrangian dual problem is given precisely by the canonical dual problem $(\mathcal{P}^\sharp)$ stated in (16).

The key to achieving this equivalence is the appropriate transformation (geometrical mapping) of the constraints into the quadratic form (3), or as in (25), and the canonical duality relations (5), which is prompted by the canonical duality approach. A detailed study on the geometrical mapping and the canonical duality relations, i.e., the so-called *constitutive laws*, appears in [15].

Remark 2. Theorem 1 shows that by the canonical duality theory, the NP-hard discrete primal problem $(\mathcal{P}_\flat)$ is actually equivalent to a continuous dual problem $(\mathcal{P}^\sharp)$ with zero duality gap. In many applications, if $G(\bar{\boldsymbol{\sigma}})$ is invertible, then the KKT point $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ of the canonical dual problem $(\mathcal{P}^\sharp)$ is a critical point of the canonical dual function $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$. If we want to find all extrema (both local minima and maxima) of the nonconvex function $P(\mathbf{x}, \mathbf{v})$ on $\mathcal{X}_v$, the constraints in $\mathcal{S}_\sharp$ can be ignored (the inequalities $\boldsymbol{\sigma} \geq 0$ and $\boldsymbol{\tau} > 0$ are constraints only for the minimization problem $(\mathcal{P}_\flat)$), i.e., for each critical point $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}})$ of the canonical dual function $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$, the vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ defined by (17) and (18) is a local extremum of the nonconvex function $P(\mathbf{x}, \mathbf{v})$ on $\mathcal{X}_v$. Particularly, for the following co-primal problem

$$(\mathcal{P}_\sharp): \quad \max\{P(\mathbf{x}, \mathbf{v}) | \ (\mathbf{x}, \mathbf{v}) \in \mathcal{X}_v\} \tag{32}$$

the associated canonical dual problem is

$$(\mathcal{P}^\flat): \quad \min\{P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) | \ (\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_\flat\}, \tag{33}$$

where

$$\mathcal{S}_\flat = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^n \times \mathbb{R}^n | \ \boldsymbol{\sigma} \leq 0, \ \boldsymbol{\tau} < 0, \quad \mathbf{c} \in \mathcal{C}_{\mathrm{col}}(G(\boldsymbol{\sigma}))\}. \tag{34}$$

Parallel to Theorem 1, we have similar canonical duality results for problems $(\mathcal{P}_\sharp)$ and $(\mathcal{P}^\flat)$.

The extremality conditions will be studied in the next section.

3 Global Optimality Criteria

In this section, we present certain global optimality conditions for the nonconvex problem $(\mathcal{P}_\flat)$. We first introduce some useful feasible spaces:

$$\mathcal{S}_\sharp^+ = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \ \boldsymbol{\sigma} \geq 0, \ \boldsymbol{\tau} > 0, \ \ G(\boldsymbol{\sigma}) \succ 0\}, \tag{35}$$

$$\mathcal{S}_\flat^- = \{(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \ \boldsymbol{\sigma} \leq 0, \ \boldsymbol{\tau} < 0, \ \ G(\boldsymbol{\sigma}) \prec 0\}. \tag{36}$$

By the *triality theory* developed in [15], we have the following results, where $\mathbf{y}^* = (\boldsymbol{\sigma}, \boldsymbol{\tau})$.

Theorem 2. *Suppose that the vector* $\bar{\mathbf{y}}^* = (\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_\sharp^+ \cup \mathcal{S}_\flat^-$ *is a critical point of the dual function* $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$. *Let* $(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \big(-G^{-1}(\bar{\boldsymbol{\sigma}})\mathbf{c},\ \frac{1}{2}(\mathbf{f} + \bar{\boldsymbol{\sigma}} + \bar{\boldsymbol{\tau}}) \oslash \bar{\boldsymbol{\tau}}\big)$.

If $\bar{\mathbf{y}}^* \in \mathcal{S}_\sharp^+$, *then* $\bar{\mathbf{y}}^*$ *is a global maximizer of* P^{d} *on* $\mathcal{S}_\sharp^+$, *the vector* $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ *is a global minimizer of* P *on* $\mathcal{X}_v$, *and*

$$P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \min_{(\mathbf{x},\mathbf{v}) \in \mathcal{X}_v} P(\mathbf{x}, \mathbf{v}) = \max_{(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\sharp^+} P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}). \tag{37}$$

If $\bar{\mathbf{y}}^* \in \mathcal{S}_\flat^-$, *then* $\bar{\mathbf{y}}^*$ *is a global minimizer of* P^{d} *on* $\mathcal{S}_\flat^-$, *the vector* $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ *is a global maximizer of* P *on* $\mathcal{X}_v$, *and*

$$P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \max_{(\mathbf{x},\mathbf{v}) \in \mathcal{X}_v} P(\mathbf{x}, \mathbf{v}) = \min_{(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\flat^-} P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}). \tag{38}$$

Proof. By Theorem 1 and the general results developed in [15] we know that if the vector $\bar{\mathbf{y}}^*$ is a critical point of problem $(\mathcal{P}^\sharp)$, then the vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ defined by (17) and (18) is a feasible solution to problem $(\mathcal{P}_\flat)$, and

$$P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \Xi(\bar{\mathbf{x}}, \bar{\mathbf{v}}, \bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}).$$

By the fact that the canonical dual function $P^{\mathrm{d}}(\mathbf{y}^*)$ is concave on $\mathcal{S}_\sharp^+$, the critical point $\bar{\mathbf{y}}^* \in \mathcal{S}_\sharp^+$ is a global maximizer of $P^{\mathrm{d}}(\mathbf{y}^*)$ over $\mathcal{S}_\sharp^+$, and $(\bar{\mathbf{x}}, \bar{\mathbf{v}}, \bar{\mathbf{y}}^*)$ is a saddle point of the total complementary function $\Xi(\mathbf{x}, \mathbf{v}, \mathbf{y}^*)$ on $\mathbb{R}^{2n} \times \mathcal{S}_\sharp^+$, i.e., Ξ is convex in $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^{2n} = \mathbb{R}^n \times \mathbb{R}^n$ and concave in $\mathbf{y}^* \in \mathcal{S}_\sharp^+$. Thus, by the (right) saddle min–max duality theory (see [15]), we have

$$\begin{aligned}
P^{\mathrm{d}}(\bar{\mathbf{y}}^*) &= \max_{\mathbf{y}^* \in \mathcal{S}_\sharp^+} P^{\mathrm{d}}(\mathbf{y}^*) = \max_{\mathbf{y}^* \in \mathcal{S}_\sharp^+} \min_{(\mathbf{x},\mathbf{v}) \in \mathbb{R}^{2n}} \Xi(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) \\
&= \min_{(\mathbf{x},\mathbf{v}) \in \mathbb{R}^{2n}} \max_{\mathbf{y}^* \in \mathcal{S}_\sharp^+} \Xi(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) \\
&= \min_{(\mathbf{x},\mathbf{v}) \in \mathbb{R}^{2n}} \left\{ P(\mathbf{x}, \mathbf{v}) + \max_{(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\sharp^+} \left\{ (\mathbf{x} \circ \mathbf{x} - \mathbf{v})^T \boldsymbol{\sigma} + (\mathbf{v} \circ \mathbf{v} - \mathbf{v})^T \boldsymbol{\tau} \right\} \right\} \\
&= \min_{(\mathbf{x},\mathbf{v}) \in \mathbb{R}^{2n}} \left\{ P(\mathbf{x}, \mathbf{v}) + \max_{(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\sharp^+} \left\{ \Lambda(\mathbf{x}, \mathbf{v})^T \mathbf{y}^*(\boldsymbol{\sigma}, \boldsymbol{\tau}) - V^\sharp(\mathbf{y}^*(\boldsymbol{\sigma}, \boldsymbol{\tau})) \right\} \right\} \\
&= \min_{(\mathbf{x},\mathbf{v}) \in \mathcal{X}_v} P(\mathbf{x}, \mathbf{v}) = P(\bar{\mathbf{x}}, \bar{\mathbf{v}})
\end{aligned}$$

due to the fact that

$$\begin{aligned}
V(\Lambda(\mathbf{x}, \mathbf{v})) &= \sup_{\mathbf{y}^* \in \mathcal{S}_\sharp^+} \{ \Lambda(\mathbf{x}, \mathbf{v})^T \mathbf{y}^*(\boldsymbol{\sigma}, \boldsymbol{\tau}) - V^\sharp(\mathbf{y}^*(\boldsymbol{\sigma}, \boldsymbol{\tau})) \} \\
&= \begin{cases} 0 & \text{if } (\mathbf{x}, \mathbf{v}) \in \mathcal{X}_v, \\ +\infty & \text{otherwise.} \end{cases}
\end{aligned}$$

This proves statement (37).

In order to prove statement (38), we introduce the Fenchel inf-conjugate

$$V^\flat(\mathbf{y}^*) = \inf_{\mathbf{y}\in\mathbb{R}^{2n}} \{\mathbf{y}^T\mathbf{y}^* + V(\mathbf{y})\} = \begin{cases} 0 & \text{if } \mathbf{y}^* \le 0, \\ -\infty & \text{otherwise.} \end{cases} \tag{39}$$

Therefore, the total complementary function associated with the co-primal problem $(\mathcal{P}_\sharp)$ is

$$\Xi_\flat(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) = \Lambda(\mathbf{x}, \mathbf{v})^T\mathbf{y}^* - V^\flat(\mathbf{y}^*) - U(\mathbf{x}, \mathbf{v}), \tag{40}$$

which is a *left saddle function* (see [15, Section 1.6]) on $\mathbb{R}^{2n} \times \mathcal{S}_\flat^-$, i.e., $\Xi_\flat(\mathbf{x}, \mathbf{v}, \mathbf{y}^*)$ is concave in $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^{2n}$ and convex in $\mathcal{S}_\flat^-$. Thus, by the *left saddle min–max duality theory* (see [15]), we have

$$\begin{aligned} P^{\mathrm{d}}(\bar{\mathbf{y}}^*) &= \min_{\mathbf{y}^*\in\mathcal{S}_\flat^-} P^{\mathrm{d}}(\mathbf{y}^*) = \min_{\mathbf{y}^*\in\mathcal{S}_\flat^-} \max_{(\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}} \Xi(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) \\ &= \max_{(\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}} \min_{\mathbf{y}^*\in\mathcal{S}_\flat^-} \Xi(\mathbf{x}, \mathbf{v}, \mathbf{y}^*) \\ &= \max_{(\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}} \left\{ P(\mathbf{x}, \mathbf{v}) + \min_{(\boldsymbol{\sigma},\boldsymbol{\tau})\in\mathcal{S}_\flat^-} \left\{ (\mathbf{x}\circ\mathbf{x} - \mathbf{v})^T\boldsymbol{\sigma} + (\mathbf{v}\circ\mathbf{v} - \mathbf{v})^T\boldsymbol{\tau} \right\} \right\} \\ &= \max_{(\mathbf{x},\mathbf{v})\in\mathbb{R}^{2n}} \{P(\mathbf{x}, \mathbf{v}) + V(\Lambda(\mathbf{x}, \mathbf{v}))\} \\ &= \max_{(\mathbf{x},\mathbf{v})\in\mathcal{X}_v} P(\mathbf{x}, \mathbf{v}) = P(\bar{\mathbf{x}}, \bar{\mathbf{v}}) \end{aligned}$$

due to the fact that

$$\begin{aligned} V(\Lambda(\mathbf{x}, \mathbf{v})) &= \inf_{\mathbf{y}^*\in\mathcal{S}_\flat^-} \{\Lambda(\mathbf{x}, \mathbf{v})^T\mathbf{y}^* + V^\flat(\mathbf{y}^*)\} \\ &= \begin{cases} 0 & \text{if } (\mathbf{x}, \mathbf{v}) \in \mathcal{X}_v, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned}$$

This proves statement (38) and the theorem. □

Theorem 2 shows that the nonconvex quadratic mixed-integer minimization problem $(\mathcal{P}_\flat)$ is canonically dual to the following concave maximization problem:

$$(\mathcal{P}_+^\sharp): \quad \max \left\{ P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \ : \ (\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_\sharp^+ \right\}. \tag{41}$$

Since $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is a continuous concave function over a convex feasible space $\mathcal{S}_\sharp^+$, if $(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) \in \mathcal{S}_\sharp^+$ is a critical point of $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$, it must be a global maximizer of problem $(\mathcal{P}_+^\sharp)$, and the vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \left(-G^{-1}(\bar{\boldsymbol{\sigma}})\mathbf{c},\ \frac{1}{2}(\mathbf{f} + \bar{\boldsymbol{\sigma}} + \bar{\boldsymbol{\tau}}) \oslash \bar{\boldsymbol{\tau}}\right)$ is a global minimizer of problem $(\mathcal{P}_\flat)$. Particularly, for a fixed $\boldsymbol{\sigma}$, let

$$P^{\mathrm{g}}(\boldsymbol{\sigma}) = \max_{\boldsymbol{\tau}>0} P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = -\frac{1}{2}\mathbf{c}^T G^{-1}(\boldsymbol{\sigma})\mathbf{c} - \sum_{i=1}^{n} (f_i + \sigma_i)^+, \quad \boldsymbol{\sigma} \in \mathcal{S}_\sigma^+, \tag{42}$$

where $(t_i)^+ = \max\{t_i, 0\}$ and

$$\mathcal{S}_\sigma^+ = \{\boldsymbol{\sigma} \in \mathbb{R}^n | \ \boldsymbol{\sigma} \geq 0, \ \ \boldsymbol{\sigma} \neq -\mathbf{f}, \ \ G(\boldsymbol{\sigma}) \succ 0\}. \tag{43}$$

Furthermore, we denote $\boldsymbol{\delta}(\mathbf{t})^+ = \{\delta_i(t_i)^+\}^n \in \mathbb{R}^n$, where

$$\delta_i(t_i)^+ = \begin{cases} 1 \text{ if } t_i > 0, \\ 0 \text{ if } t_i < 0, \end{cases} \ i = 1, \ldots, n. \tag{44}$$

Then the canonical dual problem $(\mathcal{P}_+^\sharp)$ can be written in the following simple form:

$$(\mathcal{P}_+^{\mathrm{g}}): \ \ \max\left\{P^{\mathrm{g}}(\boldsymbol{\sigma}) \ : \ \ \boldsymbol{\sigma} \in \mathcal{S}_\sigma^+\right\}. \tag{45}$$

Theorem 3 (Analytic solution to $(\mathcal{P}_\flat)$**)**. *For the given $A \in \mathbb{R}^{n\times n}$ and $\mathbf{c}$, $\mathbf{f} \in \mathbb{R}^n$, if $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_\sigma^+$ is a critical point of $P^g(\boldsymbol{\sigma})$, then the vector*

$$(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = (-G^{-1}(\bar{\boldsymbol{\sigma}})\mathbf{c}, \ \boldsymbol{\delta}(\mathbf{f} + \bar{\boldsymbol{\sigma}})^+) \tag{46}$$

is a global minimizer of $(\mathcal{P}_\flat)$.

This theorem can be proved easily by using Theorem 2. Similar results on analytic solution to nonconvex variational/boundary value problems were originally obtained in [13, 14, 25]. In the next section we will study certain existence and uniqueness conditions for the canonical dual problem to have a critical point in $\mathcal{S}_\sigma^+$.

4 Existence and Uniqueness Criteria

Let

$$\partial\mathcal{S}_\sigma^+ = \{\boldsymbol{\sigma} \in \mathcal{S}_\sigma^+ | \ \ \det G(\boldsymbol{\sigma}) = 0\}. \tag{47}$$

Based on the recent results given in [27, 28], we have the following theorem:

Theorem 4 (Existence and uniqueness criteria**)**. *For a given matrix $A \in \mathbb{R}^{n\times n}$ and vectors $\mathbf{c}, \ \mathbf{f} \in \mathbb{R}^n$, if for any given $\boldsymbol{\sigma} \in \mathcal{S}_\sigma^+$,*

$$\lim_{\alpha\to 0^+} \mathbf{c}^T[G(\boldsymbol{\sigma}_o + \alpha\boldsymbol{\sigma})]^+\mathbf{c} = \infty \ \text{ and } \lim_{\alpha\to\infty} \mathbf{c}^T[G(\boldsymbol{\sigma}_o + \alpha\boldsymbol{\sigma})]^+\mathbf{c} \geq 0 \ \ \forall\boldsymbol{\sigma}_o \in \partial\mathcal{S}_\sigma^+, \tag{48}$$

then the canonical dual problem $(\mathcal{P}_+^g)$ has at least one critical point $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_\sigma^+$ and the vector

$$(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \left(-G^{-1}(\bar{\boldsymbol{\sigma}})\mathbf{c}, \ \ \boldsymbol{\delta}(\mathbf{f} + \bar{\boldsymbol{\sigma}})^+\right)$$

is a global optimizer of the primal problem $(\mathcal{P}_\flat)$. Moreover, if

$$c_i \neq 0 \ \ \bar{\sigma}_i + f_i \neq 0 \ \ \forall i = 1, \ldots, n, \tag{49}$$

then the vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ is a unique global minimizer of $(\mathcal{P}_\flat)$.

Proof. By the fact that, on $\mathcal{S}_\sigma^+$, we have

$$\frac{\partial G^{-1}(\boldsymbol{\sigma})}{\partial \sigma_k} = -G^{-1}(\boldsymbol{\sigma})\frac{\partial G(\boldsymbol{\sigma})}{\partial \sigma_k}G^{-1}(\boldsymbol{\sigma}),$$

the Hessian of the quadratic form $-\frac{1}{2}\mathbf{c}^T G^{-1}(\boldsymbol{\sigma})\mathbf{c}$ is

$$H_{1\sigma^2}(\boldsymbol{\sigma}) = \left\{-4x_i(\boldsymbol{\sigma})G_{ij}^{-1}(\boldsymbol{\sigma})x_j(\boldsymbol{\sigma})\right\}, \tag{50}$$

where $\mathbf{x}(\boldsymbol{\sigma}) = -G^{-1}(\boldsymbol{\sigma})\mathbf{c}$. Therefore, the Hessian matrix of the dual objective function P^{d} is

$$H(\boldsymbol{\sigma},\boldsymbol{\tau}) = \nabla^2 P^{\mathrm{d}}(\boldsymbol{\sigma},\boldsymbol{\tau}) = \begin{pmatrix} H_{1\sigma^2} + H_{2\sigma^2} & H_{\sigma\tau} \\ H_{\tau\sigma} & H_{\tau^2} \end{pmatrix},$$

where

$$\begin{aligned} H_{2\sigma^2} &= \mathrm{Diag}\left\{-\frac{1}{2\tau_i}\right\}, \\ H_{\sigma\tau} = H_{\tau\sigma} &= \mathrm{Diag}\left\{\frac{(\sigma_i + f_i)}{2\tau_i^2}\right\}, \\ H_{\tau^2} &= \mathrm{Diag}\left\{-\frac{(\sigma_i + f_i)^2}{2\tau_i^3}\right\}. \end{aligned}$$

It is clear that

$$H_{1\sigma^2}(\boldsymbol{\sigma}) \preceq 0,\ H_{2\sigma^2}(\boldsymbol{\tau}) \prec 0,\ \ H_{\tau^2}(\boldsymbol{\sigma},\boldsymbol{\tau}) \preceq 0\ \ \forall(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\sharp^+, \tag{51}$$

$$H_{1\sigma^2}(\boldsymbol{\sigma}) \succeq 0,\ H_{2\sigma^2}(\boldsymbol{\tau}) \succ 0,\ \ H_{\tau^2}(\boldsymbol{\sigma},\boldsymbol{\tau}) \succeq 0\ \ \forall(\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\flat^-. \tag{52}$$

For any given non-zero vector $\mathbf{w} = (\mathbf{s},\mathbf{t}) \in \mathbb{R}^{2n}$, we have

$$\mathbf{w}^T H(\boldsymbol{\sigma},\boldsymbol{\tau})\mathbf{w} = \mathbf{s}^T H_{1\sigma^2}(\boldsymbol{\sigma})\mathbf{s} + \sum_{i=1}^{n} -\frac{1}{2\tau_i}\left(s_i - t_i\frac{\sigma_i + f_i}{\tau_i}\right)^2. \tag{53}$$

Thus

$$\begin{aligned} \nabla^2 P^d(\boldsymbol{\sigma},\boldsymbol{\tau}) &\preceq 0\ \text{ if } (\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\sharp^+, \\ \nabla^2 P^d(\boldsymbol{\sigma},\boldsymbol{\tau}) &\succeq 0\ \text{ if } (\boldsymbol{\sigma},\boldsymbol{\tau}) \in \mathcal{S}_\flat^-. \end{aligned}$$

Therefore, $P^{\mathrm{d}}(\boldsymbol{\sigma},\boldsymbol{\tau})$ is concave on $\mathcal{S}_\sharp^+$, convex on $\mathcal{S}_\flat^-$, and $P^{\mathrm{g}}(\boldsymbol{\sigma})$ is concave on $\mathcal{S}_\sigma^+$. From the conditions in (48), we have for any $\boldsymbol{\sigma}_0 \in \partial\mathcal{S}_\sigma^+$ that

$$\lim_{\alpha\to 0^+} P^{\mathrm{g}}(\boldsymbol{\sigma}_o + \alpha\boldsymbol{\sigma}) = -\infty\ \ \forall\boldsymbol{\sigma} \in \mathcal{S}_\sigma^+ \tag{54}$$

and

$$\lim_{\alpha\to\infty} P^{\mathrm{g}}(\boldsymbol{\sigma}_o + \alpha\boldsymbol{\sigma}) = -\infty\ \ \forall\boldsymbol{\sigma} \in \mathcal{S}_\sigma^+. \tag{55}$$

This shows that the canonical dual function $P^{\mathrm{g}}(\boldsymbol{\sigma})$ is concave and coercive on the open set $\mathcal{S}_{\sigma}^{+}$. Therefore, by the theory of convex analysis, we know that the canonical dual problem $(\mathcal{P}_{+}^{\mathrm{g}})$ has at least one critical point $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_{\sigma}^{+}$, which is a global maximizer of $P^{\mathrm{g}}(\boldsymbol{\sigma})$ over $\mathcal{S}_{\sigma}^{+}$. By Theorem 2, the corresponding vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ is a global optimizer of the primal problem $(\mathcal{P}_{\flat})$. Moreover, if the conditions in (49) hold, then $H_{1\sigma^2}(\boldsymbol{\sigma}) \prec 0$; $H_{\tau^2}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \prec 0 \ \forall (\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathcal{S}_{\sharp}^{+}$, and the Hessian $\nabla^2 P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) \prec 0$, i.e., $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ is strictly concave on $\mathcal{S}_{\sharp}^{+}$. Therefore, $(\mathcal{P}^{\sharp})$ has a unique critical point in $\mathcal{S}_{\sharp}^{+}$, which implies $(\mathcal{P}_{+}^{\mathrm{g}})$ has a unique critical point in $\mathcal{S}_{\sigma}^{+}$ and the primal problem has a unique global minimizer. □

5 Application to Decoupled Problem

We now apply the theory presented in this chapter to a decoupled system. For simplicity, let $A = \mathrm{Diag}\,(\mathbf{a})$ be a diagonal matrix with $\mathbf{a} = \{a_i\} \in \mathbb{R}^n$ being its diagonal elements and consider the following extremal problem:

$$\min / \max \quad \left\{ P(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^{n} \left(\frac{1}{2} a_i x_i^2 + c_i x_i - f_i v_i \right) \right\} \tag{56}$$

$$\text{s.t.} \quad -v_i \le x_i \le v_i, \ v_i \in \{0, 1\}, \quad i = 1, \ldots, n. \tag{57}$$

The notation $\min/\max P$ stands for finding both minima and maxima of P. For this decoupled problem, the canonical dual function has a simple form given by

$$P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{c_i^2}{a_i + 2\sigma_i} + \frac{(f_i + \sigma_i + \tau_i)^2}{2\tau_i} \right). \tag{58}$$

From the criticality condition $\nabla P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau}) = 0$, the critical points of $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ can be obtained analytically as (with corresponding components)

$$\sigma_i \in \left\{ -\frac{1}{2}(a_i \pm c_i), -f_i \right\}, \ \ \tau_i \in \left\{ f_i - \frac{1}{2}(a_i \pm c_i), 0 \right\} \ \ \forall i = 1, \ldots, n. \tag{59}$$

By Theorem 1, the corresponding primal solution is (for $\tau_i > 0 \ \forall i$):

$$(x_i, v_i) = \left(-\frac{c_i}{a_i + 2\sigma_i}, \frac{f_i + \sigma_i + \tau_i}{2\tau_i} \right), \quad \forall i = 1, 2, \ldots, n. \tag{60}$$

Since each component of $(\boldsymbol{\sigma}, \boldsymbol{\tau}) \in \mathbb{R}^{2n}$ has two possible corresponding solutions according to (60), the canonical dual function P^{d} has 2^n critical points! By Theorem 2, the global extrema of the primal problem can be determined by the following theorem:

Theorem 5. *For any given* $\mathbf{a},\ \mathbf{c}, \mathbf{f} \in \mathbb{R}^n$, *if* $c_i \neq 0,\ \forall i$, *and if*

$$\max\left\{-\frac{1}{2}(a_i \pm c_i)\right\} > 0 \text{ and } \max\left\{f_i - \frac{1}{2}(a_i \pm c_i)\right\} > 0\ \forall i = 1, \ldots, n, \tag{61}$$

the canonical dual function P^{d} *has a unique critical point*

$$(\boldsymbol{\sigma}_\sharp, \boldsymbol{\tau}_\sharp) = \left(\max\left\{-\frac{1}{2}(a_i \pm c_i)\right\}\forall i, \max\left\{f_i - \frac{1}{2}(a_i \pm c_i)\right\}\forall i\right) \in \mathcal{S}_\sharp^+, \tag{62}$$

which is a global maximizer of $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ *on* $\mathcal{S}_\sharp^+$, *and*

$$(\mathbf{x}_\sharp, \mathbf{v}_\sharp) = \left(\left\{-\frac{c_i}{|c_i|}\right\}, \mathbf{e}\right) \tag{63}$$

is a global minimizer of $P(\mathbf{x}, \mathbf{v})$ *on* $\mathcal{X}_v$.

On the other hand, if $c_i \neq 0,\ \forall i$, *and if*

$$\min\left\{-\frac{1}{2}(a_i \pm c_i)\right\} < 0 \text{ and } \min\left\{f_i - \frac{1}{2}(a_i \pm c_i)\right\} < 0\ \forall i = 1, \ldots, n, \tag{64}$$

the canonical dual function P^{d} *has a unique critical point*

$$(\boldsymbol{\sigma}_\flat, \boldsymbol{\tau}_\flat) = \left(\min\left\{-\frac{1}{2}(a_i \pm c_i)\right\}\ \forall i, \min\left\{f_i - \frac{1}{2}(a_i \pm c_i)\right\}\ \forall i\right) \in \mathcal{S}_\flat^-, \tag{65}$$

which is a global minimizer of $P^{\mathrm{d}}(\boldsymbol{\sigma}, \boldsymbol{\tau})$ *on* $\mathcal{S}_\flat^-$ *and*

$$(\mathbf{x}_\flat, \mathbf{v}_\flat) = \left(\left\{\frac{c_i}{|c_i|}\right\},\ \mathbf{e}\right) \tag{66}$$

is a global maximizer of $P(\mathbf{x}, \mathbf{v})$ *on* $\mathcal{X}_v$ $\square$

6 Examples

6.1 Two-Dimensional Decoupled Problem

Let $a_1 = -3$, $a_2 = 2$, $c_1 = 5$, $c_2 = -8$, $f_1 = -2$, and $f_2 = 2$. The canonical dual function P^{d} has a total of nine critical points $(\boldsymbol{\sigma}, \boldsymbol{\tau})_k$, $k = 1, \ldots, 9$, and the corresponding results are listed below:

$$\begin{array}{lll}
(\boldsymbol{\sigma}, \boldsymbol{\tau})_1 = (4, 3, 2, 5), & (\mathbf{x}, \mathbf{v})_1 = (-1, 1, 1, 1), & P^{\mathrm{d}}{}_1 = -13.5; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_2 = (2, 3, 0, 5), & (\mathbf{x}, \mathbf{v})_2 = (0, 1, 0, 1), & P^{\mathrm{d}}{}_2 = -9.0; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_3 = (4, -2, 2, 0), & (\mathbf{x}, \mathbf{v})_3 = (-1, 0, 1, 0), & P^{\mathrm{d}}{}_3 = -4.5; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_4 = (-1, 3, -3, 5), & (\mathbf{x}, \mathbf{v})_4 = (1, 1, 1, 1), & P^{\mathrm{d}}{}_4 = -3.5; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_5 = (2, -2, 0, 0), & (\mathbf{x}, \mathbf{v})_5 = (0, 0, 0, 0), & P^{\mathrm{d}}{}_5 = 0; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_6 = (4, -5, 2, -3), & (\mathbf{x}, \mathbf{v})_6 = (-1, -1, 1, 1), & P^{\mathrm{d}}{}_6 = 2.5; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_7 = (-1, -2, -3, 0), & (\mathbf{x}, \mathbf{v})_7 = (1, 0, 1, 0), & P^{\mathrm{d}}{}_7 = 5.5; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_8 = (2, -5, 0, -3), & (\mathbf{x}, \mathbf{v})_8 = (0, -1, 0, 1), & P^{\mathrm{d}}{}_8 = 7; \\
(\boldsymbol{\sigma}, \boldsymbol{\tau})_9 = (-1, -5, -3, -3), & (\mathbf{x}, \mathbf{v})_9 = (1, -1, 1, 1), & P^{\mathrm{d}}{}_9 = 12.5.
\end{array}$$

By the fact that $(\boldsymbol{\sigma}, \boldsymbol{\tau})_1 \in \mathcal{S}_\sharp^+$ and $(\boldsymbol{\sigma}, \boldsymbol{\tau})_9 \in \mathcal{S}_\flat^-$, Theorem 5 tells us that $(\mathbf{x}, \mathbf{v})_1$ is a global minimizer and $(\mathbf{x}, \mathbf{v})_9$ is a global maximizer of $P(\mathbf{x}, \mathbf{v})$.

6.2 General Nonconvex Problem

We let $n = 10$ and randomly choose $\mathbf{c}$, $\mathbf{f}$, and A, where

$$\mathbf{c} = \{16, -13, -12, -18, -11, 7, 11, 16, -4, 18\}^T,$$
$$\mathbf{f} = \{11, 5, 13, 18, 6, 4, -16, 16, -20, -3\}^T,$$

$$A = \begin{bmatrix} 10 & 9 & 9 & 9 & 1 & 9 & 4 & 1 & 5 & 9 \\ 2 & 5 & 7 & 3 & 2 & 10 & 7 & 2 & 8 & 2 \\ 7 & 2 & 6 & 6 & 2 & 2 & 6 & 1 & 7 & 5 \\ 5 & 5 & 2 & 9 & 6 & 3 & 9 & 5 & 7 & 8 \\ 2 & 9 & 1 & 9 & 8 & 10 & 9 & 4 & 4 & 5 \\ 8 & 2 & 1 & 9 & 7 & 3 & 7 & 3 & 1 & 4 \\ 4 & 2 & 8 & 2 & 2 & 6 & 6 & 2 & 4 & 2 \\ 4 & 7 & 7 & 10 & 2 & 5 & 7 & 5 & 6 & 3 \\ 3 & 6 & 9 & 10 & 1 & 8 & 6 & 5 & 9 & 5 \\ 7 & 7 & 2 & 7 & 7 & 3 & 7 & 7 & 8 & 6 \end{bmatrix}.$$

By solving the canonical dual problem $(\mathcal{P}_+^{\mathrm{g}})$, we obtain the global maximizer

$$\bar{\boldsymbol{\sigma}} = \{7.7, 7.3, 6.3, 9.8, 4.3, 3.6, 11.9, 9.3, 7.8, 8.5\}^T$$

and

$$\bar{\boldsymbol{\tau}} = \{18.7, 12.3, 19.3, 27.8, 10.3, 7.6, 4.1, 25.3, 12.2, 5.5\}^T.$$

The global minimizer of the primal problem $(\mathcal{P})$ is then

$$\bar{\mathbf{x}} = \{-1.0, 1.0, 1.0, 1.0, 1.0, -1.0, 0, -1.0, 0, -1.0\}^T$$
$$\bar{\mathbf{v}} = \{1, 1, 1, 1, 1, 1, 0, 1, 0, 1\}^T,$$

and $P^{\mathrm{d}}(\bar{\boldsymbol{\sigma}}, \bar{\boldsymbol{\tau}}) = -181 = P(\bar{\mathbf{x}}, \bar{\mathbf{v}})$.

7 Concluding Remarks and Open Problems

We have studied in this chapter an application of canonical duality theory to solve the mixed-integer quadratic optimization problem $(\mathcal{P}_\flat)$ and its co-problem $(\mathcal{P}_\sharp)$. Using an appropriate quadratic measure $\mathbf{y} = \Lambda(\mathbf{x}, \mathbf{v}) = (\mathbf{x} \circ \mathbf{x} - \mathbf{v},\ \mathbf{v} \circ \mathbf{v} - \mathbf{v})$, the given nonconvex mixed-integer primal problem was converted into a canonical dual problem in continuous space and its relationship with the classical Lagrangian duality under a similar transformation was revealed. As a special application of the triality theory developed in [15], Theorem 2 shows that the canonical dual problem $(\mathcal{P}^\sharp)$ is a concave maximization

over the convex dual feasible space $\mathcal{S}_\sharp^+$ and the co-dual $(\mathcal{P}^\flat)$ is a convex minimization problem on $\mathcal{S}_\flat^-$. Therefore, both problems can be solved via convex programming optimization methods under the stated conditions. Theorem 3 shows that the mixed-integer programming problem in $\mathbb{R}^{2n}$ is canonically dual to a concave maximization problem $(\mathcal{P}_+^{\rm g})$ over a convex feasible set $\mathcal{S}_\sigma^+ \subset \mathbb{R}^n$, which can be solved efficiently via well-developed convex minimization techniques. Certain existence and uniqueness conditions related to critical points belonging to a derived dual feasible space for yielding a zero duality gap were established in Theorem 4. An illustrative example using a decoupled problem was presented and analytic solutions to both problems $(\mathcal{P}_\flat)$ and $(\mathcal{P}_\sharp)$ were obtained. A detailed study on more general mixed-integer programming problems along with semi-analytic solutions is forthcoming.

The canonical duality theory developed in [15] is composed mainly of (1) a *canonical dual transformation methodology*, (2) a *complementary dual principle*, and (3) an associated *triality theory*. The canonical dual transformation can be used to formulate perfect dual problems without a duality gap. The complementary dual principle shows that the nonsmooth/discrete primal problems are equivalent to continuous dual problems and a wide class of constrained nonconvex primal problems in $\mathbb{R}^n$ can be transformed to unconstrained canonical dual problems (with zero duality gap) on convex dual feasible spaces in $\mathbb{R}^m$ with $m \ll n$ (see [17, 19, 29]). The triality theory can be used to identify both global and local extrema and to develop powerful canonical dual algorithms for solving general nonconvex/nonsmooth problems in complex systems. As mentioned in many applications of the canonical duality theory (see [7, 15, 17, 19, 21, 28, 45]), the geometrical nonlinear (quadratic) operator $\mathbf{y} = \Lambda(\mathbf{x}, \mathbf{v})$ plays a key role in the canonical duality theory. For general optimization problems in finite dimensional spaces, this quadratic operator can be viewed as an Euclidian distance type measure. For nonconvex variational problems in infinite dimensional spaces, this geometrical measure can be viewed as a Cauchy–Riemann metric tensor (see [15]), while the canonical duality relations (5) are controlled by certain constitutive laws [15]. The complementary dual principle was an open problem in nonconvex mechanics for more than 40 years (see [37]). This problem was solved partially by Gao and Strang in 1989 [30] when a complementary gap function was discovered in nonconvex variational problems. This gap function provides a sufficient condition for global optimality. The pure complementary dual principle for general nonconvex systems was finally proposed in 1998 [13] and the triality theory reveals the intrinsic duality pattern in complex systems. Generally speaking, for any given primal problem, so long as the geometrical operator Λ is chosen properly, the canonical dual problem can be formulated in a standard fashion, and the triality theory can then be used to identify both global and local extrema and to develop powerful algorithms.

The results presented in this chapter can be generalized for solving more complicated problems in global optimization (cf. [21, 28]). Recently, the canonical duality theory has been used successfully for solving a class

of nonconvex problems in both finite and infinite dimensional spaces, including integer programming [7, 45], fractional programming [8], nonconvex polynomial-exponential minimization [20, 26], nonconvex minimization with general nonconvex constraints [28], and nonconvex variational/boundary value problems in mathematical physics and material science [13, 14, 24, 25, 31].

By the fact that the canonical duality is a precise theory (no duality gap), if the canonical dual function $P^{\mathrm{g}}(\boldsymbol{\sigma})$ for the fixed cost quadratic programming problem has a critical point $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_{\sigma}^{+}$, then the primal problem $(\mathcal{P}_{\flat})$ has a unique global minimizer

$$(\bar{\mathbf{x}}, \bar{\mathbf{v}}) = \left(-G^{+}(\bar{\boldsymbol{\sigma}})\mathbf{c}, \ \ \boldsymbol{\delta}(\mathbf{f} + \bar{\boldsymbol{\sigma}})^{+}\right). \tag{67}$$

However, if problem $(\mathcal{P}_{+}^{\mathrm{g}})$ has no critical point in $\mathcal{S}_{\sigma}^{+}$, primal problem $(\mathcal{P}_{\flat})$ could be difficult to solve. In this case the canonical dual problem is given by

$$(\mathcal{P}^{\mathrm{g}}): \quad \min \mathrm{sta}\{P^{\mathrm{g}}(\boldsymbol{\sigma}): \ \ \boldsymbol{\sigma} \in \mathcal{S}_a\}, \tag{68}$$

where

$$\mathcal{S}_a = \{\boldsymbol{\sigma} \in \mathbb{R}_{+}^{n} | \ \ \mathbf{f} + \boldsymbol{\sigma} \neq 0, \quad \mathbf{c} \in \mathcal{C}_{\mathrm{ol}}(G(\boldsymbol{\sigma}))\}. \tag{69}$$

By the canonical duality theory, if $\bar{\boldsymbol{\sigma}} \in \mathcal{S}_a$ is a solution of $(\mathcal{P}^{\mathrm{g}})$, the corresponding vector $(\bar{\mathbf{x}}, \bar{\mathbf{v}})$ given by (67) is a global minimizer of the primal problem $(\mathcal{P}_{\flat})$. Since the canonical dual function $P^{\mathrm{g}}(\boldsymbol{\sigma})$ is nonconvex on $\mathcal{S}_a$, to solve the minimal stationary problem $(\mathcal{P}^{\mathrm{g}})$ could be a challenging task and many related theoretical issues remain open.

References

1. Aardal, K.: Capacitated facility location: separation algorithms and computational experience. Math. Program. 81(2, Ser. B), 149–175 (1998)
2. Akrotirianakis, I.G., Floudas, C.A.: Computational experience with a new class of convex underestimators: Box-constrained NLP problems, J. Global Optim. 29, 249–264 (2004)
3. Akrotirianakis, I.G., Floudas, C.A.: A new class of improved convex underestimators for twice continuously differentiable constrained NLPs. J. Global Optim. 30, 367–390 (2004)
4. Atamtürk, A.: Flow pack facets of the single node fixed-charge flow polytope. Oper. Res. Lett. 29(3), 107–114 (2001)
5. Barany, I., Van Roy, T.J., Wolsey, L.A.: Strong formulations for multi-item capacitated lot sizing. Manage. Sci. 30, 1255–1261 (1984)
6. Contesse, L.: Une caractérisation compléte des minima locaux en programmation quadratique. Numer. Math. 34, 315–332 (1980)
7. Fang, S.-C., Gao, D.Y., Shue, R.L., Wu, S.Y.: Canonical dual approach to solving 0-1 quadratic programming problems. J. Ind. Manage. Optim. 4(1), 125–142 (2008)
8. Fang, S.-C., Gao, D.Y., Sheu, R.-L., Xing, W.X.: Global optimization for a class of fractional programming problems. J. Global Optim. 45, 337–353 (2009)

9. Floudas, C.A.: Deterministic Optimization. Theory, Methods, and Applications, Kluwer, Dordrecht (2000)
10. Floudas, C.A., Akrotirianakis, I.G., Caratzoulas, S., Meyer, C.A., Kallrath, J.: Global optimization in the 21st century: advances and challenges. Comput. Chem. Eng. 29, 1185–1202 (2005)
11. Floudas, C.A., Visweswaran, V.: A primal-relaxed dual global optimization approach. J. Optim. Theory Appl., 78(2), 187–225 (1993)
12. Floudas, C.A., Visweswaran, V.: Quadratic optimization. In: R. Horst, P.M. Pardalos (Eds.) Handbook of Global Optimization Kluwer, Dordrecht 217–270 (1995)
13. Gao, D.Y.: Duality, triality and complementary extremum principles in nonconvex parametric variational problems with applications. IMA J. Appl. Math. 61, 199–235 (1998)
14. Gao, D.Y.: Analytic solution and triality theory for nonconvex and nonsmooth variational problems with applications. Nonlinear Anal. 42(7), 1161–1193 (2000a)
15. Gao, D.Y.: Duality Principles in Nonconvex Systems: Theory, Methods and Applications, Kluwer, Dordrecht (2000b)
16. Gao, D.Y.: Canonical dual transformation method and generalized triality theory in nonsmooth global optimization. J. Global Optim. 17(1/4), 127–160 (2000c)
17. Gao, D.Y.: Perfect duality theory and complete solutions to a class of global optimization problems. Optimization 52(4–5), 467–493 (2003a)
18. Gao, D.Y.: Nonconvex semi-linear problems and canonical dual solutions. In: D.Y. Gao, R.W. Ogden (Ed.) Advances in Mechanics and Mathematics (Vol. II, pp. 261–312), Kluwer, Dordrecht (2003b)
19. Gao, D.Y.: Canonical duality theory and solutions to constrained nonconvex quadratic programming. J. Global Optim. 29, 377–399 (2004)
20. Gao, D.Y.: Complete solutions and extremality criteria to polynomial optimization problems. J. Global Optim. 35, 131–143 (2006)
21. Gao, D.Y.: Solutions and optimality to box constrained nonconvex minimization problems. J. Ind. Manage Optim. 3(2), 293–304 (2007a)
22. Gao, D.Y.: Duality-Mathematics. Wiley Encyclopedia of Electrical and Electronics Engineering (Vol. 6, pp. 68–77 (1st ed., 1999), Electronic edition, Wiley, New York (2007b)
23. Gao, D.Y.: Canonical duality theory: Unified understanding and generalized solution for global optimization problems. Comput. Chem. Eng. 33, 1964–1972 (2009) doi: 10.1016/j.compchemeng.2009.06.009
24. Gao, D.Y., Ogden, R.W.: Closed-form solutions, extremality and nonsmoothness criteria in a large deformation elasticity problem. Zeits. Angewandte Math. Phy. 59(3), 498–517 (2008a)
25. Gao, D.Y., Ogden, R.W.: Multi-solutions to nonconvex variational problems with implications for phase transitions and numerical computation. Quart. J. Mech. Appl. Math. 61(4), 497–522 (2008b)
26. Gao, D.Y., Ruan, N.: Complete solutions and optimality criteria for nonconvex quadratic-exponential minimization problem. Math. Methods Oper. Res. 67(3), 479–496 (2008c)
27. Gao, D.Y., Ruan, N.: On the solutions to quadratic minimization problems with box and integer constraints. J. Global Optim. to appear (2009a)

28. Gao, D.Y., Ruan, N., Sherali, H.D.: Solutions and optimality criteria for nonconvex constrained global optimization problems. J. Global Optim. to appear (2009b)
29. Gao, D.Y., Sherali, H.D.: Canonical duality theory: connections between nonconvex mechanics and global optimization. In: D.Y. Gao, H.D. Sherali (Eds.), Advances in Applied Mathematics and Global Optimization, 257–326, Springer (2009)
30. Gao, D.Y., Strang, G.: Geometric nonlinearity: potential energy, complementary energy, and the gap function. Quart. Appl. Math. 47(3), 487–504 (1989)
31. Gao, D.Y., Yu, H.: Multi-scale modelling and canonical dual finite element method in phase transitions of solids. Int. J. Solids Struct. 45, 3660–3673 (2008)
32. Glover, F., Sherali H.D.: Some classes of valid inequalities and convex hull characterizations for dynamic fixed-charge problems under nested constraints. Ann. Oper. Res. 40(1), 215–234 (2005)
33. Grippo, L., Lucidi, S.: A differentiable exact penalty function for bound constrained quadratic programming problems. Optimization 22(4), 557–578 (1991)
34. Gu, Z., Nemhauser, G.L., Savelsbergh, M.W.P.: Lifted flow cover inequalities for mixed 0-1 integer programs. Math. Program. 85(3, Ser. A), 439–467 (1999)
35. Han, C.G., Pardalos, P.M., Ye, Y.: An interior point algorithm for large-scale quadratic problems with box constraints. In: A. Bensoussan, J.L. Lions (Eds.), Springer-Verlag Lecture Notes in Control and Information (Vol. 144, pp. 413–422) (1990)
36. Hansen, P., Jaumard, B., Ruiz, M., Xiong, J.: Global minimization of indefinite quadratic functions subject to box constraints. Nav. Res. Logist. 40, 373–392 (1993)
37. Li, S.F., Gupta, A.: On dual configuration forces. J. Elasticity 84, 13–31 (2006)
38. Murty, K.G., Kabadi, S.N.: Some NP-hard problems in quadratic and nonlinear programming. Math. Program. 39, 117–129 (1987)
39. Padberg, M.W., Van Roy, T.J., Wolsey, L.A.: Valid linear inequalities for fixed charge problems. Oper. Res. 33, 842-861 (1985)
40. Pardalos, P.M., Schnitger, G.: Checking local optimality in constrained quadratic and nonlinear programming. Oper. Res. Lett. 7, 33–35 (1988)
41. Sherali, H.D., Smith, J.C.: An improved linearization strategy for zero-one quadratic programming problems. Optim. Lett. 1(1), 33–47 (2007)
42. Sherali, H.D., Tuncbilek, C.H.: A global optimization algorithm for polynomial programming problem using a reformulation-linearization technique. J. Global Optim. 2, 101–112 (1992)
43. Sherali, H.D., Tuncbilek, C.H.: A reformulation-convexification approach for solving nonconvex quadratic programming problems. J. Global Optim. 7, 1–31 (1995)
44. Sherali, H.D., Tuncbilek, C.H.: New reformulation-linearization technique based relaxation for univariate and multivariate polynominal programming problems. Oper. Res. Lett. 21(1), 1–10 (1997)
45. Wang, Z.B., Fang, S.-C., Gao, D.Y., Xing, W.X.: Global extremal conditions for multi-integer quadratic programming. J. Ind. Manage. Optim. 4(2), 213–225 (2008)

Algorithms of Quasidifferentiable Optimization for the Separation of Point Sets

Bernd Luderer[1] and Denny Wagner[2]

[1]Department of Mathematics, Chemnitz University of Technology, Reichenhainer Str. 41 09126, Chemnitz, Germany
b.luderer@mathematik.tu-chemnitz.de
[2]Capgemini, Lyon, France
denny_wagner@web.de

Summary. An algorithm for finding the intersection of the convex hulls of two sets consisting of finitely many points each is proposed. The problem is modelled by means of a quasidifferentiable (in the sense of Demyanov and Rubinov) optimization problem, which is solved by a descent method for quasidifferentiable functions.

Key words: quasidifferential calculus, separation of point sets, intersection of sets, hausdorff distance, numerical methods

1 Introduction

The following problem is considered: Given two sets A and B, it is required to separate these sets. Due to the general setting, the intersection $A \cap B$ may be nonempty. In this case it is required to assign the points of the sets A and B to the difference sets $A \setminus B$ or $B \setminus A$ or to establish that they belong to $A \cap B$. This task has to be done in the best way. The best result we can obtain is a complete assignment to one of the three sets.

Problems of such a type are of great practical importance. They arise, e. g. in medical or technical diagnosis, in pattern recognition, classification. Of course, different approaches towards a solution are possible. Here we describe a way of solving the original setting by means of a nondifferentiable, or more exactly, a quasidifferentiable optimization problem. The first stimulus for such a treatment of the problem was given in the papers of Demyanov [2] and Demyanov et al. [3]. This special nonconvex problem will then be solved by means of an algorithm developed for the minimization of quasidifferentiable functions due to Bagirov [1]. Especially, we search for the intersection of the convex hulls of two sets consisting of finitely many points each.

This chapter is organized as follows. After introducing some notions needed in the following we explain basic definitions and properties of quasidifferentials

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_8,

as well as most important rules of quasidifferential calculus due to Demyanov and Rubinov [4–6]. The next section deals with a numerical algorithm for minimizing some quasidifferentiable function. This algorithm has been proposed by Bagirov [1] and is closely related to algorithms used in Luderer and Weigelt [9] as well as Herklotz and Luderer [7]. After describing and discussing the principal method, a numerical algorithm and some preliminary test results are presented.

2 Basic Notions

In the following, all sets and vectors belong to the finite-dimensional space $\mathbb{R}^n$, although some extensions to more general spaces are possible.

Definition 1. *Given two sets M, N, the* Hausdorff distance $\varrho(M,N)$ *between them is defined as*

$$\varrho(M,N) = \max\left\{\max_{n\in N}\min_{m\in M}||m-n||, \max_{m\in M}\min_{n\in N}||m-n||\right\}.$$

Note that later on the Hausdorff distance is used as a stop criterion.

Definition 2. *By*

$$d_y^C = |\max_{c\in C}\langle c,y\rangle - \min_{c\in C}\langle c,y\rangle|$$

we denote the extension *of a set C in direction y.*

Let two sets A, B, as well as a vector y be given.

Definition 3. *Under the* directional difference DD_y^{AB} *of two sets A and B with respect to the direction y we understand the number*

$$DD_y^{AB} = |\max_{a\in A}\langle a,y\rangle - \max_{b\in B}\langle b,y\rangle|.$$

This notion will serve as a basis for finding some cutting hyperplane.

Definition 4. *The directional derivative of a function f at point x in direction r is defined as*

$$f'(x;r) = \lim_{t\downarrow 0}\frac{f(x+tr)-f(x)}{t}.$$

3 Quasidifferential Calculus

This calculus has been developed and proposed by Demyanov and Rubinov (see, e. g. [4, 5]). It is designed for a large class of nondifferentiable, nonconvex functions. Quasidifferential calculus generalizes both differential calculus and convex analysis.

Definition 5. *The function f is said to be* quasidifferentiable *at $x \in \mathbb{R}^n$ if f is directionally differentiable and there exists a pair of convex compact sets $Df(x) = [\underline{\partial}(x), \overline{\partial}(x)]$ such that*

$$f'(x;r) = \max_{v \in \underline{\partial}(x)} \langle v, r \rangle + \min_{w \in \overline{\partial}(x)} \langle w, r \rangle, \tag{1}$$

where $\underline{\partial}(x)$ is the subdifferential *and $\overline{\partial}(x)$ is the* superdifferential.

Let us note that the pair of sets constituting the quasidifferential to a function at a certain point is not unique, because if $Df(x) = [\underline{\partial}(x), \overline{\partial}(x)]$ is a quasidifferential, then for any convex compact set W, the pair of sets $[\underline{\partial}(x) + W, \overline{\partial}(x) - W]$ is also a quasidifferential.

If in the class of quasidifferentials there is one of the form $Df(x) = [\underline{\partial}(x), \mathbf{0}]$ ($Df(x) = [\mathbf{0}, \overline{\partial}(x)]$, resp.), then the function f is called *subdifferentiable (superdifferentiable*, resp.) at the point x.

Remark 1. In the case of a convex function the subdifferential $\underline{\partial}(x)$ in the sense of Demyanov and Rubinov coincides with the subdifferential $\partial f(x)$ in the sense of convex analysis, and from (1) we get the well-known relation $f'(x : r) = \max_{v \in \partial f(x)} \langle v, r \rangle$. On the other hand, if f is differentiable at the point x, then $\underline{\partial}(x)$ (or $\overline{\partial}(x)$) consists of only one element, the derivative $\nabla f(x)$, so that $Df(x) = [\nabla f(x), \mathbf{0}]$ or, equivalently, $Df(x) = [\mathbf{0}, \nabla f(x)]$. Thus $f'(x;r) = \langle \nabla f(x), r \rangle$.

For deriving rules of calculation for quasidifferentials, we need the following two rules of set algebra:

- Addition of a pair of sets $U_i, V_i \subset \mathbb{R}^n$, $i = 1, 2$:

$$[U_1, V_1] + [U_2, V_2] = [U_1 + U_2, V_1 + V_2].$$

- Multiplication of $[U, V]$, $U, V \subset \mathbb{R}^n$, by a scalar $\lambda \in \mathbb{R}$:

$$\lambda \cdot [U, V] = \begin{cases} [\lambda U, \lambda V], & \lambda \geq 0, \\ [\lambda V, \lambda U], & \lambda < 0. \end{cases}$$

Using these operations, we are able to describe the following rules for operations with quasidifferentiable sets (note that the family of quasidifferentiable functions is closed with respect to addition, multiplication by a scalar, maximization, minimization, etc.):

Let the functions f_i, $i = 1, \ldots, m$, be quasidifferentiable at x and let $\lambda \in \mathbb{R}$. Then the functions $f_1 + f_2$, λf, $\varphi(x) = \max_{i=1,\ldots,n} f_i(x)$, $\xi(x) = \min_{i=1,\ldots,n} f_i(x)$ are also quasidifferentiable at x, where

$$D(f_1 + f_2)(x) = Df_1(x) + Df_2(x),$$
$$D(\lambda f)(x) = \lambda Df(x),$$
$$D\varphi(x) = [\underline{\partial}\varphi(x), \overline{\partial}\varphi(x)], \qquad D\xi(x) = [\underline{\partial}\xi(x), \overline{\partial}\xi(x)]$$

with

$$\underline{\partial}\varphi(x) = \operatorname{co} \bigcup_{k \in R(x)} \Big(\underline{\partial} f_k(x) - \sum_{\substack{i \in R(x) \\ i \neq k}} \overline{\partial} f_i(x) \Big), \quad \overline{\partial}\varphi(x) = \sum_{k \in R(x)} \overline{\partial} f_k(x),$$

$$\underline{\partial}\xi(x) = \sum_{k \in Q(x)} \underline{\partial} f_k(x), \quad \overline{\partial}\xi(x) = \operatorname{co} \bigcup_{k \in Q(x)} \Big(\overline{\partial} f_k(x) - \sum_{\substack{i \in Q(x) \\ i \neq k}} \underline{\partial} f_i(x) \Big),$$

where $[\underline{\partial} f_k(x), \overline{\partial} f_k(x)]$ are quasidifferentials of f at x, $R(x) = \{i \mid f_i(x) = \varphi(x)\}$, $Q(x) = \{i \mid f_i(x) = \xi(x)\}$.

3.1 Necessary Optimality Conditions

Consider the unconstrained problem

$$f(x) \to \min_{x \in \mathbb{R}^n}.$$

Theorem 1. (Necessary optimality condition) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be quasidifferentiable and let x^* be a local minimizer of f. Then the following inclusion holds:*

$$-\overline{\partial} f(x^*) \subset \underline{\partial} f(x^*). \tag{2}$$

For the proof, see, e.g. [5].

Points satisfying (2) are called *inf-stationary* points. Later on we also need the weakened notion of *ε-inf-stationary* points, satisfying the relation

$$-\overline{\partial} f(x) \subset \underline{\partial}_\varepsilon f(x),$$

where $\underline{\partial}_\varepsilon f(x)$ is some enlargement of the set $\underline{\partial} f(x)$.

It is an advantage of quasidifferential calculus that we are able to distinguish between inf-stationary and sup-stationary points. In case x is not inf-stationary, one can indicate a direction of descent and even compute the (possibly non-unique) direction of steepest descent.

Theorem 2. (Direction of steepest descent) *If x_0 is not inf-stationary, then the vector $r_0 = -\dfrac{v_0 + w_0}{||v_0 + w_0||}$ is the direction of steepest descent of f at x_0, where*

$$||v_0 + w_0|| = \max_{w \in \overline{\partial} f(x_0)} \min_{v \in \underline{\partial} f(x_0)} ||v + w||.$$

For the proof, see, e.g. [5].

4 Principal Algorithm of Finding the Intersection of Two Sets

Let there be given two sets A and B consisting of a finite number of points each: $A = \{a_j \mid j \in J_1\}$, $B = \{b_j \mid j \in J_2\}$. Set $\mathcal{A} = \operatorname{co} A$, $\mathcal{B} = \operatorname{co} B$. The task consists in finding (or approximating) the intersection $\mathcal{A} \cap \mathcal{B}$.

Principal algorithm

- *Step 1.* Set $k = 1$, $A_k = \{a_j \,|\, j \in J_{k1}\}$, $B_k = \{b_j \,|\, j \in J_{k2}\}$.
- *Step 2.* If $\varrho(\mathcal{A}, \mathcal{B}) < \varepsilon$, then stop: $\mathcal{A}_k \cup \mathcal{B}_k \approx \mathcal{A} \cap \mathcal{B}$.
- *Step 3.* Find a direction y_k with $DD_{y_k}^{\mathcal{A}_k \mathcal{B}_k} > 0$. Evaluate the scalar $c = \min\left\{ \max_{a_j \in A_k} \langle a_j, y_k \rangle; \max_{b_j \in B_k} \langle b_j, y_k \rangle \right\}$. Determine $c_k \in A_k \cup B_k$ satisfying the relation $\langle c_k, y_k \rangle = c$.
- *Step 4.* Set $d_k = \langle c_k, y_k \rangle$ and find the cutting hyperplane $h_k(y_k, d_k)$.
- *Step 5.* If $c_k \in A_k$, then set $A_{k+1} = A_k$, $B_{k+1} = B_k \setminus \{b_j \in B_k | \langle b_j, y_k \rangle > d_k\} \cup N$, where $N \subset \{b_j \in \mathcal{B}_k | \langle b_j, y_k \rangle = d_k\}$. Analogously for $c_k \in B_k$. Set $k := k + 1$, go to step 2.

Proposition 1. *The hyperplane $h_k(y_k, d_k)$ occurring in step 4 is supporting to the set $\mathcal{A}_k$ ($\mathcal{B}_k$, resp.) if $c_k \in \mathcal{A}_k$ ($\mathcal{B}_k$, resp.).*

Proof. Let us consider, e.g. the case $c_k \in A_k$. For h_k being a supporting hyperplane of $\mathcal{A}_k$ at c_k, we have to show that

$$\langle y_k, a \rangle \leq d_k \quad \forall a \in \mathcal{A}_k, \quad \langle y_k, c_k \rangle = d_k. \tag{3}$$

Since $\mathcal{A}_k = \text{co } A_k$, the inequality in (3) can be restricted to points of A_k, i. e.

$$\langle y_k, a \rangle \leq d_k \quad \forall a \in A_k. \tag{4}$$

Let $a_k^* \in A_k$ satisfy the relation $\langle y_k, a_k^* \rangle > d_k$. Due to the second relation in (3) which is fulfilled by definition of d_k and $c_k \in A_k$ we get a contradiction to step 3 of the principal algorithm. □

The algorithm INTERSEC described in the next section aims at finding a vector y_k which is the normal vector of a cutting hyperplane to $\mathcal{A}_k$ or $\mathcal{B}_k$ such that the number of points z satisfying $\langle y_k, z \rangle > d_k$ and being removed in the kth iteration is *as large as possible.*

Proposition 2. *Instead of $\mathcal{A}_k$, $\mathcal{B}_k$ it suffices to consider the sets A_k, B_k consisting of a finite number of points each.*

Proof. We show that there is always an element $a^* \in A_k$ with $a^* \in \text{argmax } \{\langle a, y_k \rangle \,|\, a \in \mathcal{A}_k\}$ (cases $\mathcal{B}_k$ and B_k can be dealt with analogously). Indeed, consider some $\bar{a} \in \mathcal{A}_k$, $\bar{a} \notin A_k$. Then there exist scalars $\lambda_j \geq 0$, $\sum_{j=1}^{N_{k1}} \lambda_j = 1$ as well as vectors $a_j \in A_k$, $j = 1, \ldots, N_{k1}$, such that $\bar{a} = \sum_{j=1}^{N_k} \lambda_j a_j$. Let $a^* \in A_k$ be such an element that $\langle a^*, y_k \rangle = \max_{j \in J_{k1}} \langle a_j, y_k \rangle$. Then

$$\langle \bar{a}, y_k \rangle = \sum_{j=1}^{N_{k1}} \lambda_j \langle a_j, y_k \rangle \leq \sum_{j=1}^{N_{k1}} \lambda_j \langle a^*, y_k \rangle = \langle a^*, y_k \rangle. \qquad \square$$

In order to realize the task of finding a "good" cutting hyperplane, the following optimization problem is formed:

$$\begin{gathered} F(y_k) = |\max_{a\in A_k}\langle a, y_k\rangle - \max_{b\in B_k}\langle b, y_k\rangle| \to \max_{y_k\in S} \\ S = \{y_k \in \mathbb{R}^n : ||y_k|| = 1\}. \end{gathered} \tag{5}$$

Note that in [2, 3] a different objective function is used:

$$\widetilde{F}(y) = |\max_{a\in\mathcal{A}}\langle a, y\rangle - |\max_{b\in\mathcal{B}}\langle b, y\rangle| + |\min_{a\in\mathcal{A}}\langle a, y\rangle - |\min_{b\in\mathcal{B}}\langle b, y\rangle| = d_y^{\mathcal{A}\cup\mathcal{B}} - d_y^{\mathcal{A}\cap\mathcal{B}}.$$

It describes the difference of the extension of the sets $\mathcal{A}\cup\mathcal{B}$ and $\mathcal{A}\cap\mathcal{B}$.

As will be explained later on, the function F is quasidifferentiable and its quasidifferential can be computed in a relatively easy way. For solving problem (5) we will use an algorithm due to Bagirov [1] which is similar to algorithms used in [7, 9].

5 A Minimization Method Due to Bagirov

In [1] Bagirov describes a minimization method for the unconstrained problem

$$H(y) = G(y, \varphi_1(y), \ldots, \varphi_m(y)) \to \min_{y\in\mathbb{R}^n}, \tag{6}$$

where G is continuously differentiable on $\mathbb{R}^{n+m}$, $\varphi_i : \mathbb{R}^n \to \mathbb{R}$ are semismooth with upper semicontinuous directional derivatives $\varphi_i(\cdot; r)\ \forall r \in \mathbb{R}^n$. Since in this algorithm the quasidifferential of H plays an important role, we first need a description of the quasidifferential $DH(y) = [\underline{\partial}H(y), \overline{\partial}H(y)]$:

$$\underline{\partial}H(y) = \operatorname{co}\left\{ v \in \mathbb{R}^n \,|\, v = \nabla_y G + \sum_{i\in I_+(y)} c_i(y)v_i,\ v_i \in \partial_{Cl}\varphi_i(y) \right\},$$

$$\overline{\partial}H(y) = \operatorname{co}\left\{ w \in \mathbb{R}^n \,|\, w = \sum_{i\in I_-(y)} c_i(y)w_i,\ w_i \in \partial_{Cl}\varphi_i(y) \right\}.$$

Here $c_i(y) = \frac{\partial G}{\partial \varphi_i}(y)$, and the index sets I_+ and I_- are defined as follows: $I_+ = \{i | c_i(y) > 0\}$, $I_- = \{i | c_i(y) < 0\}$. Moreover, ∂_{Cl} denotes the Clarke subdifferential.

The algorithm from [1] will now be applied to the function F from (5). Thus, we consider the special case

$$F(y) = |\varphi_1(y) - \varphi_2(y)| \tag{7}$$

with $\varphi_i(y) = \max_{j\in J_i} f_{ij}(y)$, $i = 1, 2$, and $f_{1j}(y) = \langle a_j, y\rangle$, $f_{1j}(y) = \langle a_j, y\rangle$.

We observe that all assumptions of H from (6) are fulfilled for F. Moreover,

$$\underline{\partial}\varphi_1(y)=\operatorname{co}\bigcup_{k\in R_1(y)}\underline{\partial}f_{1k}(y)=\operatorname{co}\{a_k\mid k\in R_1(y)\},\qquad \overline{\partial}\varphi_1(y)=\mathbf{0},$$

$$\underline{\partial}\varphi_2(y)=\operatorname{co}\bigcup_{k\in R_2(y)}\underline{\partial}f_{2k}(y)=\operatorname{co}\{b_k\mid k\in R_2(y)\},\qquad \overline{\partial}\varphi_2(y)=\mathbf{0},$$

$$R_i(y)=\{j\in J_i\mid f_{ij}(y)=\varphi_i(y)\},\qquad \varphi_i(y)=\max_{j\in J_i}f_{ij}(y),\ \ i=1,2,$$

$$f_{1j}(y)=\langle a_j,y\rangle,\ j\in J_1,\qquad f_{2j}(y)=\langle b_j,y\rangle,\ \ j\in J_2.$$

Because in problem (5) the function F is to be maximized, we consider the problem

$$(-F)(y)\to\min$$

and describe the quasidifferential $D(-F)(y)$. To this aim, we have to distinguish the following two cases.

Case 1. Assume $\varphi_1(y)\geq\varphi_2(y)$. Then

$$\underline{\partial}(-F)(y)=\operatorname{co}\{b_j\mid j\in R_2(y)\},\qquad \overline{\partial}(-F)(y)=\operatorname{co}\{-a_i\mid i\in R_1(y)\}.$$

Case 2. Assume $\varphi_1(y)<\varphi_2(y)$. Then

$$\underline{\partial}(-F)(y)=\operatorname{co}\{a_i\mid i\in R_1(y)\},\qquad \overline{\partial}(-F)(y)=\operatorname{co}\{-b_j\mid\ j\in R_2(y)\}.$$

For solving problem (6) in [1] Bagirov proposes some method using exact line search and finding the so-called ε-inf-stationary points satisfying $-\overline{\partial}f(y^*)\subset\underline{\partial}_\varepsilon f(y^*)$. For this reason, instead of the sub- and the superdifferential of the function H he uses some enlargements of these sets (cf. a similar algorithm by Luderer and Weigelt [9]). At the same time, the functions $\varphi_i(y)$, $i=1,2$, occurring in H are assumed to be the maximum of continuously differentiable functions (cf. (7)).

We need the following sets (ε,$\mu>0$):

$$R_{i\varepsilon}(y)=\{j\in J_i\mid f_{ij}(y)\geq\varphi_i(y)-\varepsilon\},\quad i=1,2,$$

$$\underline{\partial}_\varepsilon f(y)=\operatorname{co}\left\{v\in\mathbb{R}^n\mid v=\nabla_y G(y)+\sum_{i\in I_+(y)}c_i(y)\nabla f_{ij}(y),j\in R_{i\varepsilon}(y)\right\},$$

$$B_\mu(y)=\left\{w\in\mathbb{R}^n\mid w=\sum_{i\in I_-(y)}c_i(y)\nabla f_{ij}(y),j\in R_{i\mu}(y)\right\}.$$

Using these sets, the following algorithm is described in [1]:

Descent algorithm with exact line search

- *Step 1.* Choose any $y^0 \in \mathbb{R}^n$, set $k := 0$.
- *Step 2.* If $-\overline{\partial} f(y_k) \subset \underline{\partial}_\varepsilon f(y_k)$, then stop: y_k is ε-inf-stationary.
- *Step 3.* Find for any $w \in B_\mu(y_k)$ a vector $v_k(w)$ with

$$||w + v_k(w)|| = \min_{v \in \underline{\partial}_\varepsilon f(y_k)} ||w + v||.$$

- *Step 4.* If $w + v_k(w) \neq \mathbf{0}$, then set $g_k(w) = -\dfrac{w + v_k(w)}{||w + v_k(w)||}$.
- *Step 5.* Evaluate the step size $\alpha_k(w) \geq 0$ with

$$f(y_k + \alpha_k(w) g_k(w) = \inf_{\alpha \geq 0} f(y_k + \alpha g_k(w)).$$

  If $w + v_k(w) = \mathbf{0}$, then set $\alpha_k(w) g_k(w) = \mathbf{0}$.
- *Step 6.* Find w_k such that

$$f(y_k + \alpha_k(w_k) g_k(w_k)) = \min_{w \in B_\mu(y_k)} f(y_k + \alpha_k(w) g_k(w)).$$

  Go to step 2.

Remark 2.

1. The description of the quasidifferential of $(-F)$ given above has to be adapted in an obvious way. This is omitted here.
2. Bagirov's algorithm is designed for unconstrained minimization. However, (5) is a constrained optimization problem with "simple" constraints. Thus, projection onto S can be easily and explicitly carried out:

$$P_S(y) = \begin{cases} y, & y \in S, \\ y/||y||, & y \notin S. \end{cases}$$

   Using this projection, Rosens's gradient projection method (see [10]) will be applied to (5).
3. As a method of line search (for finding a suitable step size) we use quadratic interpolation.
4. Other algorithms suitable for solving (6) and (5), resp., are, e.g. the method of codifferential descent (see [1]) and Kiewiel's linearization method [8].
5. Let us emphasize that in the cutting process (by means of supporting hyperplanes to $\mathcal{A}_k$ and $\mathcal{B}_k$, resp.), some points of the positive half-space drop out, whereas some other points lying on the hyperplane h_k have to be added for correct construction of the next convex hull in the iteration process. These points are generated in the following way (the procedure is described for set A_k; concerning B_k the method works analogously): Consider all points of A_k lying on one side of h_k and all points lying on the other. Connect them by straight lines and take the intersection with h_k. All points constructed

in this way have to be added to A_k. Unfortunately, as a consequence the number of points belonging to A_k grows up considerably. If we succeed in finding the extreme points on h_k, then only these extreme points should be added to A_k. In this way, we have to perform the following manipulation with $\mathcal{A}_k$ (let h_k be a supporting hyperplane to $\mathcal{B}_k$):

– Set $c = \min\{\max_{j\in J_{k1}}\langle a_j, y_k\rangle, \max_{j\in J_{k2}}\langle b_j, y_k\rangle, \}$ (since h_k is supporting to $\mathcal{B}_k$, we have $c = \max_{j\in J_{k2}}\langle b_j, y_k\rangle$). Find the sets

$$P_{A,\text{out}} = \{a_j \in A_k \,|\, \langle a_j, y_k\rangle > c\}, \quad P_{A,\text{int}} = \{a_j \in A_k \,|\, \langle a_j, y_k\rangle < c\}.$$

– Define, for all $a_m \in P_{A,\text{out}}$ and $a_n \in P_{A,\text{int}}$, the quantities $a_{mn}(\alpha) = \alpha a_m + (1-\alpha)a_n$ and find numbers $\alpha_{mn} \in (0,1)$ as well as the set

$$P_{A,bd} = \{a_{mn}(\alpha_{mn}) \,|\, \langle a_{mn}(\alpha_{mn}), y_k\rangle = c\}.$$

– Set $A_{k+1} = (A_k \setminus P_{A,\text{out}}) \cup P_{A,bd}$.

6 Algorithm INTERSEC

Now we are prepared to describe an algorithm for finding the intersection of two convex hulls:

Algorithm INTERSEC

1. Set $k = 1$, $A_k = A$, $B_k = B$ and choose $\varepsilon > 0$.
2. If $\varrho(\mathcal{A}_k, \mathcal{B}_k) < \varepsilon$, then stop: $\mathcal{A}_k \cup \mathcal{B}_k$ is an approximation of $\mathcal{A}_k \cap \mathcal{B}_k$.
3. Find a direction y_k as a solution of problem (1).
4. If $\max\limits_{j\in J_{k1}}\langle a_j, y_k\rangle < \max\limits_{j\in J_{k2}}\langle b_j, y_k\rangle$, then cut $\mathcal{B}_k$ and set

$$A_{k+1} = A_k, \quad B_{k+1} = B_k \setminus P_{B,\text{out}} \cap P_{B,bd},$$

otherwise cut $\mathcal{A}_k$ and set

$$B_{k+1} = B_k, \quad A_{k+1} = A_k \setminus P_{A,\text{out}} \cap P_{A,bd}.$$

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

Remark 3. In the section process only the sets A_{k+1}, B_{k+1} are changed. After that the new convex hulls $\mathcal{A}_{k+1}$, $\mathcal{B}_{k+1}$ are formed.

Due to the inclusion

Theorem 3. *For the Hausdorff distance*

$$\varrho_k = \varrho((\mathcal{A}_k \cup \mathcal{B}_k), (\mathcal{A} \cap \mathcal{B})) = \varrho(\mathcal{A}_k, \mathcal{B}_k)$$

we have $\forall\, \varepsilon > 0\ \exists k > 0\colon\ \varrho_k < \varepsilon$.

Proof. We have $\mathcal{A}_{k+1} \subseteq \mathcal{A}_k$, $\mathcal{B}_{k+1} \subseteq \mathcal{B}_k$, where at least one inclusion is proper. Let us assume that there exists an $\varepsilon_0 > 0$ such that $\varrho_k > \varepsilon \ \forall k$. According to the method described above, for every k there exists a value $c_k = \{\text{argmax}_{a_k \in \mathcal{A}_k} \langle a_k, y_k \rangle, \text{argmax}_{b_k \in \mathcal{B}_k} \langle b_k, y_k \rangle\}$ with $\varrho(c_k, \mathcal{A}_{k+1} \cup \mathcal{B}_{k+1}) \geq \varepsilon_0$. From the above inclusions it follows that $\varrho(c_k, \mathcal{A}_s \cup \mathcal{B}_s) \geq \varepsilon_0 \ \ \forall s \geq k$. Since $c_k \in \mathcal{A}_k \cup \mathcal{B}_k$, from the last inequality we get $\varrho(c_k, c_s) \geq \varepsilon_0 \ \ \forall s \geq k$. But $\{c_k\}$ is a bounded sequence, because $\mathcal{A} \cup \mathcal{B}$ is bounded. Choosing a convergent subsequence $\{c_{k_i}\}$ for i, j sufficiently large, we obtain $||c_{k_j} - c_{k_i}|| < \varepsilon_0$, a contradiction. □

7 Preliminary Numerical Results

Using the Matlab system, preliminary tests have been carried out. The main experiences are the following:

If we use in the section process only extreme points (which can be easily done for dimensions $n = 2$, $n = 3$), then we get quite satisfactory results in approaching the intersection of two sets. In doing this, in most cases the method of codifferentiable descent (with Armijo step size; see [1]) is the best one, followed by the above-described descent method with exact line search (and quadratic interpolation for step size determination), whereas Kiwiel's linearization method (see [8]) is inferior.

The choice of initial direction vectors is very important. We tried the following approaches: begin with the last vector of the previous iteration (this is unfavourable), use a special deterministic grid (this led to good results), and find the initial vectors in a stochastic way.

For $n \geq 4$ the computing time is strongly growing. The reason is that in the cutting process we now consider all points of $P_{A,bd}$ and $P_{B,bd}$, respectively, instead of only the extreme points. Thus the number of points in A_k, B_k grows rapidly. Only if we succeed in identifying the sets A_k, B_k by a smaller number of points, then the method described above seems to be promising. Thus, further research has to be done in numerical respect.

Finally, let us note that for finding points $c \in A \cup B$ being located in $\mathcal{A} \cap \mathcal{B}$ another algorithm, which is based on Wolfe's algorithm (see [11]), works very satisfactory even for higher dimensions.

References

1. Bagirov, A.M.: Numerical methods for minimizing quasidifferentiable functions: A survey and comparison. In: V.F. Demyanov, A. Rubinov, (Eds.): Quasidifferentiability and Related Topics (pp. 33–71), Kluwer, Dordrecht (2000)
2. Demyanov, V.F.: On the identification of points of two convex sets. Vestn. St. Petersburg Univ., Math. 34(3), 14–20 (2001)

3. Demyanov, V.F., Astorino, A., Gaudioso, M.: Nonsmooth problems in mathematical diagnostics. In: N. Hadjisavvas, P.M. Pardalos (Eds.), Advances in Convex Analysis and Global Optimization (Pythagorion, 2000), Nonconvex Optimization and Applications (Vol. 54, pp. 11–30), Kluwer, Dordrecht, (2001)
4. Demyanov, V.F., Rubinov, A.M.: Quasidifferentiable functionals. Dokl. Akad. Nauk SSSR 250(1), 21–25 (1980)
5. Demyanov, V.F., Rubinov, A.M.: Quasidifferential Calculus, Optimization Software, New York, NY (1986)
6. Demyanov, V.F., Rubinov, A.M.: Quasidifferentiability and Related Topics. Kluwer, Dordrecht (2000)
7. Herklotz, A., Luderer, B.: Identification of point sets by quasidifferentiable functions. Optimization 54, 411–420 (2005)
8. Kiwiel, K.C.: A linearization method for minimizing certain quasidifferentiable functions. Math. Program. Study 29, 86–94 (1986)
9. Luderer, B., Weigelt, J.: A solution method for a special class of nondifferentiable unconstrained optimization problems. Comput. Optim. Appl. 24, 83–93 (2003)
10. Rosen, J.B.: The gradient projection method for nonlinear programming. Part II: nonlinear constraints. J. Ind. Appl. Math. 8, 514–532 (1961)
11. Wolfe, P.: Finding the nearest point in a polytope. Math. Program. 11 (2), 128–149 (1976)

A Hybrid Evolutionary Algorithm for Global Optimization

Mend-Amar Majig[1], Abdel-Rahman Hedar[2], and Masao Fukushima[3]

[1] Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
mendamarm@num.edu.mn
[2] Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
hedar@aun.edu.eg
[3] Department of Applied Mathematics and Physics, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
fuku@i.kyoto-u.ac.jp

Summary. In this work, we propose a method for finding as many as possible, hopefully all, solutions of the global optimization problem. For this purpose, we hybridize an evolutionary search algorithm with a fitness function modification procedure. Moreover, to make the method more effective, we employ some local search method and a special procedure to detect unpromising trial solutions. Numerical results for some well-known global optimization test problems show the method works well in practice.

Key words: global optimization, tunneling function, evolutionary algorithm, local search

1 Introduction

Consider the global optimization problem

$$\min \ f(x) \text{ s.t. } x \in D, \tag{1}$$

where f is a real-valued function and the set D is defined as $D := \{x \in R^n | \ l \leq x \leq u\}$. Here $l, u \in (R \cup \{\pm\infty\})^n$ are, possibly infinite, lower and upper bounds on the variable. This problem is a fundamental problem of optimization and

This research was supported in part by a Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_9,

has a large number of important applications. Many algorithms have been proposed for solving it [1–5, 7–10], but most of them are intended to find just a solution of this problem. However, in practice, it is appealing to have a method designed for finding all, or as many as possible, solutions of the problem.

The purpose of this chapter is to develop a method of finding as many as possible, hopefully all, solutions of the global optimization problem. We propose a hybrid evolutionary algorithm (HEA) with the fitness function modification procedure. An evolutionary algorithm gives us the opportunity to search multiple solutions simultaneously. But when we use an evolutionary algorithm in a simple manner, the searching process is very likely to wander around already detected solutions in vain. So we employ a fitness function modification procedure which is designed to prevent the search process from returning back to the already detected solutions. We use mainly two types of modifications, namely tunneling function and hump-tunneling function modifications.

Tunneling function method for solving global optimization problem was first proposed by Levy and Montalvo [9, 10] in 1985. The idea of tunneling is that once the iteration is entrapped in a local solution, the method constructs a new objective function which is expected to have no local solution around the point of trap and hopefully no basin around it. The next iteration point will be chosen from a neighborhood of this point and the iteration will continue with the new objective function. In our method we will use not only the tunneling function but also more importantly the hump-tunneling function in order to overcome some drawbacks of the tunneling function.

An evolutionary algorithm with similar tunneling and hump-tunneling function modifications has been proposed to solve the general variational inequality problem (VIP) by the authors [11], where the VIP is reformulated as an optimization problem whose global minima with zero objective value coincide with the solutions of the original VIP. The algorithm of [11] fully exploits the special property of the problem that the minimum objective value is known to be zero at any solution. Therefore, it cannot be applied to the general optimization problem (1) directly. The algorithm proposed in this chapter incorporates additional devices to cope with the general situation where the global minimum value of the problem is not known in advance.

The organization of this chapter is as follows: In Section 2, we first give a brief review of the evolutionary algorithm and main procedures used in it. In Section 3, we describe our HEA and its elements in detail. The fitness function modification procedures as well as classification of the modification points will be explained there. We then present numerical results in Section 4 and conclude chapter in Section 5.

2 Evolutionary Algorithm

2.1 Basic Schemes

An evolutionary algorithm is based on the idea of imitating the evolutionary process observed in nature. Encouraged by the roles of reproduction, mutation, and survival in the evolution of living things, an evolutionary algorithm tries to combine and change elements of existing solutions in order to create a new solution with some of the features of parents and selects next candidate solutions among them [3, 4, 12].

An evolutionary algorithm for optimization is different from classical optimization methods in several aspects. First of all, it depends on random sampling, i.e., the method is non-deterministic. So there is no theoretical guarantee for the method to find an optimal solution.

Second, an evolutionary algorithm works with a *population* of candidate solutions, meanwhile classical optimization methods usually maintain a single best solution found so far. The use of population sets helps the evolutionary algorithm avoid being trapped at a local solution.

Moreover, we will never know whether we have found a true global minimizer or not unless we already knew the global minimum value of the problem beforehand. So, in general, in order to terminate the evolutionary algorithm we usually use the upper limit on the number of function evaluations. Once the number of function evaluations hits this upper limit, the algorithm stops, and the best solution found so far is regarded as a global minimum.

Basic scheme of an evolutionary algorithm is given in Fig. 1. It relies on procedures like parents selection, crossover and mutation, and survival selection [3, 4]. Next we will discuss these procedures in detail.

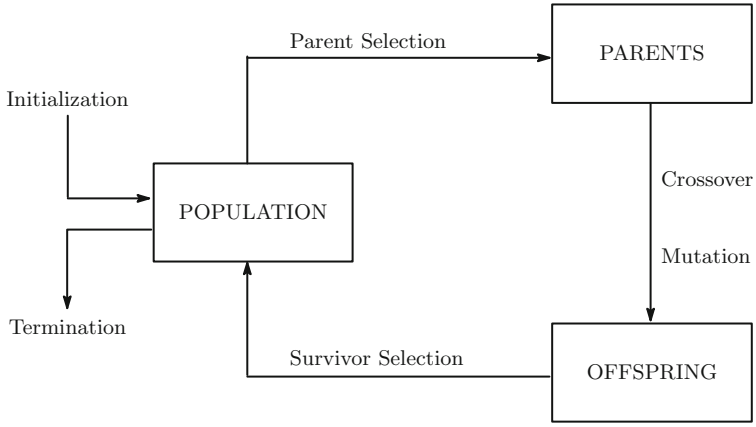


Fig. 1. Basic scheme of an evolutionary algorithm

2.2 Procedures Used in Evolutionary Algorithm

Now we elaborate the procedures shown in Fig. 1.

Initialization. We choose the parameters, the fitness function and an initial population set. To generate the initial population set, we use either a random distribution or a controlled random distribution. For example, the following procedure gives us a good diverse population set.

Diversification Generation Method: The purpose of the diversification generation [7, 8] is to generate a well-distributed set of trial solutions. The basic diversification generation method uses controlled randomization and frequency memory to generate a set of diverse solutions. This can be accomplished by dividing the range $[l_i, u_i]$ of each variable into four subranges of equal size. Then, a solution is constructed in two steps. First, a subrange is randomly selected. The probability of selecting a subrange is determined to be inversely proportional to its frequency count. Then a value is randomly generated within the selected subrange.

Crossover and Mutation. The purpose of *crossover* is to produce children who are expected to possess better properties than their parents. Good results can be obtained with a random matching of the individuals [3, 4]. Moreover, random changes or *mutations* are made periodically for some members of the current population, thereby yielding a new candidate solution. Some well-known crossovers are the following [6].

Single-point crossover: One crossover position (coordinate) in the vector of variables (genes) is randomly selected and the variables situated after this point are exchanged between individuals, thus producing two offsprings.

Multi-point crossover: Some crossover positions are chosen, and then the variables between successive crossover points are exchanged among the two parents to produce new offsprings.

Intermediate recombination: The values of the offspring variables are chosen from the values of the parents variables according to some rule.

Survival Selection. An evolutionary algorithm performs a selection process in which the most fit members of the population survive and the least fit members are eliminated. This process is done with the help of the fitness function and leads the population toward ever-better solutions.

3 Hybrid Evolutionary Algorithm

Now we describe our hybrid evolutionary algorithm HEA for global optimization. First, we will discuss the features of our algorithm that distinguish it from ordinary evolutionary algorithms.

If we use an evolutionary algorithm directly to search for multiple global solutions, it is very likely that the iteration process wanders around the already detected solutions without further advance. Since we are searching for all possible solutions, we need to prevent this kind of hindrance and go further for other solutions. To this end we propose here the fitness function modification procedure, which gives us an opportunity to go after the other solutions. The modification utilizes the tunneling function technique [1, 5, 9, 10] so that, once a local or global solution is detected during the computation, a new function is constructed to escape from the region of this solution in the further search. The new function has hopefully no solution near the point of tunneling and no basin around it. In our algorithm we use not only the tunneling function idea but also more importantly the hump-tunneling function technique [11] which is designed to overcome some drawbacks of the previous function. Details of these modifications are described in Section 3.1. Moreover, to make the method more effective, we apply a local optimization method starting from the best points in the population set. Local optimization will always be applied to the original objective function, since it will not affect the local search process even if the fitness function has been modified to a complicated function. Also, using local search will help us to detect solutions in the population set which are useless in the further search.

Another idea we use in our algorithm is intended to keep diversity of the population set. In ordinary evolutionary algorithms, a newly produced trial solution is usually accepted to survive and replace some solution in the population set, if it is better in values of the fitness function [3, 4]. Because of this selection rule, most evolutionary algorithms have the tendency that population sets eventually cluster around only a few solutions. Although some algorithms such as scatter search method [7, 8] try to keep diversity, the number of different good candidate points in the population set is still small, and the remaining points are usually just diversity points. The HEA uses the Population Update Rules (see Section 3.2), which are new types of criteria for accepting new trial solutions to survive in the population set, and tries to keep diversity while searching for promising points. The main idea is to utilize the distances between newly produced points and former members of the population set.

In an ordinary evolutionary algorithm, the upper limit on the number of function evaluations is used to terminate it [3, 4, 8]. Our HEA uses the upper limit not only for the number of function evaluations but also for the number of global solutions to be detected. Otherwise, since the problem may have infinitely many solutions, it is hardly possible to enumerate them in such a case.

3.1 Modification of the Fitness Function

First, let us consider the following two types of functions.
Tunneling function. Let f be our objective function and $\bar{x}$ be a point around which f is to be modified. Define

$$f_{\mathrm{t}}(x,\bar{x}) := (f(x) - f(\bar{x})) \cdot \exp\left(\frac{1}{\varepsilon_t + \frac{1}{\rho_{\mathrm{t}}^2}\|x-\bar{x}\|^2}\right), \tag{2}$$

where ε_{t} and ρ_{t} are positive parameters that control the degree and the range of modification. This function is called a *tunneling function* because of its behavior around the point $\bar{x}$ [1, 9, 10].

If $\bar{x}$ is not a global minimum of the function f, then $f_{\mathrm{t}}(\bar{x},\bar{x}) = 0$, and there must be at least one point on which the modified function $f_{\mathrm{t}}(x,\bar{x})$ has a negative value.

Now let $\bar{x}$ be an isolated global minimum of f. If $\bar{x}$ is an exact global solution, then the function $f_{\mathrm{t}}(x,\bar{x})$ has now the zero global minimum value. But, if $\bar{x}$ is just an approximation of a global solution $\bar{x}^*$, as one may expect in practice, then it may not be appropriate to use the tunneling function modification $f_{\mathrm{t}}(x,\bar{x})$, because we cannot fully escape from the point $\bar{x}$ in the next search (see Fig. 2).

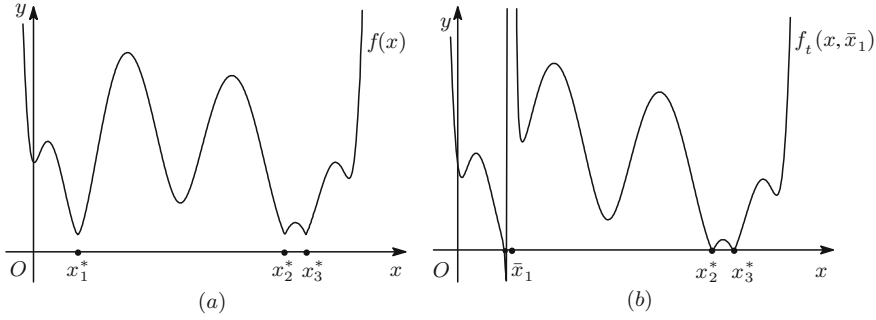


Fig. 2. (**a**) The original function and (**b**) its tunneling modification at an approximate solution $\bar{x}_1$.

We propose the following approach to overcome the above-mentioned drawback of the tunneling modification.

Hump-tunneling function. We first choose a positive scalar ρ_{h} and define a *hump function* $f_{\mathrm{h}}(x,\bar{x})$ as follows:

$$f_{\mathrm{h}}(x,\bar{x}) := f(x) - f(\bar{x}) + \alpha_{\mathrm{h}} \max\left\{0, 1 - \frac{1}{\rho_{\mathrm{h}}^2}\|x-\bar{x}\|^2\right\}, \tag{3}$$

where $\alpha_{\mathrm{h}} > 0$ is some parameter. Although this modification yields a non-differentiable function even when the original function is differentiable, it will not affect our local search procedure. Then we construct the following function:

$$\begin{aligned}\bar{f}_{\mathrm{ht}}(x,\bar{x}) :&= f_{\mathrm{h}}(x,\bar{x})\cdot\exp\left(\frac{1}{\varepsilon_{\mathrm{t}}+\frac{1}{\rho_{\mathrm{t}}^2}\|x-\bar{x}\|^2}\right)\\ &= \left(f(x)-f(\bar{x})+\alpha_{\mathrm{h}}\max\left\{0,1-\frac{1}{\rho_{\mathrm{h}}^2}\|x-\bar{x}\|^2\right\}\right)\cdot \qquad (4)\\ &\quad \exp\left(\frac{1}{\varepsilon_{\mathrm{t}}+\frac{1}{\rho_{\mathrm{t}}^2}\|x-\bar{x}\|^2}\right).\end{aligned}$$

We call this function the *hump-tunneling function* and global minimizers of this function coincide with those of the function $f(x)$ except for those minimizers in $B(\bar{x},\rho_{\mathrm{h}})$. An improper choice for the humping parameter ρ_{h} may result in the loss of some other global solutions near $\bar{x}$ (see Fig. 3a). By choosing ρ_{h} small enough in the hump-tunneling function, we can avoid this kind of difficulty (see Fig. 3b).

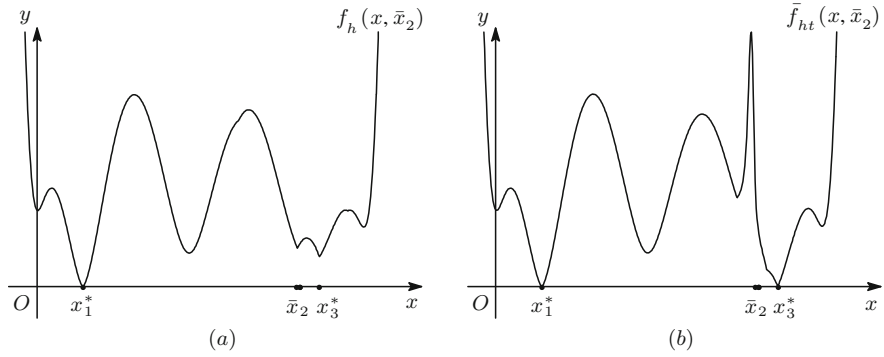


Fig. 3. (**a**) An unappropriate hump function of the function of Fig. 2a and (**b**) an appropriate hump-tunneling function constructed through modification at an approximate solution $\bar{x}_2$

In our HEA, we will mainly use these two modifications. Now we will discuss when and how we employ these modifications.

Modification and classification of modification points. The HEA collects the detected global or local solutions or unpromising trial points in the set S of modification points. Once one of those points is detected, the HEA adds it to S and modifies the objective function around this point in order to avoid returning to it in the further search. Let $f_{\mathrm{c}}(x)$ be the current fitness function used in the HEA and S be a set of modification points. Let $\bar{x}$ be a point around which the function $f_{\mathrm{c}}(x)$ is to be modified. Depending on the type of point $\bar{x}$, we use different modifications.

Definition 1. *If after a certain number of evolutionary generations and local searches, the best candidate solution in the population set P has not been*

improved, then we say the point is a semi-global solution. Moreover, a semi-global solution who has the lowest known fitness function value will be classified as an incumbent solution.

Incumbent solutions are the best points detected up to date. If we cannot find better solutions than these after a certain amount of explorations, they will be regarded as global solutions of the problem. We will also collect the incumbent solutions in the set S_{inc} and it will play an important role in the algorithm. Now we consider the modifications.

1. If $\bar{x}$ is an incumbent solution, then we set

$$S := S \cup \{\bar{x}\},\ S_{\mathrm{inc}} := S_{\mathrm{inc}} \cup \{\bar{x}\},$$

$$f_{\mathrm{c}}(x) := \left(f(x) - f(\bar{x}) + \alpha_{\mathrm{h}} \sum_{x_{\mathrm{g}} \in S_{\mathrm{inc}}} \max\left\{0, 1 - \frac{1}{\bar{\rho}_{\mathrm{h}}^2} \|x - x_{\mathrm{g}}\|^2\right\} \right) \cdot \exp\left(\sum_{x_{\mathrm{m}} \in S} \frac{1}{\varepsilon_t + \frac{1}{\rho_{\mathrm{t}}^2} \|x - x_{\mathrm{m}}\|^2} \right).$$

   After this modification, the new fitness function will have non-negative values at points no better than the incumbent solutions.
2. Suppose $S_{\mathrm{inc}} \neq \emptyset$ and $f_{\mathrm{c}}(\bar{x}) < 0$. Note that $S_{\mathrm{inc}} \neq \emptyset$ means we already have an incumbent solution and have modified the original fitness function. As mentioned above the new fitness function has non-negative values at points worse than the incumbent solutions. But since $f_{\mathrm{c}}(\bar{x}) < 0$, $\bar{x}$ is better than the current incumbent solutions and hence those incumbent solutions are not global minimizers. So setting

$$S_{\mathrm{inc}} := \emptyset,\ f_{\mathrm{c}}(x) := f(x),$$

   and including the point $\bar{x}$ in the population set, we try to find a new incumbent solution better than the previous incumbent solution with the new fitness function. Note that the set of modification points S remains the same and will be on effect after an incumbent solution is detected.

Before considering the last type of modification, let us introduce the concept of unpromising trial points.

Definition 2. *Let $f(x)$ and $f_{\mathrm{c}}(x)$ be the original and the current objective functions, respectively, and $\bar{x}$ be a trial point. Suppose a local search is executed on the original objective function f with the starting point $\bar{x}$. If the current fitness function value increases, then we say that $\bar{x}$ is an unpromising trial point.*

Figure 4 illustrates an unpromising trial point. Let $\bar{x}_1$ be an incumbent solution and $\hat{x}_1$ be obtained by local search applied to the original function from the starting point $\bar{x}_1$. Then since the modified function value increases after the local search, $\bar{x}_1$ is unpromising.

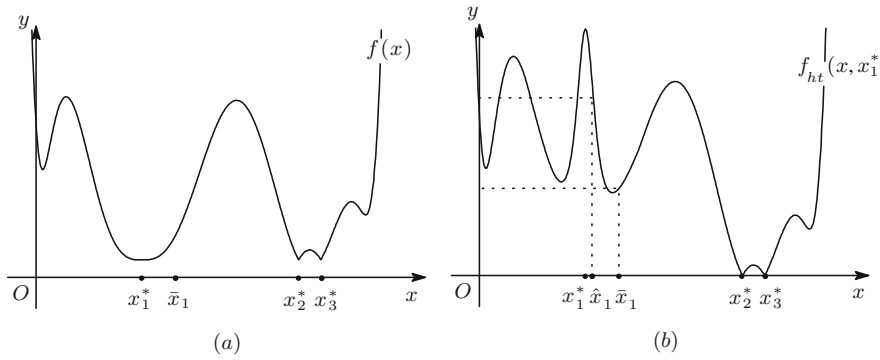


Fig. 4. (**a**) The original objective function and (**b**) a modified function on the global solution x_1^*

3. Suppose $\bar{x}$ is just a semi-global point and it is not an incumbent solution. Then it is quite likely that the point is a local solution. Since it may still attract the population set, we need to modify the function around this point. A similar observation applies when $\bar{x}$ is an unpromising trial point, and we also modify the function. In either case, we set

$$S := S \cup \{\bar{x}\},\ f_{\mathrm{c}}(x) := f_{\mathrm{c}}(x) \cdot \exp\left(\frac{1}{\varepsilon_{\mathrm{t}} + \frac{1}{\rho_{\mathrm{t}}^2}\|x - \bar{x}\|^2}\right).$$

After the modification in the fitness function, the population set P may still have some elements in a neighborhood of the point of modification. So by updating the population set P with some randomly generated points in the search space, we may remove the points lying around the point of modification. Specifically, we double the population set by adding some randomly generated points and redefine the population set by choosing the best half elements of it according to the new fitness function values.

Collecting all the procedures given in this section, we denote by **MOF** $(f_{\mathrm{c}}, \bar{x}, S, S_{\mathrm{inc}}, P)$ the fitness function modification procedure. This procedure yields a new fitness function, which is a modification of the former fitness function f_{c} on $\bar{x}$, with the corresponding changes in the sets S, S_{inc}, and P.

3.2 Population Update Rules

As mentioned earlier, most evolutionary algorithms have the property that the population set tends to cluster around only a few global solutions. Here we propose two different techniques to update the population set, which are aimed to keep diversity while searching for global solutions. The first one is heuristic and depends on the structure of the population set. The second one makes use of some tolerance parameter for the distance between trial points.

Population Update 1. Consider a set of points $X = \{x^1, x^2, \ldots, x^M\}$ sorted according to their objective function values so that $f_c(x^1) \leq f_c(x^2) \leq \cdots \leq f_c(x^M)$. Let x be a trial solution used to update the population set.

1. If $f(x) \geq f(x^M)$, i.e., x is worse than the worst element in X, then discard x.
2. If $f(x) \leq f(x^1)$, i.e., x is better than the best element in X, then add x to X and delete the closest point to x in X.
3. If $f(x^i) \leq f(x) < f(x^{i+1})$, then let

$$k := \underset{1 \leq j \leq i}{\operatorname{argmin}} \|x - x^j\|, \ l := \underset{i+1 \leq j \leq M}{\operatorname{argmin}} \|x - x^j\|,$$

namely, x^k is the closest point to x among those points in X whose objective function values are smaller than $f(x)$, while x^l is the closest point to x among those points in X whose objective function values are greater than $f(x)$.

If $\|x - x^k\| \leq \|x^k - x^l\|$, then discard x.
If $\|x - x^k\| > \|x^k - x^l\|$ and $\|x - x^l\| \leq \|x^k - x^l\|$, then delete x^l from X and add x to X in the $(i+1)$th position. Otherwise, delete x^M from X and add x to X in the $(i+1)$th position.

Population Update 2. Let $X = \{x^1, x^2, \ldots, x^M\}$ be a set of points sorted according to their function values as above, and $\varepsilon_D > 0$ be a fixed tolerance for the distance. Let x be a trial solution. Define

$$B(x, \varepsilon) := \{y \in R^n | \ \|x - y\| < \varepsilon\}, \ k(i) := \underset{1 \leq j \leq i}{\operatorname{argmin}} \|x - x^j\|.$$

1. If $f(x) \leq f(x^1)$, then add x to the set X and delete from X all the points x^j satisfying $x^j \in B(x, \varepsilon_D)$. If there is no such element in X, then delete x^M from X. If there are many, add new trial solutions generated by using the diversification generation method [7] to X to keep the size of the population set P equal to M.
2. If $f(x^i) < f(x) \leq f(x^{i+1})$, then do the following:
   If $x \in B(x^{k(i)}, \varepsilon_D)$, then discard x. Otherwise, add the point x to X, and delete all the elements $x^j, \ j = i+1, \ldots, M$ of X satisfying $x^j \in B(x, \varepsilon_D)$. If there is no such element in X, then delete x^M from X. If there are many, then add new trial solutions generated using the diversification generation method to X to keep the size of the population set P equal to M.

If $\varepsilon_D = 0$, then the Population Update Rule 2 will coincide with the ordinary update rule used in the genetic algorithm that accepts a child to survive if it is better than an element in the population. We denote by *Population Update Rule* $[X, x', x'', ...]$ the procedure of updating the population set by one of the above two rules, where X is the new set obtained by the update using the points x', x'', etc . As we see in the updating process, it always keeps the order of points in the population set.

3.3 HEA Algorithm

We first discuss parameters and procedures that will be used in the algorithm.

M – number of elements in the population set,

m – number of best points to which local search is applied,

l_{s} – maximum number of steps per local search,

$\bar{N}, \beta$ – parameters used to determine semi-global solutions,

$Crossover[(p^1, p^2)]+mutation$ – the mating procedure for the pair (p^1, p^2) and possible mutation for the resulted children pair,

Local search $(f(x), \bar{x}, l_{\mathrm{s}})$ – a local search process for the function $f(x)$ starting from the point $\bar{x}$ with the number of steps l_{s}.

To check whether a point is semi-global or not, we use $\bar{N}$ evolutionary generations and a local search step. Here we use a set B whose elements represent the historical data of the best points in the population set during the last $\bar{N}$ generations.

To terminate the HEA, we use the following three different criteria.

S1. The number of function evaluations exceeds the pre-determined upper limit,

S2. The number of detected global solutions exceeds the pre-determined number,

S3. Let N_s be a pre-specified positive integer. If the most recently added N_{s} elements of the set S of modification points were not new global solutions. If one of those criteria is satisfied, then we terminate the main algorithm. The main loop of the proposed algorithm is stated as follows.

1. Initialization Choose parameters $M, m, l_{\mathrm{s}}, \bar{N}$, and $\beta \in (0, 1)$. Generate the population set P by using the *diversity generation method*. Let the set of modification points and the set of incumbent solutions be $S := \emptyset$ and $S_{\mathrm{inc}} := \emptyset$, respectively. Define the current fitness function as

$$f_{\mathrm{c}}(x) := f(x).$$

Sort the elements in P in ascending order of their current fitness function values, i.e.,

$$f_{\mathrm{c}}(x^1) \leq f_{\mathrm{c}}(x^2) \leq \cdots \leq f_{\mathrm{c}}(x^M).$$

Set the generation counters $t := 1$ and $s := 1$.

2. Parents Pool Generation Generate a parents pool

$$P' := \{(x^i, x^j) | x^i, x^j \in P,\ x^i \neq x^j\}.$$

3. Crossover and Mutation Select a pair $(p^1, p^2) \in P'$ and generate a pair as

$$(c^1, c^2) \longleftarrow \text{Crossover}[(p^1, p^2)] + \text{mutation}.$$

4. Population Update Update the population set by

$$P \longleftarrow \text{Population Update Rule}\,[P, c^1, c^2],\ P' := P' \backslash \{(p^1, p^2)\}.$$

If $P' = \emptyset$, then let

$$N := \min\{s, \bar{N}\},\ B := \{b^1, b^2, \ldots, b^N\} \leftarrow \{x^1, b^1, \ldots, b^{(N-1)}\},\ s := s + 1$$

and go to Step 5; otherwise go to Step 3.
5. Intensification If, during the last $\bar{N}$ generations of evolution, the fitness function has not been modified and the best point in the population set has not been improved enough, i.e.,

$$s \geq \bar{N} \text{ and } \left| f_{\mathrm{c}}(b^{\bar{N}}) - f_{\mathrm{c}}(b^1) \right| \leq \beta\Big(1 + |f_{\mathrm{c}}(b^1)|\Big),$$

then choose $x^1, x^2, ..., x^m \in P$ and for each $x^i,\ i = 1, 2, ..., m$ perform the following procedure:

$$\bar{x}^i \longleftarrow \text{Local search}\,(f(x), x^i, l_{\mathrm{s}}).$$

If x^i is an unpromising trial point, then construct a new fitness function by

$$f_{\mathrm{c}}(x) := \mathbf{MOF}(f_{\mathrm{c}}, x^i, S, S_{\mathrm{inc}}, P).$$

Otherwise, $P := P \backslash \{x^i\}$ and $P \longleftarrow$ Population Update Rule$[P, \bar{x}^i]$. If the fitness function is modified at least once during the above procedure, then set $s := 1$. Go to Step 6.
6. Semi-global Solutions and Modification If $x^1 \in P$ is a semi- global solution, i.e.,

$$s \geq \bar{N} \text{ and } \left| f_{\mathrm{c}}(b^{\bar{N}}) - f_{\mathrm{c}}(x^1) \right| \leq \beta\Big(1 + |f_{\mathrm{c}}(x^1)|\Big),$$

then construct a new fitness function by

$$f_{\mathrm{c}}(x) := \mathbf{MOF}(f_{\mathrm{c}}, \bar{x}, S, S_{\mathrm{inc}}, P) \text{ and set } s := 1.$$

Otherwise, let $B := \{b^1, b^2, \ldots, b^{\bar{N}}\} \leftarrow \{x^1, b^1, \ldots, b^{(\bar{N}-1)}\}$. Proceed to Step 7 with $(f_{\mathrm{c}}(x), P)$.
7. Stopping Condition If one of the stopping conditions holds, then terminate the algorithm and refine the global solutions in S_{inc} by some local search method. Otherwise, set $t := t + 1$ and go to Step 2.

4 Numerical Experiments

The performance of the HEA was tested on a number of well-known global optimization test problems, most of which have multiple solutions. For each problem we made 20 trials with different initial populations. The programming

code for the algorithm was written in MATLAB and run on a computer with Pentium 4 Microprocessor.

For local search in the HEA, we employ MATLAB's command `fmincon`. Unless we provide the gradient or Jacobian of the function, this command performs some derivative-free search. In general, it is difficult to determine universally suitable values of HEA parameters for every problem, because they are highly problem dependent. Nevertheless, through testing many times on various test problems, we suggest possible choices of the parameters as shown in Table 1.

Table 1. Parameter settings

Parameters	definition	Value
M	Number of elements in the population set	$\min\{2n+5, 20\}$
m	Number of best points for which local search is used	2
l_{s}	Maximum number of steps per local search	$\min\{2n, 20\}$
N, β	Parameters controlling local search in HEA	3, 0.001
$N_{\max}$	Maximum number of ineffective local transformations	10
$N_{\mathrm{g\,max}}$	Maximum number of global solutions to be found	20
$NF_{\max}$	Maximum number of function evaluations	$5n10^4$
ε_D	Distance tolerance used in Population Update Rule 2	$n/5$
$\varepsilon_{\mathrm{t}}, \rho_{\mathrm{t}}$	Tunneling parameters used in (2) and (5)	0.1, 2
$\alpha_{\mathrm{h}}, \rho_{\mathrm{h}}$	Humping parameters used in (3) and (5)	1, 0.3

We have two versions of the HEA; HEA_1 and HEA_2 that use Population Update Rule 1 and Rule 2, respectively. We ran the HEA versions for all the chosen test problems with the general parameter settings given in Table 1 and obtained the numerical results shown in Tables 2 and 3. The columns in these tables have the following meanings:

Problem:	name of the test problem,
n:	dimension of the test problem,
$K_{\min}, K_{\mathrm{av}}, K_{\max}$:	minimum, average, maximum numbers of solutions found by the algorithm,
N_{gen}:	average number of generations,
N_{loc}:	average number of local steps taken,
NF:	average number of function evaluations,
N_{f}:	average number of function evaluations when the last global solution is obtained.

The results reported in Tables 2 and 3 indicate that the HEA is promising. For most of the test problems, the average numbers of obtained global solutions (K_{av}) are close to the maximum numbers of obtained global solutions

($K_{\max}$), and this implies that the HEA versions are capable of finding multiple solutions. Moreover, the average numbers of generations are reasonable compared with the problem dimensions and the numbers of obtained global solutions. We observe in both tables that the HEA versions find global solutions in a relatively small number of function evaluations (N_{f}), and after that, the algorithms were still running in order to check whether or not there remains any other solution undiscovered, until one of the termination conditions is met.

Table 2. Numerical results for the HEA with Population Update Rule 1

Problem	n	$K_{\min}$	K_{av}	$K_{\max}$	N_{gen}	N_{loc}	NF	N_{f}
Ackley	5	0	0.7	1	74	336	31,361	10,246
Branin	2	3	3	3	29	48	3,081	1,116
Dixon & price	2	2	2	2	54	139	6,163	1,460
Dixon & price	10	0	1.4	2	103	968	74,387	36,964
Hump	2	2	2	2	46	80	4,918	951
Levy	10	0	0.8	1	149	1312	99,450	23,830
Perm	2	2	2	2	34	81	3,790	1,263
Rosenbrock	10	1	1	1	74	230	47,336	5,987
Shubert	2	14	16.9	18	212	552	24,690	17,796
Trid	6	1	1	1	53	201	12,852	1,665

Table 3. Numerical results for the HEA with Population Update Rule 2

Problem	n	$K_{\min}$	K_{av}	$K_{\max}$	N_{gen}	N_{loc}	NF	N_{f}
Ackley	5	1	1	1	40	140	32,771	21,136
Branin	2	3	3	3	30	49	5,946	2,157
Dixon & price	2	2	2	2	40	70	7,638	3,170
Dixon & price	10	0	1.6	2	70	688	98,166	46,840
Hump	2	2	2	2	45	72	9,216	2,127
Levy	10	1	1	1	96	288	124,414	58,753
Perm	2	2	2	2	20	47	3,552	1,530
Rosenbrock	10	1	1	1	42	152	46,751	9,208
Shubert	2	17	17.8	18	213	524	37,646	25,761
Trid	6	1	1	1	63	168	29,134	15,974

Finally, we make some remarks on the comparison between the results shown in Tables 2 and 3 in terms of the numbers of obtained global solutions and computational costs. Generally, the HEA_1 outperforms its counterpart in the number of function evaluations, while the HEA_2 shows better results than the other in the average number of detected global solutions. For problems with only one solution, the HEA_1 works much better than the HEA_2 for

detecting the global solution, and we can see this fact, for example, by comparing the last columns of the two tables. It is due to the fact that for those problems the whole population set tends to converge to the only solution of the problem after a certain number of generations. However, since the HEA is designed for locating multiple solutions, it tries to keep diversity and removes many points around the solution from the population set. This phenomenon happens repeatedly, and it makes the HEA_2 require more function evaluations. As for HEA_1 the process of keeping diversity works differently, and it depends on the structure of the population set. Moreover, this fact shows that the HEA_2 better fits to problems with multiple solutions. As for locating all solutions of the problem, the HEA_2 is a little more reliable than the HEA_1 as shown in the K_{av} columns of the both tables. Moreover, the HEA_2 requires fewer generations than the HEA_1 in 6 problems out of 10 and requires almost the same amounts for other two problems. For the problem *trid*, the HEA_1 works better than HEA_2 in every aspect, especially in the number of function evaluations. For the problem *levy*, HEA_2 requires fewer generations, but more local searches and function evaluations than the HEA_1. Thus we conclude that HEA_1 and HEA_2 have their own advantages.

5 Conclusions

In this chapter, we have presented a population-based method that aims at finding as many as possible solutions of the global optimization problem. By controlling appropriately the sets of incumbent and modification points, the algorithm is designed to avoid searching in a region around a global solution that has already been obtained. Numerical results for some well-known test problems show that the method can detect multiple global solutions successfully in an acceptable number of function evaluations.

References

1. Barron, C., Gomez, S.: The exponential tunneling method. Reporte de Investigacio'n IIMAS 1(3), 1–23 (1991)
2. Chak, C.K., Feng, G.: Accelerated genetic algorithms: combined with local search techniques for fast and accurate global search. IEEE International Conference on Evolutionary Computation. ICEC95, Perth, Australia, 378–383 (1995)
3. De Jong, K.A.: Evolutionary Computation, MIT Press, Cambridge, MA (2005)
4. Goldberg, D.E.: Genetic Algorithm in Search, Optimization and Machine Learning, Addison-Wesley, Reading, MA (1989)
5. Gomez, S., Solorzano, J., Castellanos, L., Quintana, M.I.: Tunneling and genetic algorithms for global optimization. In: N.Hadjisavvas, P.ardalos, (Eds.), Advances in Convex Analysis and Global Optimization (pp. 553-567), Kluwer Dordrecht (2001)

6. Herrera, F., Lozono, M., Verdegay, J.L.: Tackling real-coded genetic algorithms: operators and tools for behavioural analysis. Artif Intell Rev. 12, 265–319 (1998)
7. Laguna, M., Marti, R.: Scatter Search: Methodology and Implementation in C, Kluwer, Boston, MA (2003)
8. Laguna, M., Marti, R.: Experimental testing of advanced scatter search designs for global optimization of multimodal functions. J. Global Optimi. 33, 235–255 (2005)
9. Levy, A.V., Montalvo, A.: The tunneling algorithm for the global minimization of functions. SIAM J. Sci. Stat. Comput. 6, 15–29 (1985)
10. Levy, A.V., Gomez, S.: The tunneling method applied to global optimization. In: P.T. Boggs, R.H. Byrd, R.B. Schnabel, (Eds.), Numerical Optimization (pp. 213–244), SIAM, Philadelphia, PA (1985)
11. Majig, M., Hedar, A.R., Fukushima, M.: Hybrid evolutionary algorithm for solving general variational inequalities. J. Global Optim. 38, 637–651 (2007)
12. Talbi, E.: A taxonomy of hybrid metaheuristics. J. Heuristics 8, 541–564 (2002)

Gap Functions for Vector Equilibrium Problems via Conjugate Duality

Lkhamsuren Altangerel[1] and Gert Wanka[2]

[1] School of Mathematics and Computer Science, National University of Mongolia, Mongolia
lkal@mathematik.tu-chemnitz.de
[2] Faculty of Mathematics, Chemnitz University of Technology, Germany
gert.wanka@mathematik.tu-chemnitz.de

Summary. This chapter deals with the so-called perturbation approach in the conjugate duality for vector optimization on the basis of weak orderings. As applications, we investigate some new set-valued gap functions for vector equilibrium problems.

Key words: conjugate duality, perturbation approach, vector equilibrium problems, set-valued gap functions

1 Introduction

Tanino and Sawaragi [12] (see also [9]) developed conjugate duality for vector optimization by introducing new concepts of conjugate maps and set-valued subgradients based on Pareto efficiency. Furthermore, by using the concept of the supremum of a set on the basis of weak orderings, the conjugate duality theory was extended to a partially ordered topological vector space by Tanino [14] and to set-valued vector optimization problems by Song [10, 11], respectively.

Dealing with conjugacy notions in the framework of set-valued optimization, the so-called perturbation approach in the conjugate duality (see [15]) has been extended to the constrained vector optimization problems (cf. [2]). As applications, rewriting the vector variational inequality in the form of a vector optimization problem, new set-valued gap functions for the vector variational inequality have been introduced.

By using a special perturbation function, the Fenchel-type dual problem for vector optimization has been obtained and based on this investigation some set-valued mappings have been introduced in order to apply them to variational principles for vector equilibrium problems (see [3]). Notice that variational principles for vector equilibrium problems have been investigated first in [4] and [5]. Some related results in the scalar case can be found in [1] and [6].

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_10,

In this chapter we consider two additional perturbation functions implying the Lagrange and Fenchel–Lagrange type dual problems, respectively.

This chapter is organized as follows. In Section 2 we give some preliminary results dealing with conjugate duality for vector optimization and stability criteria. On the basis of two special perturbation functions different dual problems are introduced in Section 3. In order to state the strong duality, we use in Section 3 general results due to Song. Finally, as applications some new gap functions for vector equilibrium problems related to conjugate duality are introduced in Section 4.

2 Mathematical Preliminaries

Let Y be a real topological vector space partially ordered by a pointed closed convex cone C with a nonempty interior int C in Y. For any $\xi, \mu \in Y$, we use the following ordering relations:

$$\begin{aligned}
&\xi \leq \mu \Leftrightarrow \mu - \xi \in C;\\
&\xi < \mu \Leftrightarrow \mu - \xi \in \operatorname{int} C;\\
&\xi \not< \mu \Leftrightarrow \mu - \xi \notin \operatorname{int} C.
\end{aligned}$$

The relations $\geq$, $>$, and $\not>$ are defined similarly. Let us now introduce the weak maximum and weak supremum of a set Z in the space $\overline{Y}$ induced by adding to Y two imaginary points $+\infty$ and $-\infty$. We suppose that $-\infty < y < +\infty$ for $y \in Y$. Moreover, we use the following conventions

$$\begin{aligned}
&(\pm\infty) + y = y + (\pm\infty) = \pm\infty \text{ for all } y \in Y,\ (\pm\infty) + (\pm\infty) = \pm\infty,\\
&\lambda(\pm\infty) = \pm\infty \text{ for } \lambda > 0, \text{ and } \lambda(\pm\infty) = \mp\infty \text{ for } \lambda < 0.
\end{aligned}$$

The sum $+\infty + (-\infty)$ is not considered, since we can avoid it.

For a given set $Z \subseteq \overline{Y}$, we define *the set $A(Z)$ of all points above Z* and *the set $B(Z)$ of all points below Z* by

$$A(Z) = \left\{ y \in \overline{Y} |\ y > y' \text{ for some } y' \in Z \right\}$$

and

$$B(Z) = \left\{ y \in \overline{Y} |\ y < y' \text{ for some } y' \in Z \right\},$$

respectively. Clearly $A(Z) \subseteq Y \cup \{+\infty\}$ and $B(Z) \subseteq Y \cup \{-\infty\}$.

Definition 2.1

(i) A point $\widehat{y} \in \overline{Y}$ is said to be a weak maximal point of $Z \subseteq \overline{Y}$ if $\widehat{y} \in Z$ and $\widehat{y} \notin B(Z)$, that is, if $\widehat{y} \in Z$ and there is no $y' \in Z$ such that $\widehat{y} < y'$.

(ii) A point $\widehat{y} \in \overline{Y}$ is said to be a weak supremal point of $Z \subseteq \overline{Y}$ if $\widehat{y} \notin B(Z)$ and $B(\{\widehat{y}\}) \subseteq B(Z)$, that is, if there is no $y \in Z$ such that $\widehat{y} < y$ and if the relation $y' < \widehat{y}$ implies the existence of some $y \in Z$ such that $y' < y$.

Weak minimal and weak infimal points can be defined analogously. The set of all weak maximal (minimal) and weak supremal (infimal) points of Z is denoted by $\operatorname{WMax} Z$ ($\operatorname{WMin} Z$) and $\operatorname{WSup} Z$ ($\operatorname{WInf} Z$), respectively. Remark that $\operatorname{WMax} Z = Z \cap \operatorname{WSup} Z$. Moreover, $-\operatorname{WMax}(-Z) = \operatorname{WMin} Z$ and $-\operatorname{WSup}(-Z) = \operatorname{WInf} Z$ hold. For more properties of these sets we refer to [13] and [14].

Now we give some definitions of the conjugate mapping and the subgradient of a set-valued mapping based on the weak supremum and the weak maximum of a set. Let X be another real topological vector space and let $\mathcal{L}(\mathrm{X}, \mathrm{Y})$ be the space of all linear continuous operators from X to Y. For $x \in X$ and $l \in \mathcal{L}(\mathrm{X}, \mathrm{Y})$, $\langle l, x\rangle$ denotes the value of l at x.

Definition 2.2 *(Tanino [14]). Let $G : X \rightrightarrows \overline{Y}$ be a set-valued mapping.*

(i) A set-valued mapping $G^* : \mathcal{L}(\mathrm{X}, \mathrm{Y}) \rightrightarrows \overline{Y}$ defined by

$$G^*(T) = \operatorname{WSup} \bigcup_{x \in X} \left[\langle T, x\rangle - G(x)\right], \text{ for } T \in \mathcal{L}(\mathrm{X}, \mathrm{Y})$$

is called the conjugate mapping of G.

(ii) A set-valued mapping $G^{**} : X \rightrightarrows \overline{Y}$ defined by

$$G^{**}(x) = \operatorname{WSup} \bigcup_{T \in \mathcal{L}(\mathrm{X}, \mathrm{Y})} \left[\langle T, x\rangle - G^*(T)\right], \text{ for } x \in X$$

is called the biconjugate mapping of G.

(iii) $T \in \mathcal{L}(\mathrm{X}, \mathrm{Y})$ is said to be a subgradient of the set-valued mapping G at $(x_0; y_0)$ if $y_0 \in G(x_0)$ and

$$\langle T, x_0\rangle - y_0 \in \operatorname{WMax} \bigcup_{x \in X} \left[\langle T, x\rangle - G(x)\right].$$

The set of all subgradients of G at $(x_0; y_0)$ is called *the subdifferential of G at $(x_0; y_0)$* and is denoted by $\partial G(x_0; y_0)$. If $\partial G(x_0; y_0) \neq \varnothing$ for every $y_0 \in G(x_0)$, then G is said to be *subdifferentiable at x_0*.

Let X and Y be real topological vector spaces. Assume that $\overline{Y}$ is the extended space of Y and $h : X \to Y \cup \{+\infty\}$ is a given function. We consider the vector optimization problem

$$(P) \qquad \operatorname{WInf}\{h(x) \,|\, x \in X\}.$$

Based on a perturbation approach (see [14]), a dual problem to (P) can be defined as follows:

$$(D) \qquad \operatorname{WSup} \bigcup_{\Lambda \in \mathcal{L}(U, Y)} \left[-\Phi^*(0, \Lambda)\right],$$

where $\Phi : X \times U \to Y \cup \{+\infty\}$ is called a perturbation function having the property that

$$\Phi(x, 0) = h(x) \ \forall x \in X.$$

Here, U is another real topological vector space. Moreover, the conjugate mapping of Φ is

$$\Phi^*(T, \Lambda) = \mathrm{WSup}\left\{\langle T, x\rangle + \langle \Lambda, u\rangle - \Phi(x, u)|\ x \in X,\ u \in U\right\}$$

for $T \in \mathcal{L}(\mathrm{X}, \mathrm{Y})$ and $\Lambda \in \mathcal{L}(\mathrm{U}, \mathrm{Y})$.

Proposition 2.1 (*Tanino [14]*) (*Weak duality*)
For any $x \in X$ and $\Lambda \in \mathcal{L}(\mathrm{U}, \mathrm{Y})$ it holds

$$\Phi(x, 0) \notin B\left(-\Phi^*(0, \Lambda)\right).$$

Definition 2.3 *(Tanino [14]). The primal problem (P) is said to be stable with respect to Φ if the value mapping $\Psi : U \rightrightarrows \overline{Y}$ defined by*

$$\Psi(u) = \mathrm{WInf}\left\{\Phi(x, u)|\ x \in X\right\}$$

is subdifferentiable at 0.

Theorem 2.1 *(Tanino [14], Song [10]). If the problem (P) is stable with respect to Φ, then*

$$\mathrm{WInf}(P) = \mathrm{WSup}(D) = \mathrm{WMax}(D).$$

Let us now mention some definitions and assertions related to the stability. For a given set-valued mapping $G : X \rightrightarrows Y \cup \{+\infty\}$, we have

- *effective domain of G*: $\mathrm{dom}\ G = \{x \in X|\ G(x) \neq \varnothing,\ G(x) \neq \{+\infty\}\}$,
- *epigraph of G*: $\mathrm{epi}\ G = \{(x, y) \in X \times Y|\ y \in G(x) + C\}$.

In particular, if $g : X \to Y \cup \{+\infty\}$ is a vector-valued function, then its effective domain and epigraph are defined as

$$\mathrm{epi}\ g = \{(x, y) \in X \times Y|\ g(x) \leq y\},$$
$$\mathrm{dom}\ g = \{x \in X|\ g(x) \neq +\infty\},$$

respectively. The function g is said to be *proper* if $g(x) \in X \cup \{+\infty\}$ and $g \not\equiv +\infty$.

A set-valued mapping $G : X \rightrightarrows Y \cup \{+\infty\}$ is said to be C-convex if its epigraph is convex. A given set-valued mapping $G : X \rightrightarrows Y \cup \{+\infty\}$ is C-convex if and only if for all $\lambda \in [0, 1]$ and $x_1, x_2 \in X$

$$\lambda G(x_1) \cap Y + (1 - \lambda)G(x_2) \cap Y \subseteq G(\lambda x_1 + (1 - \lambda)x_2) \cap Y + C.$$

In particular, if $g : X \to Y \cup \{+\infty\}$ is a proper vector-valued function, then g is C-convex if and only if for all $\lambda \in (0, 1)$ and $x_1, x_2 \in X,\ x_1 \neq x_2$

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \in g(\lambda x_1 + (1 - \lambda)x_2) + C.$$

Proposition 2.2 *(Song [10]). Let $G : X \rightrightarrows Y \cup \{+\infty\}$ be a C-convex set-valued mapping with $int(epi\ \ G) \neq \varnothing$. If $x_0 \in int(dom\ \ G)$ and $G(x_0) \subseteq \mathrm{WInf}\, G(x_0)$, then G is subdifferentiable at x_0.*

Definition 2.4

(i) A set-valued mapping $G : X \rightrightarrows Y \cup \{+\infty\}$ is said to be C-Hausdorff lower continuous at $x_0 \in X$ if for every neighborhood V of zero in Y there exists a neighborhood U of zero in X such that

$$G(x_0) \subseteq G(x) + V + C\ \ \forall x \in (x_0 + U) \cap dom\ \ G.$$

(ii) A set-valued mapping $G : X \rightrightarrows Y \cup \{+\infty\}$ is said to be weakly C-upper bounded on a set $A \subseteq X$ if there exists a point $b \in Y$ such that $(x, b) \in epi\ \ G,\ \forall x \in A$.

Let us remark that G is weakly C-upper bounded on a set $A \subseteq X$ if and only if there exists a point $b \in Y$ such that $G(x) \cap (b - C) \neq \varnothing\ \forall x \in A$.

Proposition 2.3 *(Song [10]). Let $G : X \rightrightarrows Y \cup \{+\infty\}$ be a set-valued mapping.*

1. *Then the following assertions are equivalent.*
   (i) $int(epi\ \ G) \neq \varnothing$.
   (ii) $\exists x_0 \in int(dom\ \ G)$ such that G is weakly C-upper bounded on some neighborhood of x_0.
2. *If G is C-Hausdorff lower continuous on $int(dom\ \ G)$, then (i) and (ii) hold.*

Proposition 2.4 *(Tanino [14]). If the perturbation function $\Phi : X \times U \to Y \cup \{+\infty\}$ is C-convex, then the value mapping Ψ is a C-convex set-valued mapping.*

Proposition 2.5 *(Song [11]). Let $\Phi : X \times U \to Y \cup \{+\infty\}$ be a C-convex vector-valued function and the value mapping Ψ be weakly C-upper bounded on a neighborhood of zero in U. Then the problem (P) is stable with respect to Φ.*

Remark 1. Proposition 2.5 was proved in [11] in the more general case when $\Phi : X \times U \to Y \cup \{+\infty\}$ is a set-valued mapping.

3 The Constrained Vector Optimization Problem

3.1 Different Dual Problems

Assume that $h : X \to Y \cup \{+\infty\}$ is a given function and $G \subseteq X$. We consider the constrained vector optimization problem

$$(P_{\mathrm{c}}) \qquad \mathrm{WInf}\{h(x)|\, x \in G\}.$$

By using the perturbation function $\Phi_{\mathrm{F}} : X \times X \to Y \cup \{+\infty\}$ defined by

$$\Phi_{\mathrm{F}}(x,u) = \begin{cases} h(x+u), \text{ if } x \in G, \\ +\infty, \qquad \text{otherwise,} \end{cases}$$

the Fenchel dual problem to (P_c) has been stated as follows (cf. [3]):

$$(D_{\mathrm{F}}) \qquad \mathrm{WSup} \bigcup_{T \in \mathcal{L}(\mathrm{X},\mathrm{Y})} \mathrm{WInf}\left\{-h^*(T) + \{\langle T, x\rangle|\ x \in G\}\right\}.$$

Proposition 3.1 (*Weak duality*)
For any $x \in G$ and $T \in \mathcal{L}(\mathrm{X},\mathrm{Y})$ it holds

$$h(x) \notin B\left(-\Phi_{\mathrm{F}}^*(0,T)\right).$$

Let U be a real topological vector space, $D \subseteq U$ be a pointed closed convex cone, $M \subseteq X$, and $g : X \to U \cup \{+\infty\}$. If the feasible set G is given by

$$G = \{x \in M|\ g(x) \in -D\},$$

then one can consider the following two perturbation functions (cf. [2] and [15])

$$\Phi_{\mathrm{L}} : X \times U \to Y \cup \{+\infty\},\ \Phi_{\mathrm{L}}(x,u) = \begin{cases} h(x), & x \in M,\ g(x) \in -D + u, \\ +\infty, & \text{otherwise,} \end{cases}$$

and

$$\Phi_{\mathrm{FL}} : X \times X \times U \to Y\ \cup\{+\infty\},$$
$$\Phi_{\mathrm{FL}}(x,v,u) = \begin{cases} h(x+v), & x \in M,\ g(x) \in -D + u, \\ +\infty, & \text{otherwise.} \end{cases}$$

In analogy to Proposition 3.3 and Proposition 3.11 in [2], the following assertion can be shown easily.

Proposition 3.2 *Let $\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})$ and $T \in \mathcal{L}(\mathrm{X},\mathrm{Y})$. Then*

(i) $\Phi_{\mathrm{L}}^*(0,\Lambda) = \mathrm{WSup}\left\{\{\langle \Lambda, u\rangle|\ u \in D\} + \{\langle \Lambda, g(x)\rangle - h(x)|\ x \in M\}\right\}.$

(ii) $\Phi_{\mathrm{FL}}^*(0,T,\Lambda) = \mathrm{WSup}\left\{\{\langle \Lambda, u\rangle|\ u \in D\}\right.$
$\left. + \{\langle T, v\rangle - h(v)|\ v \in X\} + \{\langle \Lambda, g(x)\rangle - \langle T, x\rangle|\ x \in M\}\right\}.$

Remark 2. According to Proposition 2.6 in [14], we can use for $\Phi_{\mathrm{L}}^*(0,\Lambda)$ and $\Phi_{\mathrm{FL}}^*(0,T,\Lambda)$ some equivalent formulations. For instance, for $\Phi_{FL}^*(0,T,\Lambda)$ we have

$$\begin{aligned} \Phi_{\mathrm{FL}}^*(0,T,\Lambda) &= \mathrm{WSup}\left\{\{\langle \Lambda, u\rangle|\ u \in D\}\right. \\ &\quad \left. + \{\langle T, v\rangle - h(v)|\ v \in X\} + \{\langle \Lambda, g(x)\rangle - \langle T, x\rangle|\ x \in M\}\right\} \\ &= \mathrm{WSup}\left\{\mathrm{WSup}\{\langle \Lambda, u\rangle|\ u \in D\}\right. \\ &\quad \left. + h^*(T) + \{\langle \Lambda, g(x)\rangle - \langle T, x\rangle|\ x \in M\}\right\}. \end{aligned}$$

As a consequence of Proposition 3.2 can be stated the Lagrange dual problem to (P_c)

$$(D_{\mathrm{L}}) \quad \mathrm{WSup} \bigcup_{\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})} [-\Phi_{\mathrm{L}}^*(0,\Lambda)]$$
$$= \mathrm{WSup} \bigcup_{\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})} \mathrm{WInf}\left\{\{-\langle \Lambda, u\rangle |\ u \in D\} + \{h(x) - \langle \Lambda, g(x)\rangle |\ x \in M\}\right\}$$

and the Fenchel–Lagrange dual problem

$$(D_{\mathrm{FL}}) \quad \mathrm{WSup} \bigcup_{(T,\Lambda) \in \mathcal{L}(\mathrm{X},\mathrm{Y}) \times \mathcal{L}(\mathrm{U},\mathrm{Y})} [-\Phi_{\mathrm{FL}}^*(0,T,\Lambda)]$$
$$= \mathrm{WSup} \bigcup_{(T,\Lambda) \in \mathcal{L}(\mathrm{X},\mathrm{Y}) \times \mathcal{L}(\mathrm{U},\mathrm{Y})} \mathrm{WInf}\left\{\{h(v) - \langle T, v\rangle |\ v \in X\}\right.$$
$$\left. + \{-\langle \Lambda, u\rangle |\ u \in D\} + \{\langle T, x\rangle - \langle \Lambda, g(x)\rangle |\ x \in M\}\right\},$$

respectively.

Proposition 3.3 (*Weak duality*)

(i) For any $x \in G$ and $T \in \mathcal{L}(\mathrm{X},\mathrm{Y})$ it holds

$$h(x) \notin B\left(-\Phi_{\mathrm{L}}^*(0,\Lambda)\right).$$

(ii) For any $x \in G$ and $(T,\Lambda) \in \mathcal{L}(\mathrm{X},\mathrm{Y}) \times \mathcal{L}(\mathrm{U},\mathrm{Y})$ it holds

$$h(x) \notin B\left(-\Phi_{\mathrm{FL}}^*(0,T,\Lambda)\right).$$

3.2 Stability and Strong Duality

This section deals with some stability assertions associated with the presented perturbation functions as special cases of general results due to Song [10] and [11]. In order to investigate stability criteria, let us notice that the value mappings with respect to Φ_{F}, Φ_{L}, and Φ_{FL} turn out to be

$$\begin{aligned}
\Psi_{\mathrm{L}} : U \rightrightarrows \overline{Y},\ \Psi_{\mathrm{L}}(u) &= \mathrm{WInf}\left\{\Phi_{\mathrm{L}}(x,u) |\ x \in X\right\} \\
&= \mathrm{WInf}\left\{h(x) |\ x \in M,\ g(x) \in -D + u\right\}, \\
\Psi_{F} : X \rightrightarrows \overline{Y},\ \Psi_{\mathrm{F}}(v) &= \mathrm{WInf}\left\{\Phi_{\mathrm{F}}(x,v) |\ x \in X\right\} \\
&= \mathrm{WInf}\left\{h(x+v) |\ x \in G\right\}, \\
\Psi_{\mathrm{FL}} : X \times U \rightrightarrows \overline{Y},\ \Psi_{\mathrm{FL}}(v,u) &= \mathrm{WInf}\left\{\Phi_{\mathrm{FL}}(x,v,u) |\ x \in X\right\} \\
&= \mathrm{WInf}\left\{h(x+v) |\ x \in M,\ g(x) \in -D + u\right\},
\end{aligned}$$

respectively.

Proposition 3.4 *Let $M \subseteq X$ be a convex set and $h : X \to Y \cup \{+\infty\}$, $g : X \to U$ be C- and D-convex functions, respectively. Then the value mappings Ψ_{L}, Ψ_{F}, and Ψ_{FL} are convex.*

Proof. Under the stated assumptions of convexity one can easily verify that the perturbation functions Φ_{L}, Φ_{F}, and Φ_{FL} are convex. Then the desired assertions follow from Proposition 2.4. □

Theorem 3.1 *Let $M \subseteq X$ be a convex set and $h : X \to Y \cup \{+\infty\}$, $g : X \to U$ be C- and D-convex functions, respectively. Suppose that the value mapping Ψ_{F} (resp. Ψ_{L} and Ψ_{FL}) is weakly C-upper bounded on a neighborhood of zero in X. Then the problem (P_{c}) is stable with respect to Φ_{F} (resp. Φ_{L} and Φ_{FL}).*

Proof. By Proposition 3.4 the value mapping Ψ_F (resp. Ψ_L and Ψ_{FL}) is convex. Then the stability of the problem (P_{c}) follows from Proposition 2.5. □

Proposition 3.5 *If there exists some $x_0 \in \mathrm{dom}\ \ h \cap G$ such that the function h is weakly C-upper bounded on some neighborhood of x_0, then the value mapping Ψ_{F} is weakly C-upper bounded on some neighborhood of zero in X.*

Proof. Since h is weakly C-upper bounded on some neighborhood of $x_0 \in \mathrm{dom}\ \ h \cap G$, there exists a neighborhood $V_0 \subseteq X$ of zero and $\exists b \in Y$ such that

$$(x_0 + v, b) \in \mathrm{epi}\ \ h\ \forall v \in V_0,$$

or, equivalently,

$$h(x_0 + v) \leq b\ \forall v \in V_0.$$

Hence $h(x_0 + v) \in b - C,\ \forall v \in V_0$.
On the other hand, by Corollary 2.1 in [14], we obtain that for any $v \in V_0$

$$\{h(x + v) |\ x \in G\} \subseteq \Psi_{\mathrm{F}}(v) \cup A(\Psi_{\mathrm{F}}(v)).$$

In particular, $h(x_0 + v) \in \Psi_{\mathrm{F}}(v) \cup A(\Psi_{\mathrm{F}}(v))\ \forall v \in V_0$ holds.

a. If $h(x_0 + v) \in \Psi_{\mathrm{F}}(v)$, then $(b - C) \cap \Psi_{\mathrm{F}}(v) \neq \varnothing\ \forall v \in V_0$.
b. If $h(x_0 + v) \in A(\Psi_{\mathrm{F}}(v))$, then $\exists \bar{y} \in \Psi_{\mathrm{F}}(v)$ such that $h(x_0 + v) > \bar{y}$.

Therefore,

$$\bar{y} \in h(x_0 + v) - \mathrm{int}\ \ C \subseteq h(x_0 + v) - C \subseteq b - C - C \subseteq b - C,$$

which means that also $(b - C) \cap \Psi_{\mathrm{F}}(v) \neq \varnothing\ \forall v \in V_0$. The proof is completed. □

Proposition 3.6 *If there exists some $x_0 \in \mathrm{dom}\ \ h \cap M$ such that $0 \in \mathrm{int}(g(x_0) + D)$, then the value mapping Ψ_{L} is weakly C-upper bounded on some neighborhood of zero in X.*

Proof. As $0 \in \text{int}(g(x_0) + D)$, there exists a neighborhood U_0 of zero such that $u \in g(x_0) + D,\ \forall u \in U_0 \subseteq U$. This means that $g(x_0) \in -D + u\ \forall u \in U_0$. Let us notice that because $h(x_0) \neq +\infty,\ \exists b \in Y$ such that $h(x_0) \leq b$.
By Corollary 2.1 in [14], for any $u \in U_0$ one has

$$\{h(x)|\ x \in M,\ g(x) \in -D + u\} \subseteq \Psi_{\mathrm{L}(u)} \cup A(\Psi_{\mathrm{L}}(u)).$$

In particular, it holds $h(x_0) \in \Psi_{\mathrm{L}}(u) \cup A(\Psi_{\mathrm{L}}(u))\ \forall u \in U_0$.

a. If $h(x_0) \in \Psi_{\mathrm{L}}(u)$, then $(b - C) \cap \Psi_{\mathrm{L}}(u) \neq \varnothing\ \forall u \in U_0$.
b. If $h(x_0) \in A(\Psi_{\mathrm{L}}(u))$, then $\exists \bar{y} \in \Psi_{\mathrm{L}}(u)$ such that $h(x_0) > \bar{y}$. Therefore,

$$\bar{y} \in h(x_0) - \text{int}\ C \subseteq b - C - \text{int}\ C \subseteq b - C,$$

which means that also $(b - C) \cap \Psi_{\mathrm{L}}(u) \neq \varnothing\ \forall u \in U_0$.

□

Combining the assumptions of Propositions 3.5 and 3.6, we easily show the following assertion.

Proposition 3.7 *If there exists some $x_0 \in \text{dom}\ \ h \cap M$ such that $0 \in \text{int}(g(x_0) + D)$ and the function h is weakly C-upper bounded on some neighborhood of x_0, then the value mapping Ψ_{FL} is weakly C-upper bounded on some neighborhood of zero in X.*

Theorem 3.2 *Let $M \subseteq X$ be a convex set and $h : X \to Y \cup \{+\infty\},\ g : X \to U$ be C- and D-convex functions, respectively.*

(i) If there exists some $x_0 \in \text{dom}\ \ h \cap G$ such that the function h is weakly C-upper bounded on some neighborhood of x_0, then

$$\text{WInf}(P_c) = \text{WSup}(D_F) = \text{WMax}(D_F).$$

(ii) If there exists some $x_0 \in \text{dom}\ \ h \cap M$ such that $0 \in \text{int}(g(x_0) + D)$, then

$$\text{WInf}(P_c) = \text{WSup}(D_L) = \text{WMax}(D_L).$$

(iii) If there exists some $x_0 \in \text{dom}\ \ h \cap M$ such that $0 \in \text{int}(g(x_0) + D)$ and the function h is weakly C-upper bounded on some neighborhood of x_0, then

$$\begin{aligned}\text{WInf}(P_{\mathrm{c}}) &= \text{WSup}(D_{\mathrm{F}}) = \text{WSup}(D_{\mathrm{L}}) = \text{WSup}(D_{\mathrm{FL}}) \\ &= \text{WMax}(D_{\mathrm{F}}) = \text{WMax}(D_{\mathrm{L}}) = \text{WMax}(D_{\mathrm{FL}}).\end{aligned}$$

Proof. Under the assumptions and by Theorem 3.1 the problem (P_{c}) is stable with respect to Φ_{F} (resp. Φ_{L} and Φ_{FL}). Therefore, according to Theorem 2.1 one obtains the desired assertions. □

4 Gap Functions for Vector Equilibrium Problems

Let X and Y be real topological vector spaces. Assume that K is a nonempty convex set in X and $f : K \times K \to Y$ is a bifunction such that $f(x,x) = 0\ \forall x \in K$. We consider the vector equilibrium problem which consists in finding $x \in K$ such that

$$(VEP) \qquad f(x,y) \not< 0\ \forall y \in K.$$

By K^p we denote the solution set of (VEP). In analogy to the vector variational inequality, we can give the definition of a gap function for (VEP).

Definition 4.1 *(Chen et al. [7] and Goh and Yang [8]) A set-valued mapping $\gamma : K \rightrightarrows Y \cup \{+\infty\}$ is said to be a gap function for (VEP) if it satisfies the following conditions:*

(i) $0 \in \gamma(x)$ *if and only if* $x \in K$ *solves the problem* (VEP);
(ii) $0 \not> \gamma(y)\ \forall y \in K$.

According to [3], let us remark that $\bar{x} \in K$ is a solution to (VEP) if and only if 0 is a weak minimal point of the set $\{f(\bar{x},y)|\ y \in K\}$. Rewriting the problem (VEP) into the vector optimization problem

$$(P^{VEP};x) \qquad \mathrm{WInf}\,\{f(x,y)|\ y \in K\},$$

where $x \in X$ is fixed, and using the Fenchel dual problem to $(P^{VEP};x)$, let us introduce the following mapping

$$\gamma_{\mathrm{F}}^{VEP}(x) := \bigcup_{T \in \mathcal{L}(\mathrm{X},\mathrm{Y})} \widetilde{\Phi}_{\mathrm{F}}^{*}(0,T;x),$$

where $\widetilde{\Phi}_{\mathrm{F}}^{*}(0,T;x) = \mathrm{WSup}\,\{\{\langle T,y\rangle - f(x,y)|\ y \in K\} + \{-\langle T,y\rangle|\ y \in K\}\}$, that is,

$$\gamma_{\mathrm{F}}^{VEP}(x) = \bigcup_{T \in \mathcal{L}(\mathrm{X},\mathrm{Y})} \mathrm{WSup}\,\{\{\langle T,y\rangle - f(x,y)|\ y \in K\} + \{-\langle T,y\rangle|\ y \in K\}\}.$$

Theorem 4.1 *Let $f(x,\cdot) : K \to Y$ be a convex function for all $x \in K$. Assume that for all $x \in K^p$ there exists some $y_0 \in K$ such that the function $f(x,\cdot)$ is weakly C-upper bounded on some neighborhood of y_0. Then $\gamma_{\mathrm{F}}^{VEP}$ is a gap function for (VEP).*

Proof. Under the assumptions it is clear that the problem $(P^{VEP};x)$ is stable. Consequently, the desired assertion follows from Lemma 1 and Theorem 1(i) in [3]. □

Let the ground set K be nonempty and given by

$$K = \{x \in X |\ g(x) \in -D\}, \tag{1}$$

where $D \subseteq U$ is a pointed closed convex cone, U is a real topological vector space and $g : X \to U \cup \{+\infty\}$. Let $x \in X$ be fixed. Taking $f(x, \cdot)$ instead of h in (D_{L}) and (D_{FL}), respectively, the Lagrange and the Fenchel–Lagrange dual problems can be written as follows:

$$(D_{\mathrm{L}}^{VEP}; x) \ \ \mathrm{WSup} \bigcup_{\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})} \left[-\widetilde{\Phi}_{\mathrm{L}}^{*}(0, \Lambda; x)\right]$$

$$(D_{\mathrm{FL}}^{VEP}; x) \ \ \mathrm{WSup} \bigcup_{(T,\Lambda) \in \mathcal{L}(\mathrm{X},\mathrm{Y}) \times \mathcal{L}(\mathrm{U},\mathrm{Y})} \left[-\widetilde{\Phi}_{\mathrm{FL}}^{*}(0, T, \Lambda; x)\right],$$

where

$$\widetilde{\Phi}_{\mathrm{L}}^{*}(0, \Lambda; x) := \mathrm{WSup}\left\{\{\langle \Lambda, u\rangle |\ u \in D\} + \{\langle \Lambda, g(y)\rangle - f(x, y) |\ y \in X\}\right\}, \tag{2}$$

and

$$\begin{aligned}\widetilde{\Phi}_{\mathrm{FL}}^{*}(0, T, \Lambda; x) :=\ & \mathrm{WSup}\left\{\{\langle T, y\rangle - f(x, y) |\ y \in X\}\right. \\ & \left. + \{\langle \Lambda, u\rangle |\ u \in D\} + \{\langle \Lambda, g(y)\rangle - \langle T, y\rangle |\ y \in X\}\right\}. \end{aligned} \tag{3}$$

Consequently, we can introduce two set-valued mappings

$$\gamma_{\mathrm{L}}^{VEP}(x) := \bigcup_{\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})} \widetilde{\Phi}_{\mathrm{L}}^{*}(0, \Lambda; x)$$

and

$$\gamma_{\mathrm{FL}}^{VEP}(x) := \bigcup_{(T,\Lambda) \in \mathcal{L}(\mathrm{X},\mathrm{Y}) \times \mathcal{L}(\mathrm{U},\mathrm{Y})} \widetilde{\Phi}_{\mathrm{FL}}^{*}(0, T, \Lambda; x).$$

Theorem 4.2 *Let the functions $f(x, \cdot) : K \to Y$, $x \in K$ and $g : X \to Y$ be convex. Assume that there exists $y_0 \in K$ such that $0 \in int(g(y_0) + D)$. Then $\gamma_{\mathrm{L}}^{VEP}$ is a gap function for (VEP).*

Proof. (i) Let $\bar{x} \in K$ be a solution to (VEP), then by Theorem 3.2(ii), one has

$$0 \in \mathrm{WInf}(P^{VEP}; \bar{x}) = \mathrm{WMax}\left(D_{\mathrm{L}}^{VEP}; \bar{x}\right).$$

Consequently,

$$0 \in \mathrm{WMax}[-\gamma_{L}^{VEP}(\bar{x})].$$

Whence $0 \in \gamma_{\mathrm{L}}^{VEP}(\bar{x})$. Conversely, let

$$\begin{aligned}0 \in \gamma_{\mathrm{L}}^{VEP}(\bar{x}) = \bigcup_{\Lambda \in \mathcal{L}(\mathrm{U},\mathrm{Y})} & \mathrm{WSup}\left\{\{\langle \Lambda, u\rangle |\ u \in D\} + \{\langle \Lambda, g(y)\rangle\right. \\ & \left. - f(\bar{x}, y) |\ y \in X\}\right\}. \end{aligned}$$

Then $\exists \overline{\Lambda} \in \mathcal{L}(\mathrm{U}, \mathrm{Y})$ such that

$$0 \in \mathrm{WSup}\left\{\{\langle \overline{\Lambda}, u\rangle |\ u \in D\} + \{\langle \overline{\Lambda}, g(y)\rangle - f(\bar{x}, y) |\ y \in X\}\right\},$$

or, equivalently,

$$0 \in \mathrm{WInf}\left\{\{-\langle \overline{\Lambda}, u\rangle |\ u \in D\} + \{f(\bar{x}, y) - \langle \overline{\Lambda}, g(y)\rangle |\ y \in X\}\right\}. \quad (4)$$

Assume that $0 \notin \mathrm{WMin}\{f(\bar{x}, y) |\ y \in K\}$. This means that $\exists \bar{y} \in K$ such that $f(\bar{x}, \bar{y}) < 0$. In other words, we have

$$f(\bar{x}, \bar{y}) - \langle \overline{\Lambda}, g(\bar{y})\rangle + \langle \overline{\Lambda}, g(\bar{y})\rangle < 0,$$

which contradicts (4) since $g(\bar{y}) \in -D$.

(ii) Let $x \in K$ be fixed and $z \in \gamma_{\mathrm{L}}^{VEP}(x)$. Then $\exists \overline{\Lambda} \in \mathcal{L}(\mathrm{U}, \mathrm{Y})$ such that

$$z \in \mathrm{WSup}\left\{\{\langle \overline{\Lambda}, u\rangle |\ u \in D\} + \{\langle \overline{\Lambda}, g(y)\rangle - f(x, y) |\ y \in X\}\right\}.$$

Choosing $y := x$ and $u := -g(x) \in D$, we obtain that

$$\langle \overline{\Lambda}, -g(x)\rangle + \langle \overline{\Lambda}, g(x)\rangle - f(x, x) = 0$$

is an element of the set defined within the outer braces. Therefore z as an element of the set of the weak supremal points of this set cannot be less than zero with respect to the partial ordering given by the cone C, i.e., $z \not< 0$. Consequently, one has $\gamma_{\mathrm{L}}^{VEP}(x) \not< 0\ \forall x \in K$. □

Analogously, we can verify the following assertion concerning $\gamma_{\mathrm{FL}}^{VEP}$.

Theorem 4.3 *Let the functions $f(x, \cdot) : K \to Y$, $x \in K$ and $g : X \to Y$ be convex. Assume that there exists some $y_0 \in K$ such that $0 \in int(g(y_0) + D)$ and the function $f(x, y)$ is weakly C-upper bounded with respect to y on some neighborhood of y_0. Then $\gamma_{\mathrm{FL}}^{VEP}$ is a gap function for (VEP).*

5 Conclusions

In this chapter we have proposed some new gap functions by using conjugate duality theory for vector optimization (see [14]) and the perturbation approach for conjugate duality in scalar and vector optimization (cf. [2, 15]). In order to prove the properties of a gap function, recent results related to variational principles for vector equilibrium problems (see [1]) have been used. Moreover, some stability criteria due to special perturbation functions are given.

Notice that the presented approach can be extended to set-valued problems. Moreover, one can investigate more weaker assumptions for stability criteria in the future.

Acknowledgments The research of the first author has been supported partially by Deutsche Forschungsgemeinschaft. The authors are grateful to Dr. Radu Ioan Boţ for valuable discussions.

References

1. Altangerel, L., Boţ, R.I., Wanka, G.:On gap functions for equilibrium problems via Fenchel duality. Paci. J. Optim. 2(3), 667–678 (2006)
2. Altangerel, L., Boţ, R.I., Wanka, G.:Conjugate duality in vector optimization and some applications to the vector variational inequality. J. Math. Anal. Appl. 329(2), 1010–1035 (2007a)
3. Altangerel, L., Boţ, R.I., Wanka, G. Variational principles for vector equilibrium problems related to conjugate duality. J. Nonlinear. Convex Anal. 8(2) 179–196 (2007b)
4. Ansari, Q.H., Konnov, I.V., Yao, J.C. Existence of a solution and variational principles for vector equilibrium problems. J. Optim. Theory Appl. 110(3), 481–492 (2001)
5. Ansari, Q.H., Konnov, I.V., Yao, J.C. Characterizations of solutions for vector equilibrium problems. J. Optimi. Theory Appl. 113(3), 435–447 (2002)
6. Blum, E., Oettli, W. Variational principles for equilibrium problems. In: J. Guddat, (Ed.) et al., Parametric Optimization and Related Topics. III. Proceedings of the 3rd conference held in Güstrow, Germany, Frankfurt am Main: Peter Lang Verlag. Approximation Optimization. 3, 79–88 (1993)
7. Chen, G.Y., Goh, C.J., Yang, X.Q. On gap functions for vector variational inequalities. In: F. Giannessi Vector Variational Inequalities and Vector Equilibria, Mathematical Theories (pp. 55-72), (Ed.), Kluwer, Dordrecht (2000)
8. Goh, C.J., Yang, X.Q. Duality in Optimization and Variational Inequalities, Taylor and Francis, London (2002)
9. Sawaragi, Y., Nakayama, H., Tanino, T. Theory of Multiobjective Optimization, Mathematics in Science and Engineering, (Vol. 176), Academic, Orlando etc. (1985)
10. Song, W. Conjugate duality in set-valued vector optimization. J. Math. Anal. Appl. 216(1), 265–283 (1997)
11. Song, W. A generalization of Fenchel duality in set-valued vector optimization. Mathe Methods Oper. Res. 48(2) 259–272 (1998)
12. Tanino, T., Sawaragi, Y. Conjugate maps and duality in multiobjective optimization. J. Optim. Theory Appl. 31, 473–499 (1980)
13. Tanino, T. On supremum of a set in a multidimensional space. J. Math. Anal. Appl. 130(2), 386–397 (1988)
14. Tanino, T. Conjugate duality in vector optimization. J. Math. Anal. Appl. 167(1), 84–97 (1992)
15. Wanka, G., Boţ, R.I. On the relations between different dual problems in convex mathematical programming. In: P. Chamoni, R. Leisten, A. Martin, J. Minnemann, H. Stadler (Eds.), Operations Research Proceedings 2001 (pp. 255–262), Springer, Berlin, (2002)

Polynomially Solvable Cases of Binary Quadratic Programs

Duan Li[1], Xiaoling Sun[2], Shenshen Gu[3], Jianjun Gao[4], and Chunli Liu[5]

[1] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
dli@se.cuhk.edu.hk
[2] Department of Management Science, School of Management, Fudan University, Shanghai 200433, P. R. China
xls@fudan.edu.cn
[3] Department of Automation, Shanghai University, Shanghai 200072, China
gushenshen@shu.edu.cn
[4] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong
jjgaog@se.cuhk.edu.hk
[5] Department of Applied Mathematics, Shanghai University of Finance and Economics, Shanghai 200433, P. R. China
liu.chunli@shufe.edu.cn

Summary. We summarize in this chapter polynomially solvable subclasses of binary quadratic programming problems studied in the literature and report some new polynomially solvable subclasses revealed in our recent research. It is well known that the binary quadratic programming program is NP-hard in general. Identifying polynomially solvable subclasses of binary quadratic programming problems not only offers theoretical insight into the complicated nature of the problem but also provides platforms to design relaxation schemes for exact solution methods. We discuss and analyze in this chapter six polynomially solvable subclasses of binary quadratic programs, including problems with special structures in the matrix Q of the quadratic objective function, problems defined by a special graph or a logic circuit, and problems characterized by zero duality gap of the SDP relaxation. Examples and geometric illustrations are presented to provide algorithmic and intuitive insights into the problems.

Key words: binary quadratic programming, polynomial solvability, series-parallel graph, logic circuit, lagrangian dual, SDP relaxation

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_11,

1 Introduction

We consider in this chapter the following unconstrained 0–1 quadratic programming or binary quadratic programming problem:

$$(0\text{–}1QP) \quad \min_{x\in\{0,1\}^n} x^T Qx + c^T x,$$

where $Q = (q_{ij})_{n\times n}$ is symmetric and $c \in \mathbb{R}^n$. Termed also as the *pseudo-Boolean programming*, problem $(0\text{–}1QP)$ is a classical combinatorial optimization problem and is well known to be NP-hard (see [15]).

There exist many real-world applications of 0–1 quadratic programming, including financial analysis [24], molecular conformation problem [27], and cellular radio channel assignment [10]. Many combinatorial optimization problems, such as the max-cut problem (see e.g., [12, 16]), are special cases of the 0–1 quadratic programming problems. Various exact solution methods of a branch-and-bound framework for solving $(0\text{–}1QP)$ and its variants have been proposed in the literature (see, e.g., [4, 7, 10, 21–23, 26, 29] and references therein).

We focus in this chapter on the polynomially solvable cases of the quadratic binary programming problems. Identifying polynomially solvable subclasses of binary quadratic programming problems not only offers theoretical insight into the complicated nature of the problem but also provides useful information for designing efficient algorithms for finding optimal solution to $(0\text{–}1QP)$. More specifically, the properties of the polynomially solvable subclasses of $(0\text{–}1QP)$ provide hints and facilitate the derivation of efficient relaxations for the general form of $(0\text{–}1QP)$. Polynomially solvable binary quadratic programs even play an important role in devising exact methods for linearly constrained quadratic 0–1 programming. For example, the Lagrangian relaxation of the quadratic 0–1 knapsack problem, which is a special case of $(0\text{–}1QP)$, turns out to be polynomially solvable and thus makes it possible to efficiently compute the Lagrangian bounds in a branch-and-bound method for the quadratic 0–1 knapsack problem.

It is sometimes more convenient to consider some equivalent forms of $(0\text{–}1QP)$. Since $x_i^2 = x_i$ for $x_i \in \{0,1\}$, $(0\text{–}1QP)$ can be reduced to the following homogenous form $(0\text{–}1QP_{\rm h})$ without the linear term, using the substitution $Q := Q + \mathrm{diag}(c)$, where $\mathrm{diag}(c)$ is the diagonal matrix formed by vector c,

$$(0\text{–}1QP_{\rm h}) \quad \min_{x\in\{0,1\}^n} x^T Qx.$$

In many binary quadratic programming models arising from combinatorial optimization, the decision variables take values -1 or 1. The resulting binary quadratic programs take the following form:

$$(BQP) \quad \min_{x\in\{-1,1\}^n} x^T Qx + c^T x.$$

It can be seen that (0–1QP) with 0–1 variables (in x-space) can be reduced to a form of (BQP) with $(-1, 1)$ variables (in y-space) using transformation $x_i = \frac{1}{2}(y_i + 1)$.

As $x_i^2 = 1$, for both $x_i = 1$ and -1, we can assume, without loss of generality, that all diagonal elements of Q in (BQP) are zero. Thus, we can write the objective function in (BQP) as

$$\sum_{1 \le i < j \le n} 2q_{ij}x_ix_j + \sum_{i=0}^{n} c_ix_i.$$

By introducing an artificial variable $x_0 = 1$, we further have

$$f(x) = \sum_{0 \le i < j \le n} 2q_{ij}x_ix_j,$$

where $q_{0i} = \frac{1}{2}c_i$, $i = 1, \ldots, n$. Since for any $x \in \{-1, 1\}^{n+1}$, $f(x) = f(-x)$, we can relax the domain of x_0 to $\{-1, 1\}$ and (BQP) now takes the following equivalent homogenous form:

$$(BQP_{\rm h}) \quad \min_{x \in \{-1,1\}^{n+1}} x^T Q x,$$

where $Q := \begin{pmatrix} 0 & \frac{1}{2}c^T \\ \frac{1}{2}c & Q \end{pmatrix}$.

The well-known max-cut problem, which has being attracting remarkable attentions in recent years in combinatorial optimization, can be expressed in the form of ($BQP_{\rm h}$). Consider a graph $G = (E, V)$ with vertex set $V = \{1, \ldots, n\}$ and edge set $E = \{ij \mid 1 \le i < j \le n\}$. For every edge $ij \in E$, there is an associated weight w_{ij}. For a given set $S \subseteq V$, a *cut* $\delta(S)$ is the set of all edges with one end point in S and the other in $V \setminus S$, and the weight of cut $\delta(S)$ is then given by $\sum_{ij \in \delta(S)} w_{ij}$. The max-cut problem is to find a cut $\delta(S)$ with the maximum weight. Note that each $x \in \{-1, 1\}^n$ corresponds to a partition that divides V into $S = \{i \in V \mid x_i = 1\}$ and $V \setminus S = \{i \in V \mid x_i = -1\}$. We can now express the max-cut problem as the following binary quadratic problem,

$$\begin{aligned} \text{(Max-cut)} \quad & \max \frac{1}{2} \sum_{1 \le i < j \le n} w_{ij}(1 - x_ix_j) \\ & \text{s.t. } x \in \{-1, 1\}^n. \end{aligned}$$

While all the weights in the conventional definition for the max-cut problem considered in graph theory are assumed to be nonnegative, we consider here a more general setting of the max-cut problem without confining the weights to be nonnegative.

This chapter aims to give a systematic survey of the polynomially solvable subclasses of (0–1QP) and its variants studied in the literature and to report

some recent progress in this subject. Our goal is to present a self-contained writing and to provide step-by-step examples and geometric illustrations in an effort to capture the essence of the polynomial solvability of certain subclasses of binary quadratic programming problems. In Section 2, we discuss the problem (0–1QP) with all off-diagonal elements of Q being non-positive. This subclass of problems has been known for long time to be polynomially solvable due to the total unimodularity of the constraint matrix in its linear integer programming reformulation. Its relation to the maximum flow problem is also discussed. In Section 3, we analyze the polynomial solvability of problem (0–1$QP_{\rm h}$) with a fixed rank Q using the properties of zonotope in discrete geometry. The relationship between zonotope and hyperplane arrangement is exploited to derive an efficient procedure to enumerate all extreme points of a zonotope. In Section 4, we show that the problem (0–1QP) with a tridiagonal matrix Q can be solved by the basic algorithm in polynomial time. Sections 5 and 6 devote to problems defined by a special graph or a logic circuit. Relations between the polynomial solvability and the special properties of the series-parallel graph and logic circuit are studied. We investigate in Section 7 a possible zero duality gap between problem (BQP) and its SDP relaxation. A sufficient condition for the polynomial solvability of (BQP) via the SDP relaxation is presented. We conclude this chapter in Section 8 with a brief summary.

2 Problem (0–1QP) with All Off-Diagonal Elements of Q Being Non-positive

Consider a subclass of problem (0–1QP) where all off-diagonal elements of Q are non-positive. It is easy to see that $x_i x_j = \min(x_i, x_j)$ when $x_i,\ x_j \in \{0, 1\}$. Since $x_i^2 = x_i$, we can assume, without of loss of generality, $q_{ii} = 0$, $i = 1, \ldots, n$. Let $z_{ij} = x_i x_j$. If $q_{ij} \le 0$ for $1 \le i < j \le n$, then (0–1QP) is equivalent to the following linear integer programming problem:

$$\min \quad \sum_{i=1}^{n} c_i x_i + 2 \sum_{1 \le i < j \le n} q_{ij} z_{ij} \tag{1}$$

$$\text{s.t.} \quad z_{ij} \le x_i,\ 1 \le i < j \le n, \tag{2}$$

$$z_{ij} \le x_j,\ 1 \le i < j \le n, \tag{3}$$

$$x_i,\ x_j,\ z_{ij} \in \{0, 1\},\ 1 \le i < j \le n. \tag{4}$$

Consider the linear programming relaxation of the above problem by replacing constraint (4) with

$$x_i,\ x_j,\ z_{ij} \in [0, 1],\ 1 \le i < j \le n. \tag{5}$$

Recall that a matrix $A = (a_{ij})$ is called *totally unimodular* (TU) if every square sub-matrix of A has determinant $+1$, -1, or 0. It is well known that a linear programming problem with a totally unimodular constraint matrix and an integral right-hand side has an integral optimal solution. Recall also that a matrix A is TU if (i) $a_{ij} \in \{+1, -1, 0\}$ for all i, j; (ii) each column contains at most two non-zero coefficients ($\sum_{i=1}^{m} |a_{ij}| \leq 2$); and (iii) there exists a partition (M_1, M_2) of the set M consisting of the rows of A such that each column j contains two non-zero coefficients satisfying $\sum_{i \in M_1} a_{ij} - \sum_{i \in M_2} a_{ij} = 0$.

Note that the constraint matrix in the linear programming relaxation problem (1), (2), (3), and (5) is of the form $\begin{pmatrix} C \\ I \end{pmatrix}$ where C comes from these inequalities of $z_{ij} \leq x_i$ and $z_{ij} \leq x_j$. It suffices to show C is TU as a matrix A is TU iff $(A^T, I)^T$ is TU. Recall that a matrix A is TU iff A^T is TU. Note that there is one 1 and one -1 in each row of C and the third sufficient condition mentioned above can be satisfied by selecting $M_1 = C$ and $M_2 = \emptyset$.

In conclusion, (0–1QP) with all off-diagonal elements of Q being non-positive can be reduced to a linear programming problem and thus can be solved in polynomial time [18, 30].

The polynomial solvability of this subclass of (0–1QP) can be also shown by associating the problem with a graph and reducing the problem to a maximum flow problem. Consider a directed graph $G = (V, E)$ with $V = (s, 1, 2, \ldots, n, t)$, where s denotes the source and t the sink, and with $E = E_{\mathrm{s}} \cup E_{\mathrm{Q}} \cup E_{\mathrm{t}}$, where

$$
\begin{aligned}
E_{\mathrm{s}} &= \{sj \mid j = 1, \ldots, n\}, \\
E_{\mathrm{Q}} &= \{ij \mid q_{ij} < 0,\ 1 \leq i < j \leq n\}, \\
E_{\mathrm{t}} &= \{jt \mid j = 1, \ldots, n\}.
\end{aligned}
$$

The capacities of the arcs in E are defined as follows:

$$
e_{sj} = \max \left\{ 0, -2 \sum_{i=j+1}^{n} q_{ji} - c_j \right\},\ sj \in E_s, \tag{6}
$$

$$
e_{ij} = -2q_{ij},\ ij \in E_{\mathrm{Q}}, \tag{7}
$$

$$
e_{jt} = \max \left(0, 2 \sum_{i=j+1}^{n} q_{ji} + c_j \right),\ jt \in E_{\mathrm{t}}. \tag{8}
$$

Let $(U, \overline{U})$ be a partition of G with $s \in U$ and $t \in \overline{U}$. The set of arcs $\delta^+(U) = \{ij \mid i \in U,\ j \in \overline{U}\}$ is called an $s - t$ cut. The capacity of $\delta^+(U)$ is $\sum_{ij \in \delta^+(U)} e_{ij}$. The *minimum-cut* problem is to find a cut with the minimum capacity. Let Ψ be the capacity of the minimum-cut of G. Then

$\Psi = \min_U \sum_{ij\in\delta^+(U)} e_{ij}$. Associate each cut $\delta^+(U)$ of G with a 0–1 vector $(1, x_1, \ldots, x_n, 0)$ satisfying $x_i = 1$ if $i \in U$ and $x_i = 0$ otherwise. Similar to the proof for Property 6 in [8], we prove the following result which is also stated in [28].

Theorem 1. *Problem* (0–1QP) *with all off-diagonal elements of* Q *being non-positive can be reduced to the minimum-cut problem of the graph* $G = (V, E)$ *via the following relation:*

$$\min_{x\in\{0,1\}^n} \{x^T Q x + c^T x\} = \Psi - \sum_{j=1}^{n} e_{sj}.$$

Proof. By (6), (7) and (8), we have

$$\begin{aligned}
\Psi &= \min_{x\in\{0,1\}^n} \left\{ \sum_{j=1}^{n} e_{sj}(1-x_j) + \sum_{1\le i<j\le n} e_{ij}x_i(1-x_j) + \sum_{j=1}^{n} e_{jt}x_j \right\} \\
&= \sum_{j=1}^{n} e_{sj} + \min_{x\in\{0,1\}^n} \left\{ \sum_{j=1}^{n} \min(0, 2\sum_{i=j+1}^{n} q_{ji} + c_j)x_j - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij}x_i \right. \\
&\quad \left. +2 \sum_{1\le i<j\le n} q_{ij}x_i x_j + \sum_{j=1}^{n} \max(0, 2\sum_{i=j+1}^{n} q_{ji} + c_j)x_j \right\} \\
&= \sum_{j=1}^{n} e_{sj} + \min_{x\in\{0,1\}^n} \left\{ \sum_{j=1}^{n} \left(2\sum_{i=j+1}^{n} q_{ji} + c_j \right) x_j \right. \\
&\quad \left. -2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij}x_i + 2 \sum_{1\le i<j\le n} q_{ij}x_i x_j \right\} \\
&= \sum_{j=1}^{n} e_{sj} + \min_{x\in\{0,1\}^n} \left\{ \sum_{j=1}^{n} c_j x_j + 2 \sum_{1\le i<j\le n} q_{ij}x_i x_j \right\} \\
&= \sum_{j=1}^{n} e_{sj} + \min_{x\in\{0,1\}^n} \left\{ x^T Q x + c^T x \right\}.
\end{aligned}$$

This proves the theorem. □

It is well known that the minimum-cut problem is equivalent to the maximum-flow problem that can be solved in polynomial time (see [25]). Therefore, problem (0–1QP) with all off-diagonal elements of Q being non-positive can be solved by computing the maximum flow of a graph with $n+2$ vertices and $2n + n(n-1)/2$ arcs. Algorithms with different complexity bounds have been proposed for finding a maximum flow in G (see [14, 17, 25]), for example, an $O(n^3)$ maximum-flow algorithm proposed in [14] or [17].

3 Problem (0–1$QP_{\rm h}$) with Fixed Rank Q

We consider now a subclass of problem (0–1$QP_{\rm h}$) where Q is negative semidefinite and $rank(Q) = d$. Let $G = -Q$. In this situation, there exists a row full rank $d \times n$ matrix, V, such that $G = V^T V$, where the rows of V are suitably scaled eigenvectors of G. Problem (0–1$QP_{\rm h}$) can be thus expressed as

$$(BQP_{\rm fr}) \quad \max_{x\in\{0,1\}^n} x^T G x = x^T V^T V x = \sum_{i=1}^{d} (v_i x)^2,$$

where v_i is the ith row of matrix V.

If d is equal to 1, i.e., the matrix G is of rank one with $G = v_1^T v_1$, the solution to $(BQP_{\rm fr})$ can be easily found by inspection. More specifically, we only need to select x such that the absolute value of $v_1 x$ is maximized on $\{0,1\}^n$.

In general cases with $rank(G) = d > 1$, we consider a linear map Φ: $x \in \mathbb{R}^n \to z = Vx \in \mathbb{R}^d$, in which Φ maps the hypercube $[0,1]^n$ into a convex polytope $Z(V) = \Phi([0,1]^n) = \{z \in \mathbb{R}^d \mid z = Vx,\ x \in [0,1]^n\}$, known as a *zonotope*. Note that

$$\max_{x\in\{0,1\}^n} x^T G x = \max_{x\in\{0,1\}^n} \sum_{i=1}^{d} (v_i x)^2 = \max_{z\in Z(V)} \sum_{i=1}^{d} z_i^2 = \max_{z\in Z(V)} \|z\|^2,$$

where the second equality is due to that the maximization of a convex function over a convex set is always achieved at the vertices. Based on the same argument, the convex function $\|z\|^2$ achieves its maximum over the convex set $Z(V)$ at some extreme point $\tilde{z}$. Thus, $(BQP_{\rm fr})$ reduces to a problem of finding the maximum norm in a zonotope.

Theorem 2. *For any extreme point $\tilde{z}$ of the zonotope $Z(V)$, there is a point $\tilde{x} \in \{0,1\}^n$ such that $\tilde{z} = V\tilde{x}$.*

Proof. Since V is row full rank, we can assume that $V = (\hat{V}, V_1)$, where $\hat{V}$ is a $d \times d$ nonsingular matrix. Let $x = \begin{pmatrix} \hat{x} \\ \bar{x} \end{pmatrix}$, where $\hat{x}$ is a d-dimensional vector corresponding to the columns of $\hat{V}$. Letting $\bar{x} = 0$ in the equation $\tilde{z} = Vx$, we obtain $\tilde{z} = \hat{V}\hat{x}$. Then $\tilde{x} = \begin{pmatrix} \hat{V}^{-1}\tilde{z} \\ 0 \end{pmatrix}$ satisfies $\tilde{z} = V\tilde{x}$ and is an extreme point of $[0,1]^n$. Indeed, suppose that there exist $\tilde{x}_1$ and $\tilde{x}_2$ with $\tilde{x}_1 \neq \tilde{x}_2$ such that $\tilde{x} = \lambda\tilde{x}_1 + (1-\lambda)\tilde{x}_2$ for some $\lambda \in (0,1)$. Then $\tilde{x}_1 = \begin{pmatrix} \hat{x}_1 \\ 0 \end{pmatrix}$ and $\tilde{x}_2 = \begin{pmatrix} \hat{x}_2 \\ 0 \end{pmatrix}$ for some $\hat{x}_1,\ \hat{x}_2 \in [0,1]^d$ with $\hat{x}_1 \neq \hat{x}_2$. Thus, $\tilde{z} = \lambda\hat{V}\hat{x}_1 + (1-\lambda)\hat{V}\hat{x}_2$. Since $\hat{V}$ is nonsingular and $\hat{x}_1 \neq \hat{x}_2$, we deduce that $\hat{V}\hat{x}_1,\ \hat{V}\hat{x}_2 \in Z(V)$ and $\hat{V}\hat{x}_1 \neq \hat{V}\hat{x}_2$, which in turns implies that $\tilde{z}$ is not an extreme point of $Z(V)$, a contradiction.

The following is a classical result in discrete geometry (see, e.g., [34]) which gives a polynomial upper bound of the number of extreme points of $Z(V)$ for fixed d.

Theorem 3. *Let $N_{\rm ep}(Z)$ denote the number of extreme points of the zonotope $Z(V)$. Then $N_{\rm ep}(Z) = O(n^{d-1})$.*

An immediate implication of Theorems 2 and 3 is that problem (0–1$QP_{\rm h}$) with fixed rank Q is polynomially solvable.

We now discuss how to enumerate all the extreme points of the zonotope $Z(V)$. Let v^j denote the jth column vector of V. Assume that the regularity condition is satisfied for the zonotope $Z(V)$, i.e., each column of V is non-zero and $v^i \neq kv^j$ for any $i \neq j$ and $k \neq 0$. Associated with $Z(V)$, we define a set of hyperplanes in $\mathbb{R}^d$ with v^j $(j = 1, \ldots, n)$ being normal vectors:

$$\mathcal{A}(V) = \{h_j \mid j = 1, \ldots, n\},$$

where $h_j = \{y \in \mathbb{R}^d \mid (v^j)^T y = 0\}$ for $j = 1, \ldots, n$. The set $\mathcal{A}(V)$ is called *central arrangement* of V. Denote $h_j^+ = \{y \in \mathbb{R}^d \mid (v^j)^T y > 0\}$ and $h_j^- = \{y \in \mathbb{R}^d \mid (v^j)^T y < 0\}$. For any $c \in \mathbb{R}^d$, define the location vector $\gamma(c) \in \{+, 0, -\}^n$ by

$$\gamma(c)_j = \begin{cases} +, \text{ if } c \in h_j^+, \\ 0, \text{ if } c \in h_j, \\ -, \text{ if } c \in h_j^-. \end{cases}$$

Let $c \in \mathbb{R}^d$ be such that $\gamma(c)_j \neq 0$ for $j = 1, \ldots, n$. A *cell* of the arrangement $\mathcal{A}(V)$ is defined as the following d-dimensional subset:

$$C_c = \{y \in \mathbb{R}^d \mid \gamma(y) = \gamma(c)\}. \tag{9}$$

Obviously, C_c is invariant for any $y \in C_c$. Thus, a cell can be represented by its sign vector. Denote by $C(V)$ the set of all cells of the arrangement $\mathcal{A}(V)$, i.e.,

$$C(V) = \{C_c \mid c \in \mathbb{R}^d\}.$$

For any cell $C_c \in C(V)$, denote $\gamma^+(c) = \{j \mid \gamma(c)_j = +\}$ and $\gamma^-(c) = \{j \mid \gamma(c)_j = -\}$.

Theorem 4. *There is a one-to-one correspondence between the extreme points of $Z(V)$ and the cells of $\mathcal{A}(V)$.*

Proof. For each cell $C_c \in C(V)$, define x_c by

$$(x_c)_j = \begin{cases} 1, \text{ if } j \in \gamma^+(c) \\ 0, \text{ if } j \in \gamma^-(c). \end{cases} \tag{10}$$

Let $z_c = Vx_c$, then $c^T z_c = \sum_{j \in \gamma^+(c)} c^T v^j$. Since $c^T v^j > 0$ for $j \in \gamma^+(c)$ and $c^T v^j < 0$ for $j \in \gamma^-(c)$, z_c is the unique optimal solution to the linear program $\max_{z \in Z(V)} c^T z$. Thus z_c is an extreme point of the polytope $Z(V)$.

Conversely, for any extreme point $\tilde{z}$ of $Z(V)$, there is a $c \in \mathbb{R}^d$ such that $\tilde{z}$ is the unique optimal solution to the linear program $\max_{z \in Z(V)} c^T z$. Notice that

$$\max_{z \in Z(V)} c^T z = \max_{x \in [0,1]^n} \sum_{j=1}^{n} x_j (c^T v^j).$$

So $\tilde{z}$ must be of the form Vx_c with x_c being defined in (10). There must be no j such that $c^T v^j = 0$, i.e., $\gamma(c)_j \neq 0$ for any j, since otherwise the optimal solution to the linear program $\max_{z \in Z(V)} c^T z$ is not unique. The cell C_c defined in (9) is then the cell in $C(V)$ corresponding to $\tilde{z}$. The one-to-one property of the above correspondence can be easily established by noting that V is row full rank.

Theorem 4 implies that enumeration of all the extreme points of the zonotope $Z(V)$ is equivalent to the enumeration of all the cells of the arrangement $\mathcal{A}(V)$ for which various procedures have been proposed (see [1, 2, 13, 32]).

Note that the central arrangement $\mathcal{A}(V)$ satisfies $\cap_{j=1}^{n} h_j = \{0\}$ and the cells of $\mathcal{A}(V)$ are symmetric to the origin. We thus only need to generate half of the cells or the corresponding sign vectors. Consider a shift of the last hyperplane $h = \{x \in \mathbb{R}^d \mid (v^n)^T y = b\}$, where $b \neq 0$. The intersection of $\mathcal{A}(V)$ and h is a general arrangement of $n-1$ hyperplanes in $\mathbb{R}^{d-1}$. It can be seen that the sign vectors (cells) of $\mathcal{A}'(V) = \mathcal{A}(V) \cap h$ corresponds to the half of the sign vectors of $\mathcal{A}(V)$ with the last element being $+$ or $-$.

Now, consider a general arrangement $\mathcal{A} = \{h_j \mid j = 1, \ldots, m\}$, where $h_j = \{y \in \mathbb{R}^d \mid a_j^T y = b_j,\ j = 1, \ldots, m\}$. The sign vector of a cell in a general arrangement can be defined similarly as for the central arrangement. A *root* cell is the cell with all $+$ elements in the sign vector. A root cell can be found by selecting any cell and reversing the orientation of some of the hyperplanes if necessary. Two cells are called neighbors if only one of the hyperplanes separates them, i.e., the sign vectors differ only in exactly one element. A *parent* cell of c is a unique neighbor of c which contains one more $+$ in its sign vector. Any cell with c being its parent is called a *child* of c. If a unique parent of each cell (except for the root cell) is assigned, then a directed tree structure can be obtained for the cells and the reverse search algorithm can be used to traverse this tree backward, enumerating all the cells exactly once. A procedure to search all the adjacent cells of a cell c is needed in the reverse search algorithm. The procedure of cell enumeration can be described as follows.

Procedure 1 (Cell Enumeration)

Input: a cell c represented by its sign vector, and the hyperplanes represented by (A, b)

Output: a set $C(A)$ containing all the cells of the arrangement (rooted at c)

begin

(i) output c to $C(A)$.

(ii) call a subroutine to list all adjacent cells of c
(iii) **for** each cell e of c **do**
if c is the unique parent of e **then**
recurse the procedure with e as the input cell
endif
endfor
end

The above recursive procedure starts from the root cell and terminates when all the cells are enumerated. The details of the procedures for finding all neighbors of a cell and searching for the unique parent of a cell can be found in [2, 32]. To illustrate the cell enumeration procedure, letus consider an instance of (0–1QP_{h}) where $Q = -V^T V$ and

$$V = \begin{pmatrix} -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}.$$

Using the parallel translation of the last hyperplane of the arrangement $y_3 = 1$, the reduced general arrangement contains four hyperplanes in $\mathbb{R}^2$ and is represented by

$$A = \begin{pmatrix} -1 & -1 & 0 & 1 \\ -1 & 0 & 1 & -1 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Figure 1 illustrates the cell enumerating process of the arrangement (A, b), where each cell is represented by its sign vector and the number indicates the order of the cell enumeration process in Procedure 1.

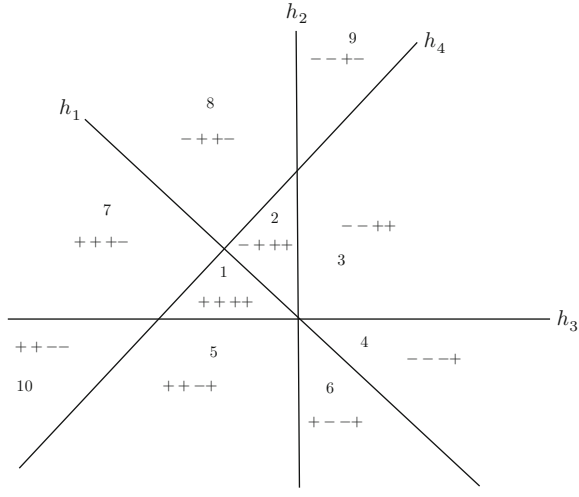


Fig. 1. Illustration for cell enumeration process

As there are 10 cells in the reduced general arrangement, there are 20 cells in the original central arrangement. Thus, the zonotope $Z(V)$ has 20 extreme points among which $z_c = Vx_c$, where $x_c = (1, 1, 0, 1, 0)^T$, is the optimal solution to $\max_{z \in Z(V)} \|z\|^2$. Therefore, $x_c = (1, 1, 0, 1, 0)^T$ is the optimal solution to the original problem (0–1$QP_{\rm h}$) with optimal value 6.

4 Problem (0–1QP) with Q Being a Tridiagonal Matrix

We first consider problem (0–1QP) in its general form in this section. Denote by $\Delta_i(x)$ the ith derivative of $f(x) = x^T Qx + c^T x$ at x,

$$\begin{aligned}\Delta_i(x) &= \frac{\partial f}{\partial x_i} \\ &= f(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n).\end{aligned}$$

Denote by $\Theta_i(x)$ the ith residual

$$\begin{aligned}\Theta_i(x) &= f(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \\ &= f(x) - x_i \Delta_i(x).\end{aligned}$$

Both $\Delta_i(x)$ and $\Theta_i(x)$ are, in general, *linear* functions of $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. Moreover, f can be expressed as

$$f(x) = x_i \Delta_i(x) + \Theta_i(x). \tag{11}$$

It is clear that a point $x \in \{0, 1\}^n$ is a solution to (0–1QP) only if for all $i = 1, \ldots, n$,

$$x_i = \begin{cases} 1, & \text{if } \Delta_i(x) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The basic algorithm [11, 19] is developed based on the above necessary optimality condition. We first express $f(x)$ in (0–1QP) as

$$f(x) = x_n \Delta_n(x_1, \ldots, x_{n-1}) + \Theta_n(x_1, \ldots, x_{n-1}). \tag{13}$$

From the optimal condition (12), the global minimizer of f satisfies

$$x_n = \begin{cases} 1, & \text{if } \Delta_n(x_1, \ldots, x_{n-1}) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

Therefore, if we can express x_n defined in (14) as a polynomial of $x_1, \ldots, x_{n-1}$, $\phi_n(x_1, \ldots, x_{n-1})$, then we can eliminate x_n from the expression of $f(x)$ in (13),

$$f_{n-1}(x_1, \ldots, x_{n-1}) = \phi_n(x_1, \ldots, x_{n-1}) \Delta_n(x_1, \ldots, x_{n-1}) + \Theta_n(x_1, \ldots, x_{n-1}).$$

Note that, in general cases, $f_{n-1}(x_1, \ldots, x_{n-1})$ may not be a quadratic function, as $\phi_n(x_1, \ldots, x_{n-1})$, in general, is not a linear function. Performing the

same elimination process for f_{n-1}, we will get a function f_{n-2} of $x_1, \ldots, x_{n-2}$ and this process continues recursively until we obtain $f_1(x_1)$. Let x^* denote the optimal solution of (0–1QP). Notice that $x_1^* = 1$ if $f_1(1) < f_1(0)$ and $x_1^* = 0$ otherwise. Then $x_2^*, \ldots, x_n^*$ can be obtained by using $x_{i+1}^* = \phi_{i+1}(x_1^*, \ldots, x_i^*)$ recursively for $i = 1, \ldots, n-1$.

The basic algorithm [11, 19] can then be described as follows.

Algorithm 4.1 (Basic Algorithm for (0–1QP)).

Step 0. Set $f_n(x) = f(x)$ and $k = n$.
Step 1. Calculate

$$\Delta_k(x_1, \ldots, x_{k-1}) = \frac{\partial f_k}{\partial x_k},$$
$$\Theta_k(x_1, \ldots, x_{k-1}) = f_k(x_1, \ldots, x_{k-1}, 0).$$

Determine the polynomial expression of ϕ_k defined by

$$\phi_k(x_1, \ldots, x_{k-1}) = \begin{cases} 1, & \text{if } \Delta_k(x_1, \ldots, x_{k-1}) < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Step 2. Compute

$$f_{k-1}(x_1, \ldots, x_{k-1}) = \phi_k(x_1, \ldots, x_{k-1})\Delta_k(x_1, \ldots, x_{k-1}) + \Theta_k(x_1, \ldots, x_{k-1}).$$

Step 3. If $k > 1$, then set $k := k - 1$ and go to Step 1. Otherwise, set $x_1^* = 1$ if $f_1(1) < f_1(0)$ and $x_1^* = 0$ if $f_1(1) \geq f_1(0)$. Calculate x_k^* by $x_k^* = \phi_k(x_1^*, \ldots, x_{k-1}^*)$ for $k = 2, \ldots, n$.

It is proved in [19] that the basic algorithm produces an optimal solution x^* to (0–1QP). The following small-size example illustrates the algorithm.

Example 1.

$$\max_{x \in \{0,1\}^3} f(x) = 4x_1x_2 - x_1x_3 + 2x_2x_3.$$

By the algorithm, we have $\Delta_3(x_1, x_2) = -x_1 + 2x_2$ and thus

$$\phi_3(x_1, x_2) = \left\{ \begin{array}{l} 1, \text{ if } \Delta_3(x_1, x_2) < 0 \\ 0, \text{ otherwise} \end{array} \right\} = x_1(1 - x_2).$$

Hence we get

$$\begin{aligned} f_2(x_1, x_2) &= \phi_3(x_1, x_2)\Delta_3(x_1, x_2) + \Theta_3(x_1, x_2) \\ &= x_1(1 - x_2)(-x_1 + 2x_2) + 4x_1x_2 \\ &= 5x_1x_2 - x_1. \end{aligned}$$

Since $\Delta_2(x_1) = 5x_1$, we get

$$\phi_2(x_1) = \left\{ \begin{array}{l} 1, \text{ if } g_2(x_1) < 0, \\ 0, \text{ otherwise} \end{array} \right\} = 0.$$

Thus,

$$f_1(x_1) = \phi_2(x_1)\Delta_2(x_1) + \Theta_2(x_1) = -x_1.$$

Therefore, $x_1^* = 1$, $x_2^* = \phi_2(x_1^*) = 0$, and $x_3^* = \phi_3(x_1^*, x_2^*) = 1$. The optimal solution to the example is $x^* = (1, 0, 1)^T$ with $f(x^*) = -1$.

The key task in performing the basic algorithm is how to identify the polynomial expression of ϕ_k defined in (15). Techniques to obtain the polynomial expression ϕ_k are discussed in [11, 20]. In principle, ϕ_k can be always constructed systematically. Let us consider the following instance $\Delta_4(x_1, x_2, x_3) = 4x_1 - x_2 - 5x_3$. The first step is to find the mapping from all possible combinations of x_1, x_2, and x_3 to the value of Δ_4 which is given in the following table.

Table 1. Illustrative example of mapping Δ_k

x_1	x_2	x_3	$\Delta_4(x_1, x_2, x_3)$
0	0	0	0
1	0	0	4
0	1	0	−1
0	0	1	−5
1	1	0	3
1	0	1	−1
0	1	1	−6
1	1	1	−2

Using Boolean algebra and noticing that all possible combinations of x_1, x_2 and x_3 are mutually exclusive, we can get

$$\begin{aligned} \phi_4(x_1, x_2, x_3) &= (1 - x_1)x_2(1 - x_3) + (1 - x_1)(1 - x_2)x_3 + x_1(1 - x_2)x_3 \\ &\quad + (1 - x_1)x_2x_3 + x_1x_2x_3 \\ &= x_2 - x_3 - x_1x_2 - x_2x_3 + x_1x_2x_3. \end{aligned}$$

Note that if Δ_k involves s variables, then we need to examine 2^s combinations. In the worst case, if Δ_n involves $n - 1$ variables, calculating ϕ_n is more than enumerating 2^{n-1} possible solutions. The basic algorithm could become very powerful for (0–1QP) when interactions among variables are weak, for example, when matrix Q in (0–1QP) is tridiagonal.

We consider now a special case of problem (0–1QP) where Q is a tridiagonal symmetric matrix with zero diagonal elements,

$$Q = \begin{pmatrix} 0 & q_{12} & 0 & \dots & 0 & 0 \\ q_{12} & 0 & q_{23} & \dots & 0 & 0 \\ 0 & q_{23} & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 & q_{n-1,n} \\ 0 & 0 & 0 & \dots & q_{n-1,n} & 0 \end{pmatrix}.$$

In this special case, it can be verified that both functions Δ_k and ϕ_k are linear functions of a single variable x_{k-1}. Thus, f_k remains a quadratic form all the way through the iteration. The basic algorithm becomes polynomial in such a special case.

Algorithm 4.2 (Exact Algorithm for (0–1QP) with Q Being Tridiagonal).

Step 0. Set $f_n(x) = f(x)$ and $k = n$.
Step 1. Calculate

$$\Delta_k(x_{k-1}) = \frac{\partial f_k}{\partial x_k} = 2q_{k-1,k}x_{k-1} + c_k,$$
$$\Theta_k(x_1, \dots, x_{k-1}) = f_k(x_1, \dots, x_{k-1}, 0).$$

Determine the polynomial expression of ϕ_k defined by

$$\phi_k(x_{k-1}) = \begin{cases} 1 & \text{if } 2q_{k-1,k} + c_k < 0 \text{ and } c_k < 0, \\ 0 & \text{if } 2q_{k-1,k} + c_k \geq 0 \text{ and } c_k \geq 0, \\ x_{k-1} & \text{if } 2q_{k-1,k} + c_k < 0 \text{ and } c_k \geq 0, \\ 1 - x_{k-1} & \text{if } 2q_{k-1,k} + c_k \geq 0 \text{ and } c_k < 0. \end{cases} \tag{16}$$

Step 2. Compute

$$f_{k-1}(x_1, \dots, x_{k-1}) = \phi_k(x_{k-1})\Delta_k(x_{k-1}) + \Theta_k(x_1, \dots, x_{k-1}),$$

and simplify the expression using $x_{k-1}^2 = x_{k-1}$.
Step 3. If $k > 1$, then set $k := k - 1$ and go to Step 1. Otherwise, set $x_1^* = 1$ if $f_1(1) < f_1(0)$ and $x_1^* = 0$ if $f_1(1) \geq f_1(0)$. Calculate x_k^* by $x_k^* = \phi_k(x_{k-1}^*)$ for $k = 2, \dots, n$.

5 Problem (BQP) Defined by a Series-Parallel Graph

We consider graph $G = (E, V)$. Given a subset of vertex $T \subset V$, we use $G[T]$ to denote an induced subgraph of G, where it consists of T and all edges whose end points are contained in T. For any node $v \in V$, the degree of v is the cardinality of cut $\delta(\{v\})$, denoted as $deg(v)$.

Given two edge sets $E_1 \subset E$ and $E_2 \subset E$ in graph G such that $E_1 \cap E_2 = \emptyset$, we use $\beta(E_1, E_2, G)$ to denote the weight of a cut $\delta(U)$ such that $E_1 \subset \delta(U)$ and $E_2 \cap \delta(U) = \emptyset$ and the weight of such a cut, $w(\delta(U))$, is maximized in

G. Therefore, $\beta(E_1, E_2, G)$ can be interpreted as a *constrained* max-cut that must include all edges in E_1 and does not include any edge in E_2. Furthermore, $\beta(\emptyset, \emptyset, G)$, for short $\beta(G)$, actually is the weight of the max-cut of graph G. Note $w(\delta(\emptyset)) = 0$.

We use K_n to denote the complete graph with n vertices, where all n vertices are pairwise adjacent. A graph G is a contractible to G', if G' can be obtained from G by a sequence of elementary contractions, in which edge ij is replaced by a single vertex whose incident edges are the edges other than ij that were incident to i or j. The multiple edges arising from the contraction are merged into a single edge in such a procedure. A graph is called *series-parallel* if it is not contractible to K_4. Graph (A) in Fig. 2 is series-parallel and (B) is not.

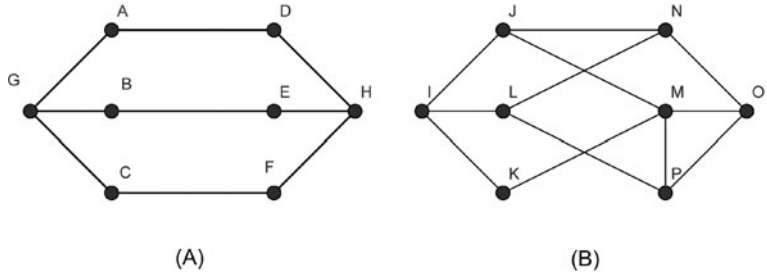


Fig. 2. Examples of series-parallel and non-series-parallel graphs

We consider problem (BQP) and reduce it first to a max-cut problem. Define a graph $G(Q) := \{V, E\}$ for problem (BQP), which is associated to $Q = \{q_{ij}\}_{n\times n}$, as follows:

$$\begin{aligned} &V = \{1, 2, \cdots, n\}, \\ &ij \in E \Leftrightarrow q_{ij} \neq 0, \\ &w_{ij} = 2q_{ij}, \end{aligned}$$

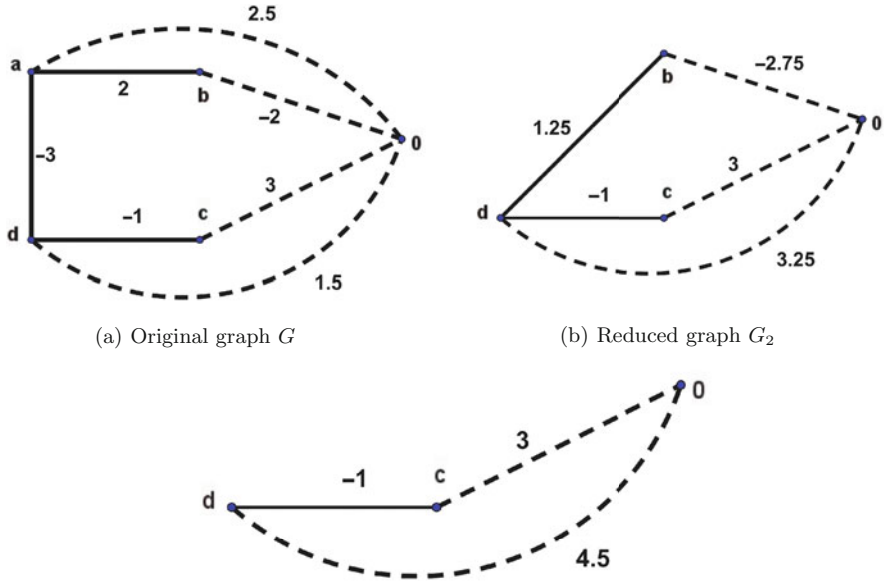
where w_{ij} is the weight assigned to edge ij. We then construct a new graph $G(Q, c)$ by adding a universal vertex $\{0\}$ which is connected to each vertex of $G(Q)$ and assign weight $w_{0j} = c_j$ to edge $0j$, for $j = 1, \ldots, n$. Clearly, $G(Q) = G(Q, c) \setminus \{0\}$. Then, solving (BQP) is equivalent to finding the max-cut of graph $G(Q, c)$:

$$\begin{aligned} \max \quad & \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \{w_{ij} | y_i = -y_j\} \\ \text{s.t.} \quad & y_i^2 = 1, \text{ for } i = 0, \cdots, n. \end{aligned}$$

Consider an instance of (BQP) with

$$Q = \begin{pmatrix} 0 & 1 & 0 & -1.5 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.5 \\ -1.5 & 0 & -0.5 & 0 \end{pmatrix}, \; c = \begin{pmatrix} 2.5 \\ -2 \\ 3 \\ 1.5 \end{pmatrix}.$$

The correspondent graph of this example problem is given in Fig. 3(a). It is easy to check that graph $G(Q)$ in this example is series-parallel.



(a) Original graph G

(b) Reduced graph G_2

(c) Reduced graph G_3

Fig. 3. The original graph and the reduced graphs of the example instance

If graph $G(Q)$ is series-parallel, then graph $G(Q, c)$ is not contractible to K_5. Recall the facts [3] that any subgraph of a series-parallel graph is still series-parallel and there always exists a vertex in a series-parallel graph that has degree not greater than 2. The main result in [3] is that if graph $G(Q)$ is series-parallel, the corresponding max-cut problem of graph $G(Q, c)$ can be solved by a linear-time algorithm which we are presenting below.

If graph $G(Q, q)$ is of three vertices or less, the max-cut problem can be solved by enumeration. Otherwise, for every vertex i in $G(Q)$, we compute its degree d_i and place all vertices with degree not greater than 2 into a list L, which can be achieved in linear time $O(n)$. In each iteration, we choose a vertex j from L and perform a reduction. We need to consider the following three different situations.

Case 1. If $deg(j) = 2$, let k and l be the vertices adjacent to j in $G(Q)$. We assume that $G(Q, q)$ contains all three edges $0k$, $0l$, and kl. Otherwise,

we can add the missing edge with weight 0. Let W be the subgraph of $G(Q,q)$ induced by $\{0,j,k,l\}$ with edge weights the same as in $G(G,q)$. See the left subgraph of Fig. 4 for graphical presentation of subgraph W. Note that any cut of W contains either two edges of $0k$, $0l$, and kl or none of them.

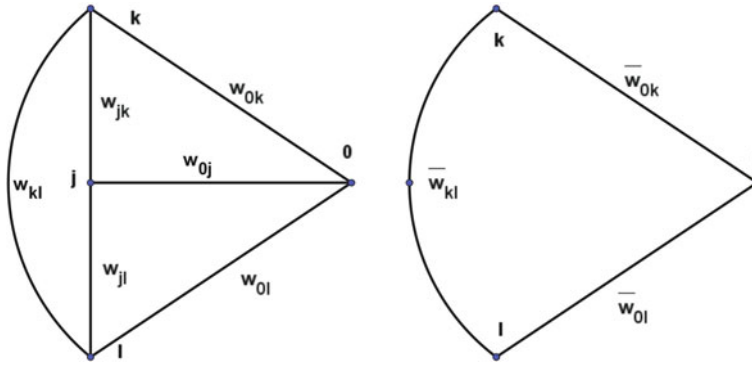


Fig. 4. The graph of $\{j,k,l,0\}$

The max-cut problem is solved by recursively generating $G' := G(Q,q) \setminus \{j\}$. All the edge weights in $G(Q,q)\setminus\{j\}$ are the same as in $G(Q,q)$, except for these edges in subgraph W, $0k$, kl, and $0l$, which need to be modified. For the reduced graph $W' = W \setminus \{j\}$ depicted in the right subgraph of Fig. 4, there are only three possible cuts, $\{0k, lk\}$, $\{0l, lk\}$, and $\{0k, 0l\}$. All of such cuts have to satisfy the following balance equations,

$$\begin{aligned}
\bar{w}_{0l} + \bar{w}_{0k} &= \beta(\{0l,0k\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W),\\
\bar{w}_{0k} + \bar{w}_{lk} &= \beta(\{0k,kl\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W),\\
\bar{w}_{0l} + \bar{w}_{lk} &= \beta(\{0l,kl\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W).
\end{aligned}$$

The meaning of the above equations is clear. For example, the weight of the cut $\{0k,kl\}$ in the reduced graph W' should be equal to that of the max-cut involving $\{0k,kl\}$ in the original graph W, while taking away the contribution of the edges leading to node j, $\beta(\emptyset,\{0k,kl,0l\},W)$. The solution to the above system of linear equations is

$$\begin{aligned}
\bar{w}_{0l} := 0.5[&\beta(\{0l,kl\},\emptyset,W) + \beta(\{0l,0k\},\emptyset,W)\\
&-\beta(\{0k,kl\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W)],\\
\bar{w}_{0k} := 0.5[&\beta(\{0k,kl\},\emptyset,W) + \beta(\{0l,0k\},\emptyset,W)\\
&-\beta(\{0l,kl\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W)],\\
\bar{w}_{lk} := 0.5[&\beta(\{0l,kl\},\emptyset,W) + \beta(\{0k,kl\},\emptyset,W)\\
&-\beta(\{0k,0l\},\emptyset,W) - \beta(\emptyset,\{0k,kl,0l\},W)].
\end{aligned}$$

It is evident that $\beta(G(Q,q)) = \beta(G') + \beta(\emptyset, \{0k, kl, 0l\}, W)$. The optimal cut in G' is extended to an optimal cut in $G(Q,q)$ by taking the appropriate cut in W. Set then $\deg(l) = \deg(l) - 1$ and $\deg(k) = \deg(k) - 1$. If $\deg(l) \leq 2$ or $\deg(k) \leq 2$, add l or k to L.

Case 2. If $\deg(j) = 1$, let k be the vertex adjacent to j in $G(Q)$. Let W be the subgraph of $G(Q,q)$ induced by $\{0, j, k\}$, in which the weights are the same as in $G(Q,q)$. In $G' := G(Q,q) \setminus \{j\}$, we only need to modify the weight of edge $0k$ to

$$\bar{w}_{0k} := \beta(\{0k\}, \emptyset, W) - \beta(\emptyset, \{0k\}, W).$$

It is clear $\beta(G(Q,q)) = \beta(G') + \beta(\emptyset, \{0k\}, W)$. Set $\deg(k) = \deg(k) - 1$. If $\deg(k) \leq 2$, we include k in L.

Case 3. If $\deg(j) = 0$, the problem can be solved in $G' := G(Q,q) \setminus \{j\}$ and in the subgraph induced by $\{j, 0\}$, separately.

In any of the above three cases, we reduce the nodes of the graph by one in each iteration. If the size of $G(Q,q)$ is n, the computational effort needed by this algorithm is bounded by $O(n)$.

We now illustrate the above solution scheme for the example given in Fig. 3(a).

Step 1 The initial list is given by $L := \{a, b, c, d\}$. As $\deg(a) = 2$, we consider a reduced graph $G_2 = G \setminus \{a\}$ given in Fig. 3(b). Let subgraph W_1 be induced by vertices $\{a, b, d, 0\}$. We calculate $\beta(\{0b, 0d\}, \emptyset, W_1)$ based on its definition. Consider two possible cuts in W_1 that include edges $0b$, $0d$ in subgraph H, $\{ab, ad, 0b, 0d\}$ and $\{0a, 0b, 0d\}$. Thus,

$$\beta(\{0b, 0d\}, \emptyset, W_1) = \max\{(2 - 3 - 2 + 1.5), (2.5 - 2 + 1.5)\} = 2.$$

Similarly, we can get $\beta(\{0b, bd\}, \emptyset, W_1) = 0$, $\beta(\{0d, bd\}, \emptyset, W_1) = 6$, and $\beta(\emptyset, \{0b, bd, 0d\}, W_1) = 1.5$. Furthermore, the modified weights for $0b$, bd, and $0d$ are given as

$$\begin{aligned}
\bar{w}_{0b} &= 0.5[\beta(\{0b, bd\}, \emptyset, W_1) + \beta(\{0b, 0d\}, \emptyset, W_1) - \beta(\{0d, bd\}, \emptyset, W_1) \\
&\quad -\beta(\emptyset, \{0b, bd, 0d\}, W_1)] = -2.75, \\
\bar{w}_{bd} &= 0.5[\beta(\{0b, bd\}, \emptyset, W_1) + \beta(\{0d, bd\}, \emptyset, W_1) - \beta(\{0b, 0d\}, \emptyset, W_1) \\
&\quad -\beta(\emptyset, \{0b, bd, 0d\}, W_1)] = 1.25, \\
\bar{w}_{0d} &= 0.5[\beta(\{0d, bd\}, \emptyset, W_1) + \beta(\{0d, 0b\}, \emptyset, W_1) - \beta(\{0b, bd\}, \emptyset, W_1) \\
&\quad -\beta(\emptyset, \{0b, bd, 0d\}, W_1)] = 3.25.
\end{aligned}$$

We also have

$$\beta(G) = \beta(G_1) + \beta(\emptyset, \{0b, bd, 0d\}, W_1).$$

After deleting a, the node list is updated to $L := \{b, d, c\}$.

Step 2 As $deg(b) = 1$ in graph G_2, we consider a reduced graph $G_3 = G_2 \backslash \{b\}$ given in Fig. 3(c). Let subgraph W_2 be induced by vertices $\{b, d, 0\}$. We have

$$\beta(\{0d\}, \emptyset, W_2) = 4.5, \ \beta(\emptyset, \{0d\}, W_2) = w(\delta(\emptyset)) = 0,$$
$$\bar{w}_{0d} = \beta(\{0d\}, \emptyset, W_2) - \beta(\emptyset, \{0d\}, W_2) = 4.5.$$

It is clear that $\beta(G_2) = \beta(G_3) + \beta(\emptyset, \{0d\}, W_2)$.

Step 3 There are only three vertices in G_3. Comparing all possible cuts yields $\beta(G_3) = 7.5$ with max-cut $\{0c, 0d\}$. Tracing back gives rise to

$$\beta(G_2) = \beta(G_3) + \beta(\emptyset, \{0d\}, W_2) = 7.5 + 0 = 7.5,$$
$$\beta(G) = \beta(G_2) + \beta(\emptyset, \{0b, bd, 0d\}, W_1) = 7.5 + 1.5 = 9.$$

The remaining problem is how to identify the optimal solution to the primal problem. The max-cut in G_3 gives rise an optimal division as $(\{c, d\}, \{0\})$. Comparing two possible "expanding" divisions of nodes in G_2, $(\{d, c\}, \{b, 0\})$ and $(\{d, c, b\}, \{0\})$ yields the optimal division in G_2, $(\{d, c\}, \{b, 0\})$. Finally, comparing two possible "expanding" divisions of nodes in G, $(\{a, d, c\}, \{b, 0\})$ and $(\{d, c\}, \{a, b, 0\})$ identifies the optimal division of the entire problem, $(\{a, d, c\}, \{b, 0\})$.

We indicate here that the solution process dictated by the above graphical method can be also produced by the basic algorithm which is also applicable to binary situations with $x \in \{-1, 1\}^n$. Expressing $f(x)$ as

$$\begin{aligned} f_4(x_1, x_2, x_3, x_4) &= 2x_1x_2 - 3x_1x_4 - x_3x_4 + 2.5x_1 - 2x_2 + 3x_3 + 1.5x_4 \\ &= x_1\Delta_4(x_2, x_3, x_4) + \Theta_4(x_2, x_3, x_4), \end{aligned}$$

where $\Delta_4 = 2x_2 - 3x_4 + 2.5$ and $\Theta_4 = -x_3x_4 - 2x_2 + 3x_3 + 1.5x_4$, we have

$$\phi_4(x_2, x_3, x_4) = \frac{1}{2}(1 - x_2)(1 + x_4) - 1,$$

which leads to a reduced form of the objective function

$$\begin{aligned} f_3(x_2, x_3, x_4) &= \phi_4(x_2, x_3, x_4)\Delta_4(x_2, x_3, x_4) + \theta_4(x_2, x_3x_4) \\ &= 1.25x_2x_4 - x_3x_4 - 2.75x_2 + 3x_3 + 3.25x_4 - 3.75. \end{aligned}$$

Note that the graphical representation of the max-cut problem corresponding to $f_3(x_2, x_3, x_4)$ is exactly Fig. 3(b). We further write f_3 in the following form,

$$f_3(x_2, x_3, x_4) = x_2\Delta_3(x_3, x_4) + \Theta_3(x_3, x_4),$$

with $\Delta_3 = 1.25x_4 - 2.75$ and $\Theta_3 = -x_3x_4 + 3x_3 + 3.25x_4 - 3.75$. We can derive $\phi_3(x_3, x_4) = 1$ which yields

$$f_2(x_3, x_4) = 3x_3 + 4.5x_4 - x_3x_4 - 6.5,$$

whose graphical representation is exactly Fig. 3(c). Minimizing $f_2(x_3, x_4)$ yields $x_3^* = -1$ and $x_4^* = -1$. We can then determine $x_2^* = \phi_3(x_3, x_4) = 1$ and $x_1^* = \phi_4(x_2, x_3, x_4) = -1$.

When the corresponding graph of problem (BQP) is serial-parallel, there are at least one row and one column in Q that have no more than two non-zero elements. This pattern remains unchanged during the reduction process. As ϕ_k is at most a quadratic function, f_k remains to be a quadratic function. In essence, if the structure of (BQP) is governed by a serial-parallel graph, the coupling among x_i's is low, and the problem can be solved efficiently by the basic algorithm.

6 Problem (0–1QP) Defined by a Logic Circuit

Let $w_{ij} = -2q_{ij}$, $I_i = -c_i$ for $i, j = 1, 2 \ldots, n$. The objective function $f(x) = x^T Qx + c^T x$ in (0–1QP) can be expressed as the following form:

$$E(x) = -\sum_{1 \le i < j \le n} w_{ij} x_i x_j - \sum_{i=1}^{n} I_i x_i,$$

which can be viewed as the energy function of a neural network where $w_{ij} \in \mathbb{R}$ is the weight associated with the connection between neurons j and i, $x_i \in \{0, 1\}$ is the activation value of neuron i, and $I_i \in \mathbb{R}$ is the threshold of neuron i. For example, the following objective function

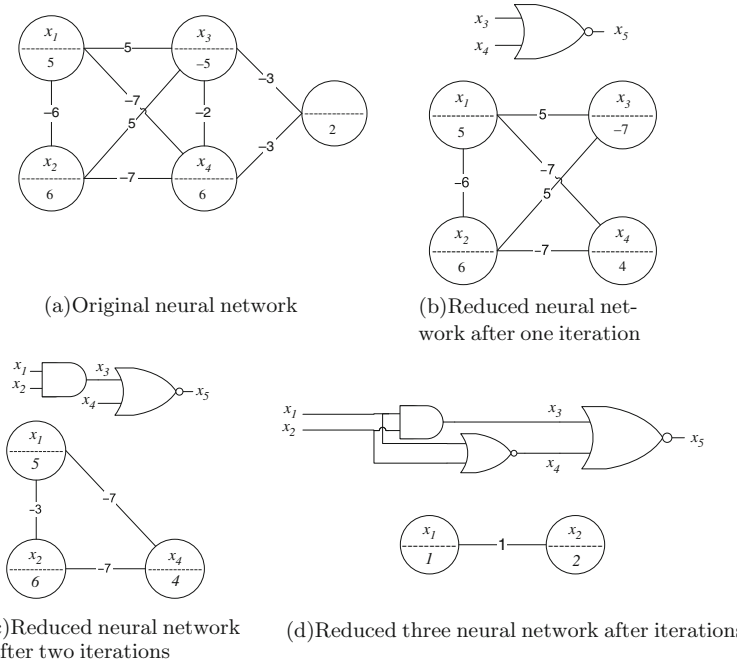
$$\begin{aligned} f = \; & -[-6x_1x_2 + 5x_1x_3 - 7x_1x_4 + 5x_2x_3 - 7x_2x_4 - 2x_3x_4 - 3x_3x_5 - 3x_4x_5] \\ & -[5x_1 + 6x_2 - 5x_3 + 6x_4 + 2x_5] \end{aligned}$$

can be expressed as the energy function of the neural network in Fig. 5(a).

It can be verified easily from the example in Fig. 5(a) that for any value of x_3 and x_4, we should assign x_5 at $x_3 \bar{\vee} x_4 = 1 - \max\{x_3, x_4\}$, i.e., optimal x_5 which minimizes the energy function should be the output of a NOR logic gate if we assign x_3 and x_4 to be the inputs of the gate. This conclusion can be also derived from our earlier discussion of the basic algorithm.

Let x_j and x_k be the inputs to a logic gate. Then x_i is the output of an AND logic gate if $x_i = x_j \wedge x_k = \min\{x_j, x_k\}$, the output of an OR logic gate if $x_i = x_j \vee x_k = \max\{x_j, x_k\}$, the output of a NAND logic gate if $x_i = x_j \bar{\wedge} x_k = 1 - \min\{x_j, x_k\}$, and the output of a NOR logic gate if $x_i = x_j \bar{\vee} x_k = 1 - \max\{x_j, x_k\}$. We can now relate the following special form of the three-variable energy function,

$$E(x_i, x_j, x_k) = -[w(x_i x_j + x_i x_k) + w_{jk} x_j x_k] - [Ix_i + I_j x_j + I_k x_k] + K,$$



(a)Original neural network

(b)Reduced neural network after one iteration

(c)Reduced neural network after two iterations

(d)Reduced three neural network after iterations

Fig. 5. The original and reduced neural networks of the example problem

with these four different logic gates. Using the basic algorithm, we can identify eight cases of different combinations of w and I and their corresponding logic gates, which are given in the following table. Figure 6 offers details in figuring out these eight cases. For example, both conditions of $w < 0$ and $-w - I < 0 < -2w - I$ give rise to case 6.

x_j	x_k	$-w(x_j + x_k) - I$	Cases 1 & 5	Case 2	Case 3	Cases 4 & 8	Case 6	Case 7
0	0	$-I$	< 0	≥ 0	≥ 0	≥ 0	< 0	< 0
0	1	$-w - I$	< 0	< 0	≥ 0	≥ 0	< 0	≥ 0
1	0	$-w - I$	< 0	< 0	≥ 0	≥ 0	< 0	≥ 0
1	1	$-2w - I$	< 0	< 0	< 0	≥ 0	≥ 0	≥ 0
		Logic Gate		OR	AND		NAND	NOR
		$\phi(x_j, x_k)$	1	$x_j + x_k - x_j x_k$	$x_j x_k$	0	$1 - x_j x_k$	$(1 - x_j)(1 - x_k)$

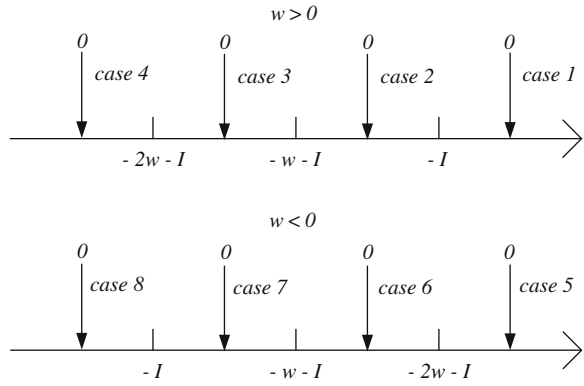


Fig. 6. Eight cases of different combinations of w and I

Replacing x_i by $\phi(x_j, x_k)$ in the four cases associated with different digital logic gates yields a reduced form for the energy function,

$$\bar{E}(x_j, x_k) = -\bar{w}_{jk} x_j x_k - [\bar{I}_j x_j + \bar{I}_k x_k] + \bar{K},$$

where the calculation of $\bar{w}_{jk}$, $\bar{I}_j$, and $\bar{I}_k$ is summarized in the following table:

AND	$\bar{I}_j = I_j$, $\bar{I}_k = I_k$ and $\bar{w}_{jk} = w_{jk} + 2w + I$
OR	$\bar{I}_j = I_j + w + I$, $\bar{I}_k = I_k + w + I$ and $\bar{w}_{jk} = w_{jk} - I$
NAND	$\bar{I}_j = I_j + w$, $\bar{I}_k = I_k + w$ and $\bar{w}_{jk} = w_{jk} - 2w - I$
NOR	$\bar{I}_j = I_j - I$, $\bar{I}_k = I_k - I$ and $\bar{w}_{jk} = w_{jk} + I$

Based on the above recognition between problem (0–1QP) and logic circuits, Chakradhar and Bushnell have designed an iterative method [9] to check whether or not a neural network corresponding to (0–1QP) can be converted into a logic circuit. If we are able to construct a logic circuit such that all the consistent input/output values together minimize the energy function of the neural network, then the original problem (0–1QP) can be solved by a linear-time algorithm.

The assumptions to ensure that the quadratic function f can be transformed into a combinational logic circuit are as follows: (i) the neural network corresponding to the energy function and all the reduced neural networks generated during the iteration have at least one vertex of degree one or two and (ii) both edges incident to the vertex with degree two have equal weights.

A satisfaction of the above assumptions will enable us, in each iteration, to identify a vertex with degree one or two by uniquely determining the corresponding logic gate.

Let us now apply this solution scheme to the example problem in Fig. 5(a). As the terms involving x_5 satisfy the condition of NOR logic gate with x_3 and x_4 being the inputs and x_5 being the output: $w_{35} = w_{45} = w = -3 < 0$, $w_{l5} = 0$, for $l \neq 3$ and 4, $-I = -2 < 0 < -w - I = 1$, we express x_5 as $(1 - x_3)(1 - x_4)$, resulting in the reduced neural network given in Fig. 5(b).

We find in Fig. 5(b) that x_1, x_2, and x_3 satisfy the condition of AND logic gate with x_1 and x_2 being the inputs and x_3 being the output: $w_{13} = w_{23} = w = -5 > 0$, $w_{34} = 0$, and $-2w - I = -3 < 0 < -w - I = 2$. Expressing x_3 as x_1x_2 results in the reduced neural network given in Fig. 5(c).

From Fig. 5(d), we can figure out that x_1, x_2, and x_4 satisfy the condition of NOR logic gate with x_1 and x_2 being the inputs and x_4 being the output: $w_{14} = w_{24} = w = -7 < 0$, $-I = -4 < 0 < -w - I = 3$. Expressing x_4 as $(1 - x_1)(1 - x_2)$ results in the reduced neural network given in Fig. 5(d).

Solving the reduced binary quadratic minimization problem,

$$\min -x_1x_2 - x_1 - 2x_2,$$

yields $x_1^* = 1$ and $x_2^* = 1$. Further calculation gives $x_3^* = x_1^*x_2^* = 1$, $x_4^* = (1 - x_1^*)(1 - x_2^*) = 0$, and $x_5^* = (1 - x_3^*)(1 - x_4^*) = 0$.

The condition to define problem (0–1QP) by a logic circuit is very strict, especially the requirement of the same weights of the edges incident to the vertex of degree two which is to be removed. If problem (0–1QP) can be defined by a logic circuit, matrix Q and its reduced forms generated during the process all have, at least, one row and one column that have no more than two non-zero elements, and when there are two, these two elements are the same. These conditions are stronger than the conditions for problems defined by the series-parallel graph.

7 SDP Representation of Lagrangian Dual and Polynomial Solvability

Based on our recent finding in [33], we discuss in this section how to identify a polynomially solvable subclass of (BQP) using Lagrangian dual. Notice that (P) can be rewritten as

$$\begin{aligned} (BQP_{\mathrm{c}}) \quad & \min f(x) = x^T Q x + c^T x \\ & \text{s.t. } x_i^2 - 1 = 0,\ i = 1, \ldots, n. \end{aligned}$$

Dualizing each $x_i^2 - 1 = 0$ by a multiplier λ_i, we get the Lagrangian relaxation problem (L_λ):

$$\begin{aligned} d(\lambda) &= \inf_{x \in \mathbb{R}^n} L(x, \lambda) := f(x) + \sum_{i=1}^{n} \lambda_i (x_i^2 - 1) \\ &= \inf_{x \in \mathbb{R}^n} \left\{ x^T (Q + \mathrm{diag}(\lambda)) x + c^T x - e^T \lambda \right\}, \end{aligned} \tag{17}$$

where $e = (1, \ldots, 1)^T$ and $\mathrm{diag}(\lambda)$ denotes the diagonal matrix with λ_i being its ith diagonal element. Obviously, the weak duality holds

$$d(\lambda) \le f(x), \text{ for any} x \in \{-1, 1\}^n.$$

The dual problem of (P_c) (or (BQP)) is

$$(D) \max_{\lambda \in \mathbb{R}^n} d(\lambda).$$

Notice that the dual problem (D) can be rewritten as

$$v(D) = \max_{\lambda \in \mathbb{R}^n} d(\lambda) = \max_{\lambda \in \mathbb{R}^n} \inf_{x \in \mathbb{R}^n} \Big\{ x^T[Q + \mathrm{diag}(\lambda)]x + c^T x - e^T\lambda \Big\},$$

which has an equivalent form:

$$\begin{aligned} v(D) = \max_{(\lambda,\tau) \in \mathbb{R}^{n+1}} \ & -\tau \\ \text{s.t. } & x^T[Q + \mathrm{diag}(\lambda)]x + c^T x - e^T\lambda \ge -\tau, \ x \in \mathbb{R}^n. \end{aligned} \tag{18}$$

Let function $g(x)$ be the constraint in problem (18),

$$g(x) = x^T[Q + \mathrm{diag}(\lambda)]x + c^T x - e^T\lambda + \tau.$$

Using homogeneous quadratic form (see [31] and Section 3.4 in [6]), we show below that $g(x) \ge 0$, $\forall\ x \in \mathbb{R}^n$, the satisfaction of the constraint in problem (18), is equivalent to

$$G(x,t) = (x^T, t) \begin{pmatrix} Q + \mathrm{diag}(\lambda) & \frac{1}{2}c \\ \frac{1}{2}c^T & \tau - e^T\lambda \end{pmatrix} \begin{pmatrix} x \\ t \end{pmatrix} \ge 0 \quad \forall\ (x,t) \in \mathbb{R}^{n+1},$$

which holds true if and only if

$$\begin{pmatrix} Q + \mathrm{diag}(\lambda) & \frac{1}{2}c \\ \frac{1}{2}c^T & \tau - e^T\lambda \end{pmatrix} \succeq 0.$$

Since $g(x) = G(x, 1)$, $G(x,t) \ge 0$ for all $(x,t) \in \mathbb{R}^{n+1}$ implies $g(x) \ge 0$ for all $x \in \mathbb{R}^n$. Now, suppose that $g(x) \ge 0$ for all $x \in \mathbb{R}^n$. Then, $g(t^{-1}x) \ge 0$ for all $x \in \mathbb{R}^n$ and $t \ne 0$, which implies

$$t^{-2}x^T[Q + \mathrm{diag}(\lambda)]x + t^{-1}c^T x - e^T\lambda + \tau \ge 0 \quad \forall x \in \mathbb{R}^n, \ t \ne 0,$$

or, equivalently,

$$G(x,t) = x^T[Q + \mathrm{diag}(\lambda)]x + c^T xt + (\tau - e^T\lambda)t^2 \ge 0 \quad \forall x \in \mathbb{R}^n, \ t \ne 0.$$

By continuity, we have

$$G(x,t) = x^T[Q + \mathrm{diag}(\lambda)]x + c^T xt + (\tau - e^T\lambda)t^2 \ge 0 \quad \forall\ (x,t) \in \mathbb{R}^{n+1}.$$

Thus, the dual problem (D) can be expressed by the following equivalent semidefinite programming formulation,

$$(D_{\rm SDP}) \quad \max_{(\lambda,\tau)\in\mathbb{R}^{n+1}} \ -\tau \qquad (19)$$
$$\text{s.t.} \ \begin{pmatrix} Q+\text{diag}(\lambda) & \frac{1}{2}c \\ \frac{1}{2}c^T & \tau - e^T\lambda \end{pmatrix} \succeq 0.$$

Since $(D_{\rm SDP})$ is a semidefinite programming problem, it is polynomially solvable. The above discussion implies that if there is no duality gap between (BQP) and $(D_{\rm SDP})$, i.e., $v(BQP) = v(D) = v(D_{\rm SDP})$, then $v(BQP)$ is polynomially computable.

The following theorem further gives a sufficient condition for the polynomial solvability of (BQP).

Theorem 5. *Assume that the optimal solution λ^* to $(D_{\rm SDP})$ satisfies $Q^* = Q + \text{diag}(\lambda^*) \succ 0$. Then $x^* = -\frac{1}{2}(Q^*)^{-1}c$ is the unique optimal solution to (BQP) and $v(BQP) = v(D) = v(D_{\rm SDP})$. Moreover, (BQP) is polynomially solvable.*

Proof. From [5], we know that, for any $\lambda \in \mathbb{R}^n$, $d(\lambda) > -\infty$ with x solving (L_λ) if and only if

(i) $Q + \text{diag}(\lambda) \succeq 0$;
(ii) $(Q + \text{diag}(\lambda))x = -\frac{1}{2}c$.

Since the optimal solution λ^* to $(D_{\rm SDP})$ satisfies $Q^* \succ 0$, we can verify that (D) or (D_{SDP}) is equivalent to the following problem:

$$(D_1) \quad \sup \ \Phi(\lambda) = -\frac{1}{4}c^T(Q+\text{diag}(\lambda))^{-1}c - e^T\lambda \qquad (20)$$
$$\text{s.t.} \ Q + \text{diag}(\lambda) \succ 0.$$

Thus, λ^*, an interior point of the feasible region of (D_1), also solves (D_1). By KKT theorem, we must have $\nabla\Phi(\lambda^*) = 0$, where Φ is defined in (20). Calculating the gradient of Φ at λ^* and setting it at zero yield the following:

$$\frac{1}{4}c^T(Q^*)^{-1}\text{diag}(e_i)(Q^*)^{-1}c = 1, \ i = 1, \ldots, n. \qquad (21)$$

This is to say $(x_i^*)^2 = 1$, for all $i = 1, \ldots, n$. Thus $x^* \in \{-1, 1\}^n$. As $Q^* \succ 0$, x^* is the unique optimal solution to (BQP) and $v(BQP) = v(D) = v(D_{\rm SDP}) = v(D_1)$. Moreover, since λ^* is polynomially computable and $x^* = -\frac{1}{2}(Q + \text{diag}(\lambda^*))^{-1}c$, we deduce that (BQP) is polynomially solvable.

8 Conclusions

We have summarized the state of the art of polynomially solvable cases for binary quadratic programming problems. Separating certain easy subclasses

from a general NP-hard class facilitates identification schemes to peel off hard covers of some seemingly intractable, but actually manageable, binary quadratic programming problems. Furthermore, investigation of this subject not only helps us better understand inherent nature of the problem but also stimulates innovative thinking for development of solution schemes for general binary quadratic programming problems.

Acknowledgment This research was partially supported by the Research Grants Council of Hong Kong under grant 414207 and by the National Natural Science Foundation of China under grants 70671064 and 70832002.

References

1. Allemand, K., Fukuda, K., Liebling, T.M., Steiner, E.: A polynomial case of unconstrained zero-one quadratic optimization. Math. Program. 91, 49–52 (2001)
2. Avis, D., Kukuda, K.: Reverse search for numeration. Discrete Appl. Math. 65, 21–46 (1996)
3. Barahona, F.: A solvable case of quadratic 0–1 programming. Discrete Appl. Math. 13, 23–26 (1986)
4. Barahona, F., Jünger, M., Reinelt, G.: Experiments in quadratic 0–1 programming. Math. Program. 44, 127–137 (1989)
5. Beck, A., Teboulle, M.: Global optimality conditions for quadratic optimization problems with binary constraints. SIAM J. Optim. 11, 179–188 (2000)
6. Ben-Tal, A.: Conic and Robust Optimization, Lecture Notes, Universita di Roma La Sapienzia, Rome, Italy (2002)
7. Billionnet, A., Elloumi, S.: Using a mixed integer quadratic programming solver for the unconstrained quadratic 0–1 problem. Math. Program. 109, 55–68 (2007)
8. Chaillou, P., Hansen, P., Mahieu, Y.: Best network flow bounds for the quadratic knapsack problem. Lect. Notes Math. 1403, 226–235 (1986)
9. Chakradhar, S.T., Bushnell, M.L.: A solvable class of quadratic 0–1 programming. Discrete Appl. Math. 36, 233–251 (1992)
10. Chardaire, P., Sutter, A.: A decomposition method for quadratic zero-one programming. Manage. Sci. 41, 704–712 (1995)
11. Crama, Y., Hansen, P., Jaumard, B.: The basic algorithm for pseudo-Boolean programming revisited. Discrete Appl. Math. 29, 171–185 (1990)
12. Delorme, C., Poljak, S.: Laplacian eigenvalues and the maximum cut problem. Math. Program. 62, 557–574 (1993)
13. Ferrez, J.A., Fukuda, K., Liebling, T.M.: Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm. Eur. J. Oper. Res. 166, 35–50 (2005)
14. Gallo, G., Grigoridis, M., Tarjan, R.E.: A fast parametric maximum flow algorithm and applications. SIAM J. Comput. 18, 30–55 (1989)
15. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. WH Freeman & Co., New York (1979)
16. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. Assoc. Comput. Mach. 42, 1115–1145 (1995)

17. Goldberg, A.V., Tarjar, R.E.: A new approach to the maximum flow problem. *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, Berkeley, CA, 136–146 (1986)
18. Hammer, P.L., Hansen, P., Simeone, B.: Roof duality, complementation and persistency in quadratic 0–1 optimization. Math. Program. 28, 121–155 (1984)
19. Hammer, P.L., Rudeanu, S.: Boolean Methods in Operations Research and Related Areas, Springer, Berlin (1968)
20. Hansen, P., Jaumard, B., Mathon, V.: Constrained nonlinear 0–1 programming. ORSA J. Comput. 5, 97–119 (1993)
21. Helmberg, C., Rendl, F.: Solving quadratic (0,1)-problems by semidefinite programs and cutting planes. Math. Program. 82, 291–315 (1998)
22. Li, D., Sun, X.L.: Nonlinear Integer Programming, Springer, New York (2006)
23. Li, D., Sun, X.L., Liu, C.L.: An exact solution method for quadratic 0–1 programming: a geometric approach. Technical Report, Chinese University of Hong Kong. Department of Systems Engineering and Engineering Management (2006)
24. Mcbride, R.D., Yormark, J.S.: An implicit enumeration algorithm for quadratic integer programming. Manage. Sci. 26, 282–296 (1980)
25. Nemhauser, G.L., Wolsey, L.A.: Integer and Combinatorial Optimization, Wiley, New York (1988).
26. Pardalos, P.M., Rodgers, G.P.: Computational aspects of a branch-and-bound algorithm for quadratic zero-one programming. Computing 45, 131–144 (1990)
27. Phillips, A.T., Rosen, J.B.: A quadratic assignment formulation of the molecular conformation problem. J. Global Optim. 4, 229–241 (1994)
28. Picard, J.C., Ratliff, H.D.: Minimum cuts and related problems. Networks 5, 357–370 (1975)
29. Rendl, F., Rinaldi, G., Wiegele, A.: Solving max-cut to optimality by intersecting semidefinite and polyhedral relaxations. Lect. Notes Comput. Sci. 4513, 295–309 (2007)
30. Rhys, J.: A selection problem of shared fixed costs and network flows. Manage. Sci. 17, 200–207 (1970)
31. Shor, N.Z.: Quadratic optimization problems. Sov. J. Comput. Syst. Sci. 25, 1–11 (1987)
32. Sleumer, N.: Output-sensitive cell enumeration in hyperplane arrangements. Nordic J. Comput. 6, 137–161 (1999)
33. Sun, X.L., Liu, C.L., Li, D., Gao, J.J.: On duality gap in binary quadratic optimization. Technical Report, Chinese University of Hong Kong. Department of Systems Engineering and Engineering Management (2007)
34. Zaslavsky, T.: Facing up to arrangements: face-count formulas for partitions of space by hyperplanes. Mem. Am. Math. Soc. 1, 1–101 (1975)

Generalized Solutions of Multi-valued Monotone Quasi-variational Inequalities

Baasansuren Jadamba[1], Akhtar A. Khan[1], Fabio Raciti[2], and Behzad Djafari Rouhani[3]

[1] School of Mathematical Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA
bxjsma,aaksma@rit.edua

[2] Dipartimento di Matematica e Informatica, Università di Catania, Viale A. Doria 6–I, 95125, Catania, Italy
fraciti@dmi.unict.it

[3] Department of Mathematical Sciences, University of Texas at El Paso, El Paso, Texas, 79968, USA
behzad@math.utep.edu

Summary. An ill-posed quasi-variational inequality with multi-valued maps can be conveniently formulated as a parameter identification problem on the graph of a variational selection. Using elliptic regularization for parametric variational inequalities, it is possible to pose another parameter identification problem that gives a stable approximation procedure for the ill-posed problem. The results are quite general and are applicable to ill-posed variational inequalities, inverse problems, split-feasibility problem, among others.

Key words: quasi-variational inequalities, parameter identification, regularization, ill-posed, multi-valued monotone maps, inverse problems

1 Introduction

Let $\mathcal{B}$ be a uniformly convex Banach space with a strictly convex topological dual $\mathcal{B}*$. We specify the duality pairing between $\mathcal{B}$ and $\mathcal{B}^*$ by $\langle\cdot,\cdot\rangle$, whereas $\|\cdot\|$ stands for the norm in $\mathcal{B}$ as well as in $\mathcal{B}^*$. Let $\mathcal{C}$ be a nonempty, closed, and convex subset of $\mathcal{B}$, and let $\mathcal{K}:\mathcal{C}\rightrightarrows 2^{\mathcal{C}}$ be a set-valued map such that for every $v\in\mathcal{C}$, the set $\mathcal{K}(v)$ is nonempty, closed, and convex. Let $\mathcal{F}:\mathcal{B}\rightrightarrows 2^{\mathcal{B}^*}$ be a given multi-valued map, and let $f\in\mathcal{B}^*$. The effective domain and the graph of any map $\mathcal{A}:\mathcal{B}\rightrightarrows 2^{\mathcal{B}^*}$ are denoted by $\mathcal{D}(\mathcal{A})$ and $\mathcal{G}(\mathcal{A})$, respectively. The strong convergence and the weak convergence in $\mathcal{B}$ as well as in $\mathcal{B}^*$ are specified by $\to$ and $\rightharpoonup$, respectively.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_12,

The present study is focused on the following multi-valued quasi-variational inequality (QVI): find $x \in \mathcal{C}$ such that $x \in \mathcal{K}(x)$, and there exists $w \in \mathcal{F}(x)$ satisfying the variational inequality

$$\langle w - f, z - x \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(x). \tag{1}$$

The above QVI includes many important problems of interest as particular cases. For example, if $\mathcal{F}$ is single valued, then (1) recovers the following QVI: find $x \in \mathcal{C}$ such that $x \in \mathcal{K}(x)$ and

$$\langle \mathcal{F}(x) - f, z - x \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(x). \tag{2}$$

The above problem was introduced by Bensoussan and Lions [3] in connection with a problem of impulse control. A general treatment of (2) was made by Mosco [27]. If additionally $\mathcal{K}(x) = \mathcal{C}$ for every $x \in \mathcal{C}$, then (1) recovers the following variational inequality: find $x \in \mathcal{C}$ such that

$$\langle \mathcal{F}(x) - f, z - x \rangle \geq 0 \quad \text{for every } z \in \mathcal{C}. \tag{3}$$

Variational inequality (3) appears as a necessary optimality condition for the output least-squares formulation in the inverse problem of identifying coefficients in partial differential equations (see [13]). Furthermore, (3) also emerges as a necessary and sufficient optimality condition for the same inverse problem through the modified output least-squares (see [14, 15]) and the equation-error approach (see [16]). Recently, Noor [28] proved the equivalence between the split-feasibility problem and (3). In recent years, the split-feasibility problem has attracted much attention due to its interesting applications in image processing and inverse problems (see [7, 8, 30]). Some regularization methods for variational inequalities are available in [10, 17, 20, 22, 25], among others.

Notice that if for every $x \in \mathcal{C}$, $\mathcal{K}(x)$ is a closed and convex cone with its apex at the origin and $f = 0$, then (1) collapses to the generalized complementarity problem: find $x \in \mathcal{C}$ such that

$$x \in \mathcal{K}(x), \quad w \in \mathcal{F}(x) \cap \mathcal{K}^*(x), \quad \langle w, x \rangle = 0, \tag{4}$$

where $\mathcal{K}^*(x)$ denotes the positive polar of $\mathcal{K}(x)$. If additionally $\mathcal{K}(x) \equiv \mathcal{C}$, then (4) recovers the classical complementarity problem (see [18]). For a detailed study of complementarity problems we refer the reader to Isac et al. [19]. The equivalence between (1) and (4) is given by Giannessi [11].

In recent years the theory of variational and quasi-variational inequalities has emerged as one of the most promising branches of pure, applied, and industrial mathematics. This theory provides us with a convenient mathematical apparatus for studying a wide range of problems arising in diverse fields such as structural mechanics, elasticity, economics, optimization, optimal control,

inverse problems, financial mathematics (see [2, 21]). The existence theory for quasi-variational inequalities is challenging and it requires that a variational inequality and a fixed point problem should be solved simultaneously. Consequently many solution techniques which are available for variational inequalities have not been extended for quasi-variational inequalities. For example, regularization and penalization methods for monotone variational inequalities have almost reached a saturation point. However, for quasi-variational inequalities these approaches have not been fully explored and there are many questions to be answered.

In this chapter, our objective is to develop a regularization theory for ill-posed quasi-variational inequalities involving multi-valued maps. The basic idea is to cast (1) as a parameter identification problem defined on the graph of a variational selection. To explain this idea, we fix an element $v \in \mathcal{C}$ and consider the following parametric variational inequality (PVI) with v as the parameter: find $x \in \mathcal{K}(v)$ such that there exists $w \in \mathcal{F}(x)$ satisfying the variational inequality

$$\langle w - f, z - x \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v). \tag{5}$$

We define the variational selection $\mathcal{S} : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$ by the condition that for any $v \in \mathcal{C}$, the set $\mathcal{S}(v)$ is the set of all solutions of the PVI with parameter v.

Consider the following parameter identification problem (PIP): find $(x, u) \in \mathcal{G}(\mathcal{S})$ such that

$$\|x - u\|^2 \leq \|y - v\|^2 \quad \text{for every } (y, v) \in \mathcal{G}(\mathcal{S}). \tag{6}$$

An element $x \in \mathcal{C}$ will be referred to as a generalized solution of QVI (1) if (x, u) is a solution to the above parameter identification problem.

Evidently, $x \in \mathcal{C}$ is a solution of (1) if and only if x is a fixed point of $\mathcal{S}$. Moreover

- If (6) is solvable, and $\|x - u\| = 0$ where (x, u) is a solution, then (1) is solvable.
- If (1) is solvable, then (6) is also solvable, and their solution sets coincide.

To exploit the advantages of a minimization formulation many researchers have focused on (6) rather than on (1). Although the above technique has its origin in the original work of Mosco [27], we would like to acknowledge the contribution of Bruckner who systematically explored the connection between (1) and (6) (see [4, 5]).

Quasi-variational inequality (1) and many of its particular cases mentioned above are in general ill-posed. That is, a small noise in the data can lead to uncontrollable errors in its solution. One of our main objectives is to develop a stable approximation scheme for (1) when instead of the exact data $(\mathcal{F}, f)$ only the noisy data are available. The key idea is to approximate (6) by a

regularized PIP. For this, we consider the following regularized parametric variational inequality (RPVI) for a fixed parameter $v \in \mathcal{C}$: find $x_\epsilon \in \mathcal{K}(v)$ such that there exists $w_\epsilon \in \mathcal{F}(x_\epsilon)$ satisfying

$$\langle w_\epsilon + \epsilon J(x_\epsilon - v) - f, z - x_\epsilon \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v), \tag{7}$$

where $\epsilon > 0$ and J is the normalized duality map. (The general case of noisy data will be studied in Section 2.) We define the variational selection $\mathcal{S}_\epsilon : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$ by the condition that for any $v \in \mathcal{C}$, the set $\mathcal{S}_\epsilon(v)$ is the set of all solutions of (7).

The regularized parameter identification problem (RPIP) then seeks $(x, u) \in \mathcal{G}(\mathcal{S}_\epsilon)$ such that

$$\|x - u\|^2 \leq \|y - v\|^2 \quad \text{for every } (y, v) \in \mathcal{G}(\mathcal{S}_\epsilon). \tag{8}$$

We will show that, under suitable conditions, a sequence of solutions of (8) converges to a solution of (6). Moreover, by making full use of the recent developments in the theory of multi-valued maps (see [1]), we present a general regularization theory for quasi-variational inequalities. Some of our results are new even for quasi-variational inequalities with single-valued monotone maps.

We conclude this introduction by stating an existence result for (6). A proof of this result (based on the classical Weierstrass theorem) can be found in Bruckner [4] where the focus is on the elliptic regularization of quasi-variational inequalities with single-valued monotone maps in the framework of parameter identification problems.

Lemma 1. *Assume that there exists* $(x_0, u_0) \in \mathcal{G}(\mathcal{S})$ *such that the set*

$$\Phi = \{(y, v) \in \mathcal{G}(\mathcal{S}) | \quad \|y - v\| \leq \|x_0 - u_0\|\} \tag{9}$$

is weakly compact. Then (6) has a nonempty solution set.

2 Main Results

We first focus on the solvability of (6) by using Lemma 1. To prove that the set Φ in (9) is weakly closed, it suffices to show that $\mathcal{G}(\mathcal{S})$ is weakly closed. For this we recall the notion of Mosco convergence (see [26]). The map $\mathcal{K} : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$ is said to be M-continuous if it satisfies the following:

(M1) For every sequence (x_n) with $x_n \rightharpoonup x$, and for each $y \in \mathcal{K}(x)$, there exists a sequence (y_n), with $y_n \in \mathcal{K}(x_n)$ and $y_n \to y$.

(M2) For $y_n \in \mathcal{K}(x_n)$ with $x_n \rightharpoonup x$ and $y_n \rightharpoonup y$, we have $y \in \mathcal{K}(x)$.

Lemma 2. *Assume that $\mathcal{F}$ is bounded and satisfies the following condition ($\mathcal{GM}$): If $(x_n, w_n) \in \mathcal{G}(\mathcal{F})$, with $x_n \rightharpoonup x$ and $w_n \rightharpoonup w$, satisfies $\limsup_{n\to\infty}\langle w_n, x_n - x\rangle \leq 0$, then $w \in \mathcal{F}(x)$ and $\langle w_n, x_n\rangle \to \langle w, x\rangle$. Assume that $\mathcal{K}$ is M-continuous. Then the graph of the variational selection $\mathcal{S}$ is weakly closed.*

Proof. Let $(y_n, v_n) \in \mathcal{G}(\mathcal{S})$ be such that $y_n \rightharpoonup y$ and $v_n \rightharpoonup v$. We will show that $(y, v) \in \mathcal{G}(\mathcal{S})$. The set $\mathcal{C}$ being convex and closed is also weakly closed and hence $v \in \mathcal{C}$. From the containment $(y_n, v_n) \in \mathcal{G}(\mathcal{S})$, we infer that $y_n \in \mathcal{K}(v_n)$ and that there exists $w_n \in \mathcal{F}(y_n)$ such that

$$\langle w_n - f, z - y_n\rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v_n). \tag{10}$$

Notice that $y_n \in \mathcal{K}(v_n)$, in view of (M2), implies that $y \in \mathcal{K}(v)$. Moreover, due to (M1), there exists $z_n \in \mathcal{K}(v_n)$ such that $z_n \to y$. By substituting $z = z_n$ in (10), rearranging the terms, and using the boundedness of $\mathcal{F}$, we obtain

$$\begin{aligned}\limsup_{n\to\infty}\langle w_n, y_n - y\rangle &\leq \limsup_{n\to\infty}\langle w_n, y_n - z_n\rangle + \limsup_{n\to\infty}\langle w_n, z_n - y\rangle,\\ &\leq \limsup_{n\to\infty}\{\langle f, y_n - z_n\rangle\},\\ &\leq 0.\end{aligned}$$

In view of ($\mathcal{GM}$), for a subsequence (w_n) such that $w_n \rightharpoonup w$ and satisfying the above inequality, we have $w \in \mathcal{F}(y)$ and $\lim_{n\to\infty}\langle w_n, y_n\rangle = \langle w, y\rangle$. We claim that

$$\langle w - f, z - y\rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v).$$

Let $z \in \mathcal{K}(v)$ be arbitrary. In view of (M1) there exists a sequence (z_n) such that $z_n \in \mathcal{K}(v_n)$, and $z_n \to z$. Therefore

$$\begin{aligned}\langle w, y - z\rangle &= \liminf_{n\to\infty}\langle w_n, y_n - z_n\rangle,\\ &\leq \limsup_{n\to\infty}\langle f, y_n - z_n\rangle,\\ &\leq \langle f, y - z\rangle.\end{aligned}$$

Since $z \in \mathcal{K}(v)$ is arbitrary, we deduce that $(y, v) \in \mathcal{G}(S)$. The proof is complete. □

Evidently, we only require that $\mathcal{F}$ is bounded on the solutions of PVIs. If $\mathcal{F}$ is monotone and contains $\mathcal{C}$ in the interior of its domain, then we do not need the boundedness assumption on $\mathcal{F}$. Moreover, if $\mathcal{F}$ is maximal monotone, with $\mathcal{D}(\mathcal{F}) = \mathcal{B}$, then it satisfies ($\mathcal{GM}$) condition (see [6]).

The following result gives conditions ensuring that the set Φ is bounded.

Lemma 3. *Assume that for $v \in \mathcal{C}$, there are $m(v) \in \mathcal{K}(v)$, and positive constants a and b such that $\|m(v)\| \leq a\|v\| + b$. Assume that for $y \in \mathcal{S}(v)$, with $\|y - v\| \leq a_1 < \infty$, we have*

$$\lim_{\|y\|\to\infty} \frac{\langle w, y - m(v)\rangle}{\|y\|} = \infty \quad \textit{for every } w \in \mathcal{F}(y). \tag{11}$$

Then the set Φ is bounded.

Proof. Let $(y, v) \in \mathcal{G}(\mathcal{S})$ be arbitrary with $\|y-v\| \leq \|y_0-x_0\|$. Since $y \in \mathcal{S}(v)$, there exists $w \in \mathcal{F}(y)$ such that

$$\langle w - f, z - y\rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v).$$

We set $z = m(v)$ in the above inequality and rearrange the terms to get a constant c such that

$$\frac{\langle w, y - m(v)\rangle}{\|y\|} \leq c.$$

If $\|y\| \to \infty$, then the above inequality contradicts (11), confirming that the set Φ is bounded. □

To ensure that $\mathcal{S}(v) \neq \emptyset$, for every $v \in \mathcal{C}$, we need to discuss the solvability of the following variational inequality: find $x \in \mathcal{C} \subset \mathcal{B}$ such that for some $w \in \mathcal{F}(x)$, we have

$$\langle w - f, z - x\rangle \geq 0 \quad \text{for every } z \in \mathcal{C}. \tag{12}$$

In the following we recall a few notions and auxiliary results concerning (12).

Definition 1. *Let $\mathcal{F} : \mathcal{B} \rightrightarrows 2^{\mathcal{B}^*}$ be a set-valued map and let $(x, x^*), (y, y^*) \in \mathcal{G}(\mathcal{F})$ be arbitrary. The map $\mathcal{F}$ is said to be*

(a) monotone, if $\langle x^* - y^*, x - y\rangle \geq 0$;
(b) m-monotone, if $\langle x^* - y^*, x - y\rangle \geq m\|x - y\|^2$;
(c) m-relaxed monotone, if $\langle x^* - y^*, x - y\rangle \geq -m\|x - y\|^2$;
(d) maximal monotone, if the graph of $\mathcal{F}$ is not included in the graph of any other monotone operator with the same domain.

The following result is a Minty formulation for multi-valued variational inequalities (12).

Lemma 4. *Let $\mathcal{F} : \mathcal{B} \rightrightarrows 2^{\mathcal{B}^*}$ be a maximal monotone map, let $\mathcal{C}$ be a nonempty closed convex subset of int*$(\mathrm{dom}(\mathcal{F}))$*, and let $f \in \mathcal{B}^*$. Then $x \in \mathcal{C}$ is a solution of (12) if and only if it solves the following Minty variational inequality: find $x \in \mathcal{C}$ such that*

$$\langle w^* - f, z - x\rangle \geq 0 \quad \textit{for every } z \in \mathcal{C}, \ \textit{for every } w^* \in \mathcal{F}(z). \tag{13}$$

Proof. See [1] or [12]. □

The following existence result can be proved by standard monotonicity arguments (see [29]).

Lemma 5. *Let $\mathcal{F}$, $\mathcal{C}$, and f be as in Lemma 4, and let J be the normalized duality map. Then there exists a unique $x \in \mathcal{C}$ and some $w \in \mathcal{F}(x)$ such that*

$$\langle w + \epsilon J(x), y - x\rangle \geq 0 \text{ for every } y \in \mathcal{C}. \tag{14}$$

We would also need the following interesting result (see [1]).

Lemma 6. *Let $\mathcal{A} : \mathcal{B} \rightrightarrows 2^{\mathcal{B}}$ be monotone. If $\bar{x} \in int(\mathcal{D}(\mathcal{A}))$, then there exists a real number $r = r(\bar{x}) > 0$ such that for every $(x, w) \in \mathcal{G}(\mathcal{A})$, we have*

$$\langle w, x - \bar{x}\rangle \geq r\|w\| - (\|x - \bar{x}\| + r)c,$$

where $c := \sup\{\|w\|\| \ \|x - \bar{x}\| \leq r, \quad \text{and} \quad w \in \mathcal{A}(x)\} < \infty$.

We now prepare for the regularization theory. We begin with by connecting the exact data $(\mathcal{F}, f)$ to the noisy data $(\mathcal{F}_n, f_n)$ by the following hypothesis:

- (A_0) For each $n \in \mathbb{N}$, the map $\mathcal{F}_n$ and the map $\mathcal{F}$ are maximal monotone and satisfy $\mathcal{C} \subset \text{int}(\mathcal{D}(\mathcal{F})) \cap \text{int}(\mathcal{D}(\mathcal{F}_n))$. The map $\mathcal{K}$ is M-continuous.
- (A_1) For any $x \in \mathcal{B}$ and for any $w \in \mathcal{F}(x)$ (resp. $w_n \in \mathcal{F}_n(x)$) there exist $w_n \in \mathcal{F}_n(x)$ (resp. $w \in \mathcal{F}(x)$) and $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ which is bounded on bounded sets, such that

$$\|w_n - w\| \leq \alpha_n \kappa(\|x\|), \quad \alpha_n > 0.$$

- (A_2) For each $n \in \mathbb{N}$, $f_n \in \mathcal{B}^*$ and satisfies $\|f_n - f\| \leq \beta_n$, where $\beta_n > 0$.
- (A_3) $\left\{\alpha_n, \beta_n, \dfrac{\alpha_n}{\epsilon_n}, \dfrac{\beta_n}{\epsilon_n}\right\} \to 0$, as $n \to 0$.

For a fixed $v \in \mathcal{C}$, we consider the following regularized parametric variational inequality (RPVI): find $x_n \in \mathcal{K}(v)$ such that there exists $w_n \in \mathcal{F}_n(x_n)$ satisfying the variational inequality

$$\langle w_n + \epsilon_n J(x_n - v) - f_n, z - x_n\rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(v), \quad \epsilon_n > 0. \tag{15}$$

We define the regularized variational selection $\mathcal{S}_n : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$ by the condition that for some $v \in \mathcal{C}$, $\mathcal{S}_n(v)$ is the unique solution of (15).

Finally we pose the following regularized parameter identification problem (RPIP): find $(x_n, u_n) \in \mathcal{G}(\mathcal{S}_n)$ such that

$$\|x_n - u_n\|^2 \leq \|y - v\|^2 \quad \text{for every } (y, v) \in \mathcal{G}(\mathcal{S}_n). \tag{16}$$

In our next result we will only focus on conditions that ensure that a solution of (6) can be approximated by solutions of (16). These conditions turn out to be slightly weaker than the ones that are needed to ensure the solvability of (6) and (16).

Theorem 1. *Assume that (6) and (16) have nonempty solution sets. Assume that for $\mathcal{S}_n : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$, the set $\mathcal{S}_n(\mathcal{C})$ is bounded. Assume that the assumptions (A_0) through (A_3) hold. Then each weak cluster point of the solution sequence $\{(x_n, y_n)\}$ of (16) is a solution of (6).*

Proof. Assume that $(x, u) \in \mathcal{G}(\mathcal{S})$ is a solution of (6). For this fixed u, consider the following RPVI: find $x_n^u \in \mathcal{K}(u)$ such that there exists $w_n^u \in \mathcal{F}_n(x_n^u)$ satisfying the variational inequality:

$$\langle w_n^u + \epsilon_n J(x_n^u - u) - f_n, z - x_n^u \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u). \tag{17}$$

We will show that (x_n^u) is bounded. Let $x^u \in \mathcal{S}(u)$ be arbitrary. Then $x^u \in \mathcal{K}(u)$ and there exists $w^u \in \mathcal{F}(x^u)$, such that

$$\langle w^u - f, z - x^u \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u). \tag{18}$$

In view of (A1), there exists $\bar{w}_n^u \in \mathcal{F}_n(x^u)$ such that $\|\bar{w}_n^u - w^u\| \leq \alpha_n \kappa(\|x^u\|)$. Therefore,

$$\begin{aligned} \langle w^u - w_n^u, x_n^u - x^u \rangle &= \langle w^u - \bar{w}_n^u, x_n^u - x^u \rangle - \langle \bar{w}_n^u - w_n^u, x^u - x_n^u \rangle \\ &\leq \alpha_n \kappa(\|x^u\|) \|x_n^u - x^u\|. \end{aligned}$$

By setting $z = x^u$ in (17), $z = x_n^u$ in (18), and rearranging the two resulting inequalities, we obtain

$$\begin{aligned} \epsilon_n \langle J(x_n^u - u), x_n^u - x^u \rangle &\leq \langle w^u - w_n^u, x_n^u - x^u \rangle + \langle f - f_n, x^u - x_n^u \rangle \\ &\leq (\alpha_n \kappa(\|x^u\|) + \beta_n) \, \|x_n^u - x^u\|. \end{aligned}$$

In view of the properties of the duality map J (see [29]), we have

$$\begin{aligned} \epsilon_n (\|x_n^u - u\| - \|x^u - u\|)^2 &\leq \epsilon_n \langle J(x_n^u - u), x_n^u - x^u \rangle - \epsilon_n \langle J(x^u - u), x_n^u - x^u \rangle \\ &\leq (\alpha_n \kappa(\|x^u\|) + \beta_n + \epsilon_n \|x^u - u\|) \, \|x_n^u - x^u\|, \end{aligned}$$

from which the boundedness of (x_n^u) ensues. Since $\mathcal{B}$ is reflexive, there exists a subsequence (x_n^u) that converges weakly to some $\bar{x}^u \in \mathcal{K}(u)$. Minty's formulation of (17) (cf. (13)) reads as follows:

$$\langle \tilde{w} + \epsilon_n J(z - u) - f_n, z - x_n^u \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u), \quad \text{for every } \tilde{w} \in \mathcal{F}_n(z), \tag{19}$$

which after some rearrangements of the terms yields

$$0 \leq \langle \bar{w} - f, z - x_n^u \rangle + \beta_n \|x_n^u - z\| + \epsilon_n \langle J(z - u), x_n^u - z \rangle,$$

where $\bar{w} \in \mathcal{F}(z)$ satisfies $\|\bar{w} - \tilde{w}\| \leq \alpha_n \kappa(\|z\|)$. The above inequality under limit $n \to \infty$ gives

$$\langle \bar{w} - f, z - \bar{x}^u \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u), \quad \text{for every } \bar{w} \in \mathcal{F}(z).$$

Using Minty's formulation once again, we ensure the existence of $w^u \in \mathcal{F}(\bar{x}^u)$ such that

$$\langle w^u - f, z - \bar{x}^u \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u),$$

confirming that $\bar{x}^u \in \mathcal{S}(u)$. Let $\tilde{x} \in \mathcal{S}(u)$ be arbitrary. Therefore, $\tilde{x} \in \mathcal{K}(u)$, and there exists $\tilde{w}_0 \in \mathcal{F}(\tilde{x})$ such that

$$\langle \tilde{w}_0 - f, z - \tilde{x} \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(u).$$

In particular,

$$\alpha_n \kappa(\|\tilde{x}\|)\|x_n^u - \tilde{x}\| + \langle \tilde{w}_n - f, x_n^u - \tilde{x} \rangle \geq 0,$$

where $\tilde{w}_n \in \mathcal{F}_n(\tilde{x})$ is such that $\|\tilde{w}_n - \tilde{w}_0\| \leq \alpha\kappa(\|\tilde{x}\|)$. By setting $z = \tilde{x}$ in (17), combining the resulting inequality with the above, we obtain

$$\epsilon_n \langle J(\tilde{x} - u), x_n^u - \tilde{x} \rangle \leq \epsilon_n \langle J(x_n^u - u), x_n^u - \tilde{x} \rangle \leq [\alpha_n \kappa(\|\tilde{x}\|) + \beta_n] \, \|x_n^u - \tilde{x}\|.$$

By using the fact that $\left\{ \dfrac{\alpha_n}{\epsilon_n}, \dfrac{\beta_n}{\epsilon_n} \right\} \to 0$ as $n \to \infty$, we obtain $\langle J(\tilde{x} - u), \bar{x}^u - \tilde{x} \rangle \leq 0$, implying

$$\langle J(\bar{x}^u - u), \tilde{x} - \bar{x}^u \rangle \geq 0 \quad \text{for every } \ \tilde{x} \in \mathcal{S}(u). \tag{20}$$

Since $\bar{x}^u$ is the unique solution of (20), the whole sequence (x_n^u) converges weakly to $\bar{x}^u$. Furthermore, the properties of the duality map (20) also confirm that (x_n^u) converges strongly to $\bar{x}^u$.

On the other hand, since (x, u) is a solution of (6), we have

$$\|x - u\|^2 \leq \|y - u\|^2 \quad \text{for every } \ y \in \mathcal{S}(u).$$

Since $\mathcal{S}(u)$ is closed and convex, the above inequality is equivalent to (20) and x is its unique solution. Therefore, $\bar{x}^u = x$, conforming that (x_n^u) converges strongly to x.

Let $(\tilde{x}_n, \tilde{u}_n)$ be a solution of (16). Then $\tilde{x}_n \in \mathcal{S}_n(\tilde{u}_n)$ and

$$\|\tilde{x}_n - \tilde{u}_n\|^2 \leq \|y - v\|^2 \quad \text{for every } (y, v) \in \mathcal{G}(\mathcal{S}_n).$$

We set $(y, v) = (x_n^u, u)$ in the above inequality and use the boundedness of (x_n^u) to ensure that $\|\tilde{x}_n - \tilde{u}_n\|$ remains bounded. Since $\mathcal{S}_n(\mathcal{C})$ is assumed to be bounded, $\tilde{u}_n$ is bounded as well. Therefore, there are subsequences $(\tilde{x}_n)$ and $(\tilde{u}_n)$ such that $\tilde{x}_n \rightharpoonup \tilde{x}$ and $\tilde{u}_n \rightharpoonup \tilde{u}$. We claim that $\tilde{x} \in \mathcal{S}(\tilde{u})$. Since $(\tilde{x}_n, \tilde{u}_n) \in \mathcal{G}(\mathcal{S}_n)$, we have $\tilde{x}_n \in \mathcal{K}(\tilde{u}_n)$, and there exists $\tilde{w}_n \in \mathcal{F}_n(\tilde{x}_n)$ such that

$$\langle \tilde{w}_n + \epsilon_n J(\tilde{x}_n - \tilde{u}_n) - f_n, z - \tilde{x}_n \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(\tilde{u}_n).$$

We will first show that $\{\tilde{w}_n\}$ is bounded. Let $\bar{x} \in \mathcal{K}(\tilde{u})$ be arbitrary. Then, there exists $z_n \in \mathcal{K}(\tilde{u}_n)$ such that $z_n \to \bar{x}$. By setting $z = z_n$ in the above inequality, we obtain

$$\langle \tilde{w}_n + \epsilon_n J(\tilde{x}_n - \tilde{u}_n) - f_n, z_n - \tilde{x}_n \rangle \geq 0.$$

Due to (A_1), there exists $w_n \in \mathcal{F}(\tilde{x}_n)$ satisfying $\|\tilde{w}_n - w_n\| \leq \alpha_n \kappa(\|\tilde{x}_n\|)$. Some rearrangements of the terms and Lemma 6 (with $\mathcal{A}(x) = \mathcal{F}(x) - f$) ensure that for sufficiently large n, we have

$$\|w_n - f\| \leq k < \infty,$$

where k is a constant, confirming $\{w_n\}$ is bounded. This further ensures that $\{\tilde{w}_n\}$ is bounded too.

We continue the pursuit of the containment $\tilde{x} \in \mathcal{S}(\tilde{u})$. Let $z \in \mathcal{K}(\tilde{u})$ be arbitrary. Then there exists a sequence (z_n) with $z_n \in \mathcal{K}(\tilde{u}_n)$ converging strongly to z such that for some $\bar{w}_n \in \mathcal{F}_n(\tilde{x}_n)$ the following inequality holds:

$$\langle \bar{w}_n - f_n + \epsilon_n J(\tilde{x}_n - \tilde{u}_n), z_n - \tilde{x}_n \rangle \geq 0.$$

Let $w_n \in \mathcal{F}(\tilde{x}_n)$ be such that $\|w_n - \bar{w}_n\| \leq \alpha_n \kappa(\|\tilde{x}_n\|)$. Then the above inequality, after some rearrangements of the terms yields

$$\begin{aligned}\langle w_z, \tilde{x}_n - z \rangle \leq{}& \langle \epsilon_n J(\tilde{x}_n - \tilde{u}_n), z_n - \tilde{x}_n \rangle + \langle f - f_n, z_n - \tilde{x}_n \rangle - \langle f, z_n - z \rangle \\ & - \langle f, z - \tilde{x}_n \rangle + \langle \bar{w}_n - w_n, z_n - \tilde{x}_n \rangle + \langle w_n, z_n - z \rangle + \langle \bar{w}_n - w_z, z - \tilde{x}_n \rangle,\end{aligned}$$

where $w_z \in \mathcal{F}(z)$. The above inequality, under the limit, implies that

$$\langle w_z, \tilde{x} - z \rangle \leq \langle f, \tilde{x} - z \rangle.$$

In view of the Minty formulation for the above, there exists $w \in \mathcal{F}(\tilde{x})$ such that

$$\langle w - f, z - \tilde{x} \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(\tilde{u}).$$

Consequently, $\tilde{x} \in \mathcal{S}(\tilde{u})$. Furthermore,

$$\begin{aligned}\|\tilde{x} - \tilde{u}\|^2 &\leq \liminf_{n \to \infty} \|\tilde{x}_n - \tilde{u}_n\|^2 \\ &\leq \limsup_{n \to \infty} \|x_n^u - u\|^2 \\ &= \|x - u\|^2,\end{aligned}$$

confirming that $(\tilde{x}, \tilde{u})$ is a solution of (6). The proof is complete. □

We conclude this section by a result ensuring the boundedness of $\mathcal{S}_n(\mathcal{C})$.

Proposition 1. *Assume that $(x_n, u_n) \in \mathcal{G}(\mathcal{S}_n)$ is such that $\{\|x_n - u_n\|\}$ is bounded. Assume that there are elements $z_n \in \mathcal{K}(u_n)$ such that $\|z_n\| \leq k_1$. Then (x_n) is bounded provided that for any sequence $(w_n, x_n) \in \mathcal{G}(\mathcal{F}_n)$, the following holds:*

$$\lim_{\|x_n\| \to \infty} \frac{\langle w_n, x_n - z_n \rangle}{\|x_n\|} = \infty$$

Proof. The proof is very similar to that of Lemma 3 and hence omitted. □

3 Applications

3.1 Quasi-hemivariational Inequalities

Let $\mathcal{B}$ be a uniformly convex Banach space with a strictly convex topological dual $\mathcal{B}*$. Let $\mathcal{C}$ be a nonempty, closed, and convex subset of $\mathcal{B}$, and let $\mathcal{K} : \mathcal{C} \rightrightarrows 2^{\mathcal{C}}$ be a set-valued map such that for every $v \in \mathcal{C}$, the set $\mathcal{K}(v)$ is nonempty, closed, and convex. Let $\mathcal{F} : \mathcal{B} \rightrightarrows 2^{\mathcal{B}^*}$ be a given multi-valued map, let $h : \mathcal{B} \to \mathbb{R}$ be a locally Lipschitz functional, and let $f \in \mathcal{B}^*$.

A genuine class of multi-valued variational and quasi-variational inequalities consists of subdifferential maps. Of particular relevance to this discussion is the Clarke's subgradient (see [9]). Given $h : \mathcal{B} \to \mathbb{R}$, locally Lipschitz near some $x \in \mathcal{B}$, the generalized derivative of h at x in direction $y \in \mathcal{B}$, denoted by $h^0(x, y)$, is defined by

$$h^0(x, y) = \limsup_{z \to x, \lambda \to 0} \lambda^{-1}[h(z + \lambda y) - h(z)],$$

where $z \in \mathcal{B}$, and λ is a positive scalar. Then the Clarke's subgradient of h at x, denoted by $\partial h(x)$, is given by

$$\partial h(x) = \{w \in \mathcal{B}^* | \quad h^0(x, y) \geq \langle w, y \rangle \ \forall \ y \in \mathcal{B}\}.$$

Let us now consider the following quasi-hemivariational inequality: find $x \in \mathcal{C}$ such that $x \in \mathcal{K}(x)$, and there exist $w \in \mathcal{F}(x)$ and $u \in \partial h(x)$ satisfying the inequality

$$\langle w + u - f, z - x \rangle \geq 0 \quad \text{for every } z \in \mathcal{K}(x). \tag{21}$$

If $\mathcal{F}$ is m-strongly monotone and ∂h is m-relaxed monotone, then the map $\mathcal{F} + \partial h$ is monotone, and our general theory can be applied to (21). A similar hemivariational inequality was studied in [23] with a single-valued $\mathcal{F}$ (see also [24]).

3.2 Inverse Problems

Assume that V is Hilbert space, $\mathcal{B}$ is a reflexive Banach space, and assume that $A \subset \mathcal{B}$ is convex and closed. We assume that $T : \mathcal{B} \times V \times V \to \mathbb{R}$ is a continuous and coercive trilinear form $T(a, u, v)$. Assume that $T(a, u, v)$ is symmetric in u, v. Finally, we assume that m is a bounded linear functional on V. Then, for any $a \in A$, it follows from the Riesz representation theorem that the following variational equation has a unique solution $u \in V$:

$$T(a, u, v) = m(v) \quad \text{for all} \quad v \in V. \tag{22}$$

We focus on the inverse problem associated with the direct problem (22) which is the following: Given some measurement of u, say z, estimate the coefficient a which together with u makes (22) true.

By the Riesz representation theorem, there is an isomorphism $E : V \to V^*$ defined by

$$(Eu)(v) = \langle u, v\rangle_V \quad \text{for all } v \in V.$$

For each $(a, u) \in A \times V$, $T(a, u, \cdot) - m(\cdot) \in V^*$. We define $e(a, u)$ to be the pre-image under E of this element:

$$\langle e(a, u), v\rangle_V = T(a, u, v) - m(v) \quad \text{for all } v \in V.$$

For a fixed $z \in V$, we consider the following minimization problem. Find $a^* \in A$ by solving

$$\min_{a\in A} J(a) = \|e(a, z)\|_V^2. \tag{23}$$

The functional J being convex, a necessary and sufficient optimality condition for (23) is a variational inequality involving the Fréchet derivative of $J(\cdot)$, defined by $\langle J'(a), b\rangle = 2\langle e(a, z), e_1(a, z)\rangle_V$, where e_1 is given by $\langle e_1(a, z), v\rangle = T(a, z, v)$ for all $v \in V$.

Since J is convex, the map J' is monotone, and hence our general theory can be applied to a perturbed analogue of the equation error approach (see [16]).

4 Concluding Remarks

In this chapter, we developed an approximation scheme for the generalized solutions of a quasi-variational inequality involving multi-valued monotone maps. The generalized solutions are defined through a parameter identification problem and they coincide with the classical solutions if the quasi-variational inequality is solvable. We have shown that the generalized solutions of a multi-valued ill-posed quasi-variational inequality are the weak cluster points of a sequence of regularized generalized solutions. As noticed, the existence criteria for the generalized solutions are quite mild, and hence a natural extension of our results would be to investigate quasi-variational inequalities with pseudo-monotone or generalized pseudo-monotone maps.

References

1. Alber, Y.I, Butnariu, D, Ryazantseva, I.: Regularization of monotone variational inequalities with Mosco approximations of the constraint sets. Set-Valued Anal. 13, 265–290 (2005)
2. Baiocchi, C., Capelo, A.: Variational and Quasivariational Inequalities. Applications to Free Boundary Problems, Wiley, New York (1984)
3. Bensoussan, A., Lions, J.L.: Nouvelle formulation de problmes de contrle impulsionnel et applications. C. R. Acad. Sci. Paris Sr. A-B 276, A1189–A1192 (1973)

4. Bruckner, G.: On abstract quasivariational inequalities. Approximation of solutions. I. Math. Nachr. 104, 209–216 (1981)
5. Bruckner, G.: On the existence of the solution of an abstract optimization problem related to a quasivariational inequality. Z. Anal. Anwendungen 3, 81–86 (1984)
6. Browder, F.E., Hess, P.: Nonlinear mappings of monotone type in Banach spaces. J. Funct. Anal. 11, 251–294 (1972)
7. Byrne, C.: A unified treatment of some iterative algorithms in signal processing and image reconstruction. Inverse Probl. 20, 103–120 (2004)
8. Censor, Y., Elfving, T., Kopf, N., Bortfeld, T.: The multiple-sets split feasibility problem and its applications for inverse problems. Inverse Probl. 21, 2071–2084 (2005)
9. Clarke, F.H.: Optimization and Nonsmooth Analysis, SIAM, Philadelphia, PA (1990)
10. Djafari Rouhani, B., Khan, A.A.: On the embedding of variational inequalities. Proc. Am. Math. Soc. 131, 3861–3871 (2003)
11. Giannessi, F.: Embedding variational inequalities and their generalizations into a separation scheme. J. Inequal. Appl. 1(2), 139–147 (1997)
12. Giannessi, F., Khan, A.A.: Regularization of non-coercive quasi variational inequalities. Control Cybernet 29, 91–110 (2000)
13. Gockenbach, M.S., Khan, A.A.: Identification of Lamé parameters in linear elasticity: a fixed point approach. J. Indust. Manag. Optim. 1, 487–497 (2005a)
14. Gockenbach, M.S., Khan, A.A.: A convex objective functional for elliptic inverse problems. In: K.M. Furuti, M.Z. Nashed, A.H. Siddiqi (Eds.), Mathematical Models and Methods for Real World Systems (pp. 389–419), Chapman & Hall/CRC (2005b)
15. Gockenbach, M.S., Khan, A.A.: An abstract framework for elliptic inverse problems. Part 1: an output least-squares approach. Math. Mech. Solids 12, 259–276 (2007)
16. Gockenbach, M.S., Jadamba, B., Khan, A.A.: Equation error approach for elliptic inverse problems with an application to the identification of Lamé parameters. Inverse Probl. Sci. Eng. 16, 349–367 (2008)
17. Gwinner, J.: Note on pseudomonotone functions, regularization, and relaxed coerciveness, Nonlinear Anal. 30, 4217–4227 (1997)
18. Isac, G.: Tikhonov regularization and the complementarity problem in Hilbert spaces. J. Math. Anal. Appl. 174, 53–66 (1993)
19. Isac, G., Bulavski, V.A., Kalashnikov, V.V.: Complementarity, Equilibrium, Efficiency and Economics, Kluwer, Dordrecht (2002)
20. Konnov, I.V., Ali, M.S.S., Mazurkevich, E.O.: Regularization of nonmonotone varoiational inequalities. Appl. Math. Optim. 53, 311–330 (2006)
21. Kravchuk, A., Neittaanmki, P.J.: Variational and Quasi-variational Inequalities in Mechanics Springer, Dordrecht (2007)
22. Liskovets, O.L.: Regularization of mixed incorrect variational inequalities of monotone type. Soviet J. Numer. Anal. Math. Model. 6, 107–119 (1991)
23. Liu, Z.: Generalized quasi variational hemi-variational inequalities. Appl. Math. Lett. 17, 741–745 (2004)
24. Liu, Z.: Browder-Tikhonov regularization of non-coercive evolution hemivariational inequalities. Inverse Probl. 21, 13–20 (2005)

25. Liu, F., Nashed, M.Z.: Regularization of nonlinear ill-posed variational inequalities and convergence rates. Set-Valued Anal. 6, 313–344 (1998)
26. Mosco, U.: Convergence of convex sets and of solutions of variational inequalities. Adv. Math. 3, 512–585 (1969)
27. Mosco, U.: Implicit variational problems and quasi variational inequalities. In: J.P. Gossez, E.J. Lami Dozo, J. Mawhin and L. Waelbroeck (Eds.), Nonlinear Operators and the Calculus of Variations, Lecture Notes in Mathematics. Vol. 543, (pp. 83–156), Springer, Berlin (1976)
28. Noor, M.A.: Some iterative algorithms for solving split feasibility problems (2008) (to appear)
29. Zeidler, E.: Nonlinear Functional Analysis and Its Applications. II/B. Nonlinear Monotone Operators, Springer, New York (1990)
30. Zhao, J., Yang, Q.: Several solution methods for the split feasibility problem. Inverse Probl. 21, 1791–1799 (2005)

Optimal Feedback Control for Stochastic Impulsive Linear Systems Subject to Poisson Processes

Zhi Guo Feng[1] and Kok Lay Teo[2]

[1] College of Mathematics and Computer Science, Chongqing Normal University, Chongqing, 400047, People's Republic of China
z.feng.scholar@gmail.com
[2] Department of Mathematics and Statistics, Curtin University of Technology, Perth, 6845, Australia
k.l.teo@curtin.edu.au

Summary. This chapter considers a class of optimal feedback control problems, where its dynamical system is described by stochastic linear systems subject to Poisson processes and with state jumps. We show that this stochastic impulsive optimal parameter selection problem is equivalent to a deterministic impulsive optimal parameter selection problem, where the times at which the jumps occurred as well as their heights are decision variables. Then, by introducing a time scaling transform, we show that this deterministic impulsive optimal parameter selection problem is transformed into an equivalent deterministic impulsive optimal parameter selection problem with fixed jump times. For the numerical computation, we derive the gradient formulae of the cost function and the constraint functions. On this basis, an efficient computational method is developed and an example is solved for illustration.

Key words: stochastic impulsive optimal parameter selection problem, Poisson process, time scaling transformation

1 Introduction

A stochastic differential equation is a differential equation of which at least one term is a stochastic process so that the solution of a stochastic differential equation is also a stochastic process. It is a powerful mathematical tool which can be applied to many real-life problems in nature, science, and engineering. The theory of Ito stochastic differential equations driven by Wiener processes and Poisson processes and their many important applications (such as filtering problems) can be found in [3–8], [11] and [13–15]. In [3–7], some sensor scheduling problems are considered, where the underlying dynamic system is

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_13,

governed by a system of linear Ito stochastic differential equations driven by Wiener processes. In [8, 11, 14], a class of stochastic optimal control problems is considered, where the dynamical systems are described by Ito stochastic differential equations driven by Wiener processes. In [13, 15], a class of optimal control problems described by linear Ito stochastic differential equations driven by Poisson processes is considered and studied. It is shown that this class of stochastic optimal control problems is equivalent to a class of deterministic optimal control problems. However, numerical solution methods available in the literature for solving such deterministic optimal control problems are only applicable to cases with low dimension.

The optimal parameter selection problems occur in many dynamical optimization models where the controls are restricted to be constant functions of time. It plays a fundamental role in the numerical computation of optimal control problems. To be more specific, after the application of the control parameterization (see [16]) or control parametrization time scaling technique (see [12]), all optimal control problems are essentially reduced to optimal parameter selection problems. Thus, the solvability of optimal parameter selection problem is crucially important for generating numerical solution methods to many complex optimal control problems. In [1] and [14], respective necessary conditions for optimality are derived for deterministic and stochastic optimal parameter selection problems. Computational methods for solving deterministic optimal parameter selection problems are reported in [2].

Stochastic model generally assumes smoothness and continuity of the phenomena of interest. However, some phenomena may experience sudden or sharp changes. Many natural and man-made systems do exhibit the phenomenon of jumps occurring at various time points along their trajectories. Examples include drug administration in cancer chemotherapy, insulin injection, and native forest ecosystems management, just to name a few. In this chapter, we consider an optimal feedback control problem, where the system dynamics are described by linear Ito stochastic differential equations driven by Poisson process, and the state jumps are to occur at various time points.

The rest of the chapter is organized as follows. In Section 2, we formulate the optimal feedback control problem as a stochastic impulsive optimal parameter selection problem. In Section 3, we show that this problem is equivalent to a deterministic impulsive optimal parameter selection problem. In Section 4, a time scaling transform is applied to map the variable jump times into pre-fixed jump times in a new timescale. In Section 5, we derive the gradient formulae of the cost function and the constraint functions. With the information on these gradients, the problem can be solved as an optimization problem by using a gradient-based algorithm. For illustration, an example is solved using the proposed method in Section 6.

2 Problem Statement

Consider a system governed by the following Ito stochastic differential equation over a finite time interval $(0, T]$:

$$d\boldsymbol{x}(t) = \boldsymbol{A}(t)\boldsymbol{x}(t)dt + \boldsymbol{B}(t)d\boldsymbol{u}(t) + \boldsymbol{D}(t)d\boldsymbol{N}(t) \tag{1a}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}^0 \tag{1b}$$

$$\boldsymbol{x}(\tau_i^+) = \boldsymbol{J}^i\boldsymbol{x}(\tau_i^-) + \boldsymbol{\Delta}_i + \boldsymbol{\gamma}^i, \quad i = 1, \ldots, m, \tag{1c}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$, $\boldsymbol{A}(t) \in \mathbb{R}^{n\times n}$, $\boldsymbol{B}(t) \in \mathbb{R}^{n\times r}$, while $\boldsymbol{u}(t) \in \mathbb{R}^r$ is a control function which is of bounded variation and hence $d\boldsymbol{u}(t)$ is a measure, $\boldsymbol{D}(t) \in \mathbb{R}^{n\times d}$, and the noise $\boldsymbol{N}(t) \in \mathbb{R}^d$ is a d-dimensional Poisson process with mean intensity $\boldsymbol{\lambda}(t)$. The initial condition $\boldsymbol{x}^0 \in \mathbb{R}^n$ is either a deterministic or a Gaussian vector. In the case when $\boldsymbol{x}^0$ is a Gaussian vector, let $\bar{\boldsymbol{x}}^0$ and $\boldsymbol{P}^0$ be its mean and covariance, respectively. Equation (1c) is condition on the state jumps, where $\boldsymbol{J}^i \in \mathbb{R}^{n\times n}$, $i = 1, \ldots, m$ are given coefficient matrices, $\tau_1, \ldots, \tau_m$, are the time points at which the state jumps are occurred, $\boldsymbol{\Delta}_i$, $i = 1, \ldots, m$, are Gaussian vectors with mean $\boldsymbol{0}$ and covariance matrices $\boldsymbol{P}^i$, $i = 1, \ldots, m$, and $\boldsymbol{\gamma}^i = [\gamma_1^i, \ldots, \gamma_n^i]^\intercal$, $i = 1, \ldots, m$, are the magnitude vectors of the jumps. Let $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_m]^\intercal$.

Along with (1a, 1b, 1c), suppose that we have an observation system described by

$$d\boldsymbol{y}(t) = \boldsymbol{G}(t)\boldsymbol{x}(t)\ dt + \boldsymbol{D}^0(t)(d\boldsymbol{N}^0(t) - \boldsymbol{\lambda}^0(t)\ dt), \tag{2a}$$

$$\boldsymbol{y}(0) = \boldsymbol{0}, \tag{2b}$$

where $\boldsymbol{y}(t) \in \mathbb{R}^p$, $\boldsymbol{G}(t) \in \mathbb{R}^{p\times n}$, $\boldsymbol{D}^0(t) \in \mathbb{R}^{p\times q}$, while $\boldsymbol{N}^0(t) \in \mathbb{R}^q$ is a q-dimensional Poisson process with mean intensity $\boldsymbol{\lambda}^0(t)$. The initial condition (2b) means that no information is available at $t = 0$.

We assume that the following conditions are satisfied.

(i) $\boldsymbol{A}(t) \in \mathbb{R}^{n\times n}$, $\boldsymbol{B}(t) \in \mathbb{R}^{n\times r}$, and $\boldsymbol{D}(t) \in \mathbb{R}^{n\times d}$ are continuous on $[0, T]$.
(ii) The Poisson processes $\boldsymbol{N}(t)$, $\boldsymbol{N}^0(t)$ and the random vectors $\boldsymbol{x}^0$, $\boldsymbol{\Delta}_i$, $i = 1, \ldots, m$, are mutually independent.
(iii) All the components of the mean intensities, $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\lambda}^0(t)$, are nonnegative and bounded measurable functions.

Suppose that the control function $\boldsymbol{u}$ is such that the corresponding measure $d\boldsymbol{u}(t)$ is of the form as given below:

$$d\boldsymbol{u}(t) = \boldsymbol{K}\boldsymbol{y}(t)\ dt + \hat{\boldsymbol{K}}d\boldsymbol{y}(t) - \boldsymbol{C}(t)\boldsymbol{D}(t)\boldsymbol{\lambda}(t)\ dt, \tag{3}$$

where $\boldsymbol{K}$, $\hat{\boldsymbol{K}} \in \mathbb{R}^{r\times p}$ are constant matrices yet to be determined, and

$$\boldsymbol{B}(t)\boldsymbol{C}(t)\boldsymbol{D}(t)\boldsymbol{\lambda}(t) = \boldsymbol{D}(t)\boldsymbol{\lambda}(t) \tag{4}$$

provided such a matrix $\boldsymbol{C}(t)$ exists. In fact, if $\boldsymbol{B}(t)$ has rank n and $n \le r$, then $\boldsymbol{C}(t)$ is just the right inverse of $\boldsymbol{B}(t)$.

Substituting (3) and (2) into (1a), we obtain

$$\begin{aligned}\mathrm{d}\boldsymbol{x}(t) =&(\boldsymbol{A}(t)+\boldsymbol{B}(t)\hat{\boldsymbol{K}}\boldsymbol{G}(t))\boldsymbol{x}(t)\ dt+\boldsymbol{B}(t)\boldsymbol{K}\boldsymbol{y}(t)\ dt\\ &+\boldsymbol{B}(t)\hat{\boldsymbol{K}}\boldsymbol{D}^0(t)(d\boldsymbol{N}^0(t)-\boldsymbol{\lambda}^0(t)\ dt)+\boldsymbol{D}(t)(d\boldsymbol{N}(t)-\boldsymbol{\lambda}(t)\ dt). \quad (5)\end{aligned}$$

Define

$$\boldsymbol{\xi}(t)=\begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)\end{bmatrix}.$$

Then, the system dynamics (5) together with the observation dynamics (2) can be jointly written as

$$d\boldsymbol{\xi}(t)=\tilde{\boldsymbol{A}}(t,\boldsymbol{\kappa})\boldsymbol{\xi}(t)\ dt+\tilde{\boldsymbol{D}}(t,\boldsymbol{\kappa})d\tilde{\boldsymbol{M}}(t), \quad (6a)$$

$$\boldsymbol{\xi}(0)=\boldsymbol{\xi}^0 \quad (6b)$$

$$\boldsymbol{\xi}(\tau_i^+)=\tilde{\boldsymbol{J}}^i\boldsymbol{\xi}(\tau_i^-)+\tilde{\boldsymbol{\Delta}}_i+\tilde{\boldsymbol{\gamma}}^i,\quad i=1,\ldots,m, \quad (6c)$$

where the vector $\boldsymbol{\kappa}\in\mathbb{R}^{2rp}$ is defined by

$$\boldsymbol{\kappa}=[K_{11},\ldots,K_{1p},\ldots,K_{r1},\ldots,K_{rp},\hat{K}_{11},\ldots,\hat{K}_{1p},\ldots,\hat{K}_{r1},\ldots,\hat{K}_{rp}]^\intercal,$$

$$\tilde{\boldsymbol{A}}(t,\boldsymbol{\kappa})=\begin{bmatrix}\boldsymbol{A}(t)+\boldsymbol{B}(t)\hat{\boldsymbol{K}}\boldsymbol{G}(t) & \boldsymbol{B}(t)\boldsymbol{K}\\ \boldsymbol{G}(t) & \boldsymbol{0}\end{bmatrix},\ \tilde{\boldsymbol{D}}(t,\boldsymbol{\kappa})=\begin{bmatrix}\boldsymbol{D}(t) & \boldsymbol{B}(t)\hat{\boldsymbol{K}}\boldsymbol{D}^0(t)\\ \boldsymbol{0} & \boldsymbol{D}^0(t)\end{bmatrix},$$

$$d\tilde{\boldsymbol{M}}(t)=\begin{bmatrix}d\boldsymbol{N}(t)-\boldsymbol{\lambda}(t)dt\\ d\boldsymbol{N}^0(t)-\boldsymbol{\lambda}^0(t)dt\end{bmatrix},\quad \boldsymbol{\xi}^0=\begin{bmatrix}\boldsymbol{x}^0 & \boldsymbol{0}\end{bmatrix}^\intercal,\quad \tilde{\boldsymbol{J}}^i=\begin{bmatrix}\boldsymbol{J}^i & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{p\times p}\end{bmatrix},$$

$$\tilde{\boldsymbol{\Delta}}_i=\begin{bmatrix}\boldsymbol{\Delta}_i & \boldsymbol{0}\end{bmatrix}^\intercal,\quad \tilde{\boldsymbol{\gamma}}^i=\begin{bmatrix}\boldsymbol{\gamma}^i & \boldsymbol{0}\end{bmatrix}^\intercal.$$

Note that $\tilde{\boldsymbol{M}}$ is a vector of zero-mean martingales.

We assume the vector $\boldsymbol{\kappa}$ is to be chosen from the set $\mathbb{K}$ defined by

$$\begin{aligned}\mathbb{K}&=\{\boldsymbol{\kappa}=[\kappa_1,\ldots,\kappa_{2rp}]^\intercal\in\mathbb{R}^{2rp}:\underline{\boldsymbol{\beta}}\le\boldsymbol{\kappa}\le\overline{\boldsymbol{\beta}}\}\\ &=\{\boldsymbol{\kappa}=[\kappa_1,\ldots,\kappa_{2rp}]^\intercal\in\mathbb{R}^{2rp}:\underline{\beta}_i\le\kappa_i\le\overline{\beta}_i,\ i=1,\ldots,2rp\}, \quad (7)\end{aligned}$$

where $\underline{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{\beta}}$ are given vectors in $\mathbb{R}^{2rp}$.

For the jump time vector $\boldsymbol{\tau}=[\tau_1,...,\tau_m]^\intercal$, it is assumed, without loss of generality, that

$$0<\tau_1<\cdots<\tau_m<T. \quad (8)$$

Let $\boldsymbol{\mathcal{T}}$ be the set of all those $\boldsymbol{\tau}=[\tau_1,\ldots,\tau_m]^\intercal$ which satisfy (4). For brevity in notation, we denote $\tau_0=0$ and $\tau_{m+1}=T$.

Let $\boldsymbol{\Gamma}$ be the set of all those magnitude vectors $\boldsymbol{\gamma}=[(\boldsymbol{\gamma}^1)^\intercal,\ldots,(\boldsymbol{\gamma}^m)^\intercal]^\intercal$ such that

$$\underline{\gamma}_j^i\le\gamma_j^i\le\overline{\gamma}_j^i,\ i=1,\ldots,m;\ \ j=1,\ldots,n. \quad (9)$$

The cost function to be minimized is given by

$$g_0(\boldsymbol{\delta}) = \psi(\boldsymbol{\gamma}) + \mathcal{E}\{(\boldsymbol{\xi}(T))^\intercal \boldsymbol{Q}_5(\boldsymbol{\delta})\boldsymbol{\xi}(T) + (\boldsymbol{Q}_4(\boldsymbol{\delta}))^\intercal \boldsymbol{\xi}(T) + \boldsymbol{Q}_3(\boldsymbol{\delta}) + \sum_{i=1}^{m+1} \int_{\tau_{i-1}}^{\tau_i} [(\boldsymbol{\xi}(t))^\intercal \boldsymbol{Q}_2(t,\boldsymbol{\delta})\boldsymbol{\xi}(t) + (\boldsymbol{Q}_1(t,\boldsymbol{\delta}))^\intercal \boldsymbol{\xi}(t) + \boldsymbol{Q}_0(t,\boldsymbol{\delta})]\, dt\}, \tag{10}$$

where $\boldsymbol{\delta} = (\boldsymbol{\kappa}, \boldsymbol{\tau}, \boldsymbol{\gamma})$, $\psi(\boldsymbol{\gamma})$ is a penalty term to prevent high jumps, and $\boldsymbol{Q}_5(\boldsymbol{\delta}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q}_2(t,\boldsymbol{\delta}) \in \mathbb{R}^{n\times n}$ are positive semi-definite matrices which are continuously differentiable with respect to their respective arguments, while $\boldsymbol{Q}_4(\boldsymbol{\delta})$ and $\boldsymbol{Q}_1(t,\boldsymbol{\delta})$ (respectively, $\boldsymbol{Q}_3(\boldsymbol{\delta})$ and $\boldsymbol{Q}_0(t,\boldsymbol{\delta})$) are n-dimensional vector-valued functions (respectively, real-valued functions) which are also continuously differentiable with respect to their respective arguments.

Then, we formulate the problem as

Problem 1. Given the system (1), the observation channel (2), and the proposed control dynamics of the form (3), find a feasible parameter vector $\boldsymbol{\delta} \in \mathbb{K} \times \boldsymbol{\mathcal{T}} \times \boldsymbol{\Gamma}$ such that the cost function (10) is minimized, subject to the constraints

$$\begin{aligned} g_i(\boldsymbol{\delta}) =& \mathcal{E}\{(\boldsymbol{\xi}(T))^\intercal \boldsymbol{S}_{i5}(\boldsymbol{\delta})\boldsymbol{\xi}(T) + (\boldsymbol{S}_{i4}(\boldsymbol{\delta}))^\intercal \boldsymbol{\xi}(T) + \boldsymbol{S}_{i3}(\boldsymbol{\delta}) \\ &+ \sum_{j=1}^{m+1} \int_{\tau_{j-1}}^{\tau_j} [(\boldsymbol{\xi}(t))^\intercal \boldsymbol{S}_{i2}(t,\boldsymbol{\delta})\boldsymbol{\xi}(t) + (\boldsymbol{S}_{i1}(t,\boldsymbol{\delta}))^\intercal \boldsymbol{\xi}(t) + \boldsymbol{S}_{i0}(t,\boldsymbol{\delta})]\, dt\} \\ \leq& 0, \end{aligned} \tag{11}$$

for $i = 1, \ldots, n_0$, where, for each i, $\boldsymbol{S}_{i5}(\boldsymbol{\delta}) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{S}_{i2}(t,\boldsymbol{\delta}) \in \mathbb{R}^{n\times n}$ are positive semi-definite matrices which are continuously differentiable with respect to their respective arguments, while $\boldsymbol{S}_{i4}(\boldsymbol{\delta})$ and $\boldsymbol{S}_{i1}(t,\boldsymbol{\delta})$ (respectively, $\boldsymbol{S}_{i3}(\boldsymbol{\delta})$ and $\boldsymbol{S}_{i0}(t,\boldsymbol{\delta})$) are n-dimensional vector-valued functions (respectively, real-valued functions) which are also continuously differentiable with respect to their respective arguments.

Problem 1 is a stochastic impulsive optimal parameter selection problem. We shall show that it is equivalent to a deterministic optimal parameter selection problem, and then a numerical computational method will be developed for solving this problem.

3 Deterministic Transformation

For each $\boldsymbol{\delta}$, it is clear from (6a) that the solution of system (6) is given, for $t \in (\tau_{i-1}, \tau_i)$ with $i = 1, \ldots, m$, by

$$\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) = \tilde{\boldsymbol{\Phi}}(t, \tau_i \mid \boldsymbol{\kappa})\boldsymbol{\xi}(\tau_{i-1}^+ \mid \boldsymbol{\delta}) + \int_{\tau_{i-1}}^{t} \tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa})\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa}) d\tilde{\boldsymbol{M}}(s), \tag{12}$$

where $\tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa}) \in \mathbb{R}^{(n+p)\times(n+p)}$ is the principal solution matrix of the homogeneous system

$$\frac{\partial \tilde{\boldsymbol{\Phi}}(t, s)}{\partial t} = \tilde{\boldsymbol{A}}(t, \boldsymbol{\kappa})\tilde{\boldsymbol{\Phi}}(t, s), \quad 0 \le s \le t < \infty \tag{12a}$$

$$\tilde{\boldsymbol{\Phi}}(t, t) = \boldsymbol{I}_{(n+p)\times(n+p)}, \tag{12b}$$

where $\boldsymbol{I}_{(n+p)\times(n+p)}$ denotes the identity matrix.

Define the mean of the process $\boldsymbol{\xi}$ as

$$\boldsymbol{\mu}(t \mid \boldsymbol{\delta}) = \mathcal{E}\{\boldsymbol{\xi}(t \mid \boldsymbol{\delta})\}.$$

It is given in the following theorem.

Theorem 1. *For each $\boldsymbol{\delta}$, the mean behavior of the corresponding solution of the coupled system (6) is determined by*

$$\frac{\mathrm{d}\boldsymbol{\mu}(t)}{dt} = \tilde{\boldsymbol{A}}(t, \boldsymbol{\kappa})\boldsymbol{\mu}(t) \tag{13a}$$

$$\boldsymbol{\mu}(0) = [\bar{\boldsymbol{x}}^0, \boldsymbol{0}]^\intercal = \boldsymbol{\mu}^0 \tag{13b}$$

$$\boldsymbol{\mu}(\tau_i^+) = \tilde{\boldsymbol{J}}^i \boldsymbol{\mu}(\tau_i^-) + \tilde{\boldsymbol{\gamma}}^i, \qquad i = 1, \ldots, m. \tag{13c}$$

Proof. Equation (13a) is derived by taking the expectation of (12) and applying (12a, 12b). (13b) and (13c) are derived by taking the expectation of (6b) and (6c), respectively.

Define the $(n + p) \times (n + p)$ covariance matrix of the process $\boldsymbol{\xi}$ as

$$\boldsymbol{\Psi}(t \mid \boldsymbol{\delta}) = \mathcal{E}\{(\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(t \mid \boldsymbol{\delta}))(\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(t \mid \boldsymbol{\delta})^\intercal\}.$$

Then, we have

Theorem 2. *For each $\boldsymbol{\delta}$, the covariance matrix of the corresponding solution of the coupled system (6) is determined by*

$$\frac{d\boldsymbol{\Psi}(t)}{dt} = \tilde{\boldsymbol{A}}(t, \boldsymbol{\kappa})\boldsymbol{\Psi}(t) + \boldsymbol{\Psi}(t)[\tilde{\boldsymbol{A}}(t, \boldsymbol{\kappa})]^\intercal + \tilde{\boldsymbol{D}}(t, \boldsymbol{\kappa})\tilde{\boldsymbol{\Lambda}}(t)[\tilde{\boldsymbol{D}}(t, \boldsymbol{\kappa})]^\intercal \tag{14a}$$

$$\boldsymbol{\Psi}(0) = \boldsymbol{\Psi}^0 \tag{14b}$$

$$\boldsymbol{\Psi}(\tau_i^+) = \tilde{\boldsymbol{J}}^i \boldsymbol{\Psi}(\tau_i^-)(\tilde{\boldsymbol{J}}^i)^\intercal + \tilde{\boldsymbol{P}}^i, \qquad i = 1, \ldots, m, \tag{14c}$$

where

$$\tilde{\boldsymbol{\Lambda}}(t) = \begin{bmatrix} \boldsymbol{\Lambda}(t) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}^0(t) \end{bmatrix}, \quad \boldsymbol{\Psi}^0 = \begin{bmatrix} \boldsymbol{P}^0 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \quad \tilde{\boldsymbol{P}}^i = \begin{bmatrix} \boldsymbol{P}^i & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},$$

with $\boldsymbol{\Lambda}(t) = \mathrm{diag}\ (\lambda_1(t), \ldots, \lambda_d(t))$ *and* $\boldsymbol{\Lambda}^0(t) = \mathrm{diag}\ (\lambda_1^0(t), \ldots, \lambda_q^0(t))$.

Proof. From (12), it follows that for each $t \in [\tau_{i-1}, \tau_i]$,

$$\begin{aligned}&\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(t \mid \boldsymbol{\delta})\\ =&\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa})(\boldsymbol{\xi}(\tau_{i-1}) - \boldsymbol{\mu}(\tau_{i-1})) + \int_{\tau_{i-1}}^{t} \tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa})\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa})d\tilde{\boldsymbol{M}}(s), \end{aligned} \tag{15}$$

where the second term on the right-hand side, which is a stochastic integral with respect to the martingale $\tilde{\boldsymbol{M}}$, is itself a martingale. Now, for any $\boldsymbol{\varphi} \in \mathbb{R}^{n+p}$, define

$$\boldsymbol{\varphi}^{\intercal}\boldsymbol{\Psi}(t \mid \boldsymbol{\delta})\boldsymbol{\varphi} = \mathcal{E}\left\{[\boldsymbol{\varphi}^{\intercal}(\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(t \mid \boldsymbol{\delta}))]^2\right\}. \tag{16}$$

From (15), it follows that

$$\begin{aligned}[\boldsymbol{\varphi}^{\intercal}(\boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(t \mid \boldsymbol{\delta}))]^2 =&\Big[\boldsymbol{\varphi}^{\intercal}\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa})(\boldsymbol{\xi}(\tau_{i-1}) - \boldsymbol{\mu}(\tau_{i-1}))\\ &+ \int_{\tau_{i-1}}^{t} \boldsymbol{\varphi}^{\intercal}\tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa})\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa})d\tilde{\boldsymbol{M}}(s)\Big]^2. \end{aligned} \tag{17}$$

Taking the expectation of both sides and then using the quadratic variation of the martingale $\tilde{\boldsymbol{M}}$ given by

$$\mathcal{E}\left\{\int_{\tau_{i-1}}^{t} \boldsymbol{a}^{\intercal}d\tilde{\boldsymbol{M}}(s)\right\}^2 = \int_{\tau_{i-1}}^{t} \boldsymbol{a}^{\intercal}\tilde{\boldsymbol{\Lambda}}\boldsymbol{a}ds, \quad \boldsymbol{a} \in \mathbb{R}^{d+q},$$

we obtain

$$\begin{aligned}\boldsymbol{\varphi}^{\intercal}\boldsymbol{\Psi}(t \mid \boldsymbol{\delta})\boldsymbol{\varphi} =& \boldsymbol{\varphi}^{\intercal}\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa})\boldsymbol{\Psi}(\tau_{i-1} \mid \boldsymbol{\delta})(\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa}))^{\intercal}\boldsymbol{\varphi}\\ &+ \int_{\tau_{i-1}}^{t} \boldsymbol{\varphi}^{\intercal}\tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa})\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa})\tilde{\boldsymbol{\Lambda}}(s)(\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa}))^{\intercal}(\tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa}))^{\intercal}\boldsymbol{\varphi}ds. \end{aligned} \tag{18}$$

Since (18) is valid for any $\boldsymbol{\varphi} \in \mathbb{R}^{n+p}$, it follows that for each $t \in [\tau_{i-1}, \tau_i]$,

$$\begin{aligned}\boldsymbol{\Psi}(t \mid \boldsymbol{\delta}) =&\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa})\boldsymbol{\Psi}(\tau_{i-1} \mid \boldsymbol{\delta})(\tilde{\boldsymbol{\Phi}}(t, \tau_{i-1} \mid \boldsymbol{\kappa}))^{\intercal}\\ &+ \int_{\tau_{i-1}}^{t} \tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa})\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa})\tilde{\boldsymbol{\Lambda}}(s)(\tilde{\boldsymbol{D}}(s, \boldsymbol{\kappa}))^{\intercal}(\tilde{\boldsymbol{\Phi}}(t, s \mid \boldsymbol{\kappa}))^{\intercal}ds. \end{aligned} \tag{19}$$

From (6b), it follows that

$$\boldsymbol{\Psi}(0) = \boldsymbol{\Psi}^0. \tag{20}$$

From (6c), it follows that for each $i = 1, \ldots, m$,

$$\begin{aligned}
&\boldsymbol{\Psi}(\tau_i^+ \mid \boldsymbol{\delta}) = \mathcal{E}\{[\boldsymbol{\xi}(\tau_i^+ \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(\tau_i^+ \mid \boldsymbol{\delta})][\boldsymbol{\xi}(\tau_i^+ \mid \boldsymbol{\delta}) - \boldsymbol{\mu}(\tau_i^+ \mid \boldsymbol{\delta})]^\intercal\} \\
&= \mathcal{E}\{[\tilde{\boldsymbol{J}}^i \boldsymbol{\xi}(\tau_i^- \mid \boldsymbol{\delta}) - \tilde{\boldsymbol{J}}^i \boldsymbol{\mu}(\tau_i^- \mid \boldsymbol{\delta}) + \tilde{\boldsymbol{\Delta}}_i][\tilde{\boldsymbol{J}}^i \boldsymbol{\xi}(\tau_i^- \mid \boldsymbol{\delta}) - \tilde{\boldsymbol{J}}^i \boldsymbol{\mu}(t_i^- \mid \boldsymbol{\delta}) + \tilde{\boldsymbol{\Delta}}_i]^\intercal\} \\
&= \mathcal{E}\{[\tilde{\boldsymbol{J}}^i \boldsymbol{\xi}(\tau_i^- \mid \boldsymbol{\delta}) - \tilde{\boldsymbol{J}}^i \boldsymbol{\mu}(\tau_i^- \mid \boldsymbol{\delta})][\tilde{\boldsymbol{J}}^i \boldsymbol{\xi}(\tau_i^- \mid \boldsymbol{\delta}) - \tilde{\boldsymbol{J}}^i \boldsymbol{\mu}(\tau_i^- \mid \boldsymbol{\delta})]^\intercal\} + \mathcal{E}\{\tilde{\boldsymbol{\Delta}}_i \tilde{\boldsymbol{\Delta}}_i^\intercal\} \\
&= \tilde{\boldsymbol{J}}^i \boldsymbol{\Psi}(\tau_i^- \mid \boldsymbol{\delta})(\tilde{\boldsymbol{J}}^i)^\intercal + \tilde{\boldsymbol{P}}^i. \qquad (21)
\end{aligned}$$

Now, by differentiating (19) and then using (20) and (21), we obtain (14a), (14b), (14c). Thus, the proof is complete. □

Consider the cost function (10). Since $\mathcal{E}\{\boldsymbol{\xi}(t)\boldsymbol{\xi}^\intercal(t)\} = \boldsymbol{\Psi}(t) + \boldsymbol{\mu}(t)(\boldsymbol{\mu}(t))^\intercal$ and

$$\begin{aligned}
\mathcal{E}\{\boldsymbol{\xi}^\intercal(t)\boldsymbol{Q}(t)\boldsymbol{\xi}(t)\} &= \mathcal{E}\{\mathrm{trace}(\boldsymbol{\xi}^\intercal(t)\boldsymbol{Q}(t)\boldsymbol{\xi}(t))\} = \mathcal{E}\{\mathrm{trace}(\boldsymbol{Q}(t)\boldsymbol{\xi}(t)\boldsymbol{\xi}^\intercal(t))\} \\
&= \mathrm{trace}\{\boldsymbol{Q}(t)(\boldsymbol{\Psi}(t) + \boldsymbol{\mu}(t)(\boldsymbol{\mu}(t))^\intercal)\},
\end{aligned}$$

it follows that (10) is equivalent to

$$\begin{aligned}
g_0(\boldsymbol{\delta}) =& \psi(\boldsymbol{\gamma}) + \mathrm{trace}\{\boldsymbol{Q}_5(\boldsymbol{\delta})(\boldsymbol{\Psi}(T) + \boldsymbol{\mu}(T)(\boldsymbol{\mu}(T))^\intercal)\} + (\boldsymbol{Q}_4(\boldsymbol{\delta}))^\intercal \boldsymbol{\mu}(T) \\
&+ \boldsymbol{Q}_3(\boldsymbol{\delta}) + \sum_{i=1}^{m+1} \int_{\tau_{i-1}}^{\tau_i} [\mathrm{trace}\{\boldsymbol{Q}_2(t,\boldsymbol{\delta})(\boldsymbol{\Psi}(t) + \boldsymbol{\mu}(t)(\boldsymbol{\mu}(t))^\intercal)\} \\
&+ (\boldsymbol{Q}_1(t,\boldsymbol{\delta}))^\intercal \boldsymbol{\mu}(t) + \boldsymbol{Q}_0(t,\boldsymbol{\delta})]\, dt. \qquad (22)
\end{aligned}$$

By the same token, we can show that, for each $i = 1, \ldots, n_0$, the constraint (11) is equivalent to

$$\begin{aligned}
g_i(\boldsymbol{\delta}) =& \mathrm{trace}\{\boldsymbol{S}_{i5}(\boldsymbol{\delta})(\boldsymbol{\Psi}(T) + \boldsymbol{\mu}(T)(\boldsymbol{\mu}(T))^\intercal\} + (\boldsymbol{S}_{i4}(\boldsymbol{\delta}))^\intercal \boldsymbol{\mu}(T) + \boldsymbol{S}_{i3}(\boldsymbol{\delta}) \\
&+ \sum_{j=1}^{m+1} \int_{\tau_{j-1}}^{\tau_j} [\mathrm{trace}\{\boldsymbol{S}_{i2}(t,\boldsymbol{\delta})(\boldsymbol{\Psi}(t) + \boldsymbol{\mu}(t)(\boldsymbol{\mu}(t))^\intercal\} \\
&+ (\boldsymbol{S}_{i1}(t,\boldsymbol{\delta}))^\intercal \boldsymbol{\mu}(t) + \boldsymbol{S}_{i0}(t,\boldsymbol{\delta})]\, dt \le 0. \qquad (23)
\end{aligned}$$

Now, we have transformed the stochastic optimal parameter selection problem into a deterministic optimal parameter selection problem defined as follows.

Problem 2. Given the dynamical system (13a), (13b), (13c) and (14a), (14b), (14c), find a parameter $\boldsymbol{\delta} \in \mathbb{K} \times \boldsymbol{\mathcal{T}} \times \boldsymbol{\Gamma}$, such that the cost function (22) is minimized, subject to the constraints (23).

We now summarize the results obtained so far below as a theorem.

Theorem 3. *Problem 1 is equivalent to Problem 2.*

4 Time Scaling Transformation

Problem 2 is a deterministic impulsive optimal parameter selection problem, where the jump times are decision variables to be determined optimally. This will encounter difficulty in numerical calculation when solving the impulsive dynamical system with varying jump times. In this section, we will use a time scaling transform reported in [12] to map these variables jump times into fixed knots in a new timescale.

We consider a new time variable s which varies from 0 to $m+1$. We re-scale $t \in [0,T]$ into $s \in [0, m+1]$. The transformation from $t \in [0,T]$ to $s \in [0, m+1]$ is defined by the differential equation

$$dt(s)/ds = \upsilon(s) = \sum_{i=1}^{m+1} \upsilon_i \chi_{[i-1,i]}(s) \tag{24a}$$

$$t(0) = 0, \tag{24b}$$

where $\upsilon_i = \tau_i - \tau_{i-1}$. Let $\boldsymbol{\Upsilon}$ be the set of all those $\boldsymbol{\upsilon} = [\upsilon_1, \ldots, \upsilon_{m+1}]^\intercal \in \mathbb{R}^{m+1}$ such that

$$\upsilon_i \geq 0, \;\; i = 1, \ldots, m+1.$$

Obviously, the following constraint must also be satisfied:

$$\sum_{i=1}^{m+1} \upsilon_i = T. \tag{25}$$

Denote $\hat{\boldsymbol{\mu}}(s) = \boldsymbol{\mu}(t(s))$ and $\hat{\boldsymbol{\Psi}}(s) = \boldsymbol{\Psi}(t(s))$. Then, (13a), (13b), (13c) and (14a), (14b), (14c) are transformed into

$$d\hat{\boldsymbol{\mu}}(s)/ds = \upsilon(s)[\tilde{\boldsymbol{A}}(t(s), \boldsymbol{\kappa})\hat{\boldsymbol{\mu}}(s)] \tag{26a}$$

$$\hat{\boldsymbol{\mu}}(0) = \boldsymbol{\mu}^0 \tag{26b}$$

$$\hat{\boldsymbol{\mu}}(i^+) = \tilde{\boldsymbol{J}}^i \hat{\boldsymbol{\mu}}(i^-) + \tilde{\boldsymbol{\gamma}}^i, \quad i = 1, \ldots, m, \tag{26c}$$

and

$$\begin{aligned} d\hat{\boldsymbol{\Psi}}(s)/ds = \upsilon(s)[&\tilde{\boldsymbol{A}}(t(s), \boldsymbol{\kappa})\hat{\boldsymbol{\Psi}}(s) + \hat{\boldsymbol{\Psi}}^\intercal(s)\tilde{\boldsymbol{A}}(t(s), \boldsymbol{\kappa}) \\ &+ \tilde{\boldsymbol{D}}(t(s), \boldsymbol{\kappa})\tilde{\boldsymbol{\Lambda}}(t(s))[\tilde{\boldsymbol{D}}(t(s), \boldsymbol{\kappa})]^\intercal] \end{aligned} \tag{27a}$$

$$\hat{\boldsymbol{\Psi}}(0) = \boldsymbol{\Psi}^0 \tag{27b}$$

$$\hat{\boldsymbol{\Psi}}(i^+) = \tilde{\boldsymbol{J}}^i \hat{\boldsymbol{\Psi}}(i^-)(\tilde{\boldsymbol{J}}^i)^\intercal + \tilde{\boldsymbol{P}}^i, \quad i = 1, \ldots, m. \tag{27c}$$

Denote $\tilde{\boldsymbol{\delta}} = (\boldsymbol{\kappa}, \boldsymbol{\upsilon}, \boldsymbol{\gamma})$. The cost function (22) is transformed into

$$\begin{aligned} \hat{g}_0(\tilde{\boldsymbol{\delta}}) = &\hat{\Phi}_0(\hat{\boldsymbol{\mu}}(m+1), \hat{\boldsymbol{\Psi}}(m+1), \tilde{\boldsymbol{\delta}}) \\ &+ \sum_{i=1}^{m+1} \int_{i-1}^{i} \hat{\mathcal{L}}_0(t(s), \hat{\boldsymbol{\mu}}(s), \hat{\boldsymbol{\Psi}}(s), \tilde{\boldsymbol{\delta}}) ds, \end{aligned} \tag{28}$$

where

$$\hat{\Phi}_0(\hat{\boldsymbol{\mu}}(m+1), \hat{\boldsymbol{\Psi}}(m+1), \tilde{\boldsymbol{\delta}}) = \psi(\boldsymbol{\gamma}) + \boldsymbol{Q}_4^{\intercal}(\boldsymbol{\delta})\hat{\boldsymbol{\mu}}(m+1) + \boldsymbol{Q}_3(\boldsymbol{\delta}) + \mathrm{trace}\{\boldsymbol{Q}_5(\boldsymbol{\delta})[\hat{\boldsymbol{\Psi}}(m+1) + \hat{\boldsymbol{\mu}}(m+1)(\hat{\boldsymbol{\mu}}(m+1))^{\intercal}]\}$$

and

$$\hat{\mathcal{L}}_0(t(s), \hat{\boldsymbol{\mu}}(s), \hat{\boldsymbol{\Psi}}(s), \tilde{\boldsymbol{\delta}}) = \upsilon_i\{\mathrm{trace}[\boldsymbol{Q}_2(t(s), \boldsymbol{\delta})(\hat{\boldsymbol{\Psi}}(s) + \hat{\boldsymbol{\mu}}(s)(\hat{\boldsymbol{\mu}}(s))^{\intercal})] + \boldsymbol{Q}_1(t(s), \boldsymbol{\delta})^{\intercal}\hat{\boldsymbol{\mu}}(s) + \boldsymbol{Q}_0(t(s), \boldsymbol{\delta})\}.$$

The constraints (23) are transformed into

$$\begin{aligned}\hat{g}_i(\tilde{\boldsymbol{\delta}}) =& \hat{\Phi}_0(\hat{\boldsymbol{\mu}}(m+1), \hat{\boldsymbol{\Psi}}(m+1), \tilde{\boldsymbol{\delta}}) + \sum_{j=1}^{m+1} \int_{j-1}^{j} \hat{\mathcal{L}}_0(t(s), \hat{\boldsymbol{\mu}}(s), \hat{\boldsymbol{\Psi}}(s), \tilde{\boldsymbol{\delta}}) ds \\ \leq& 0, \end{aligned} \tag{29}$$

where

$$\hat{\Phi}_i(\hat{\boldsymbol{\mu}}(m+1), \hat{\boldsymbol{\Psi}}(m+1), \tilde{\boldsymbol{\delta}}) = \boldsymbol{S}_{i4}^{\intercal}(\boldsymbol{\delta})\hat{\boldsymbol{\mu}}(m+1) + \boldsymbol{S}_{i3}(\boldsymbol{\delta}) + \mathrm{trace}\{\boldsymbol{S}_{i5}(\boldsymbol{\delta})[\hat{\boldsymbol{\Psi}}(m+1) + \hat{\boldsymbol{\mu}}(m+1)(\hat{\boldsymbol{\mu}}(m+1))^{\intercal}]\}$$

and

$$\hat{\mathcal{L}}_i(t(s), \hat{\boldsymbol{\mu}}(s), \hat{\boldsymbol{\Psi}}(s), \tilde{\boldsymbol{\delta}}) = \upsilon_i\{\mathrm{trace}[\boldsymbol{Q}_2(t(s), \boldsymbol{\delta})(\hat{\boldsymbol{\Psi}}(s) + \hat{\boldsymbol{\mu}}(s)(\hat{\boldsymbol{\mu}}(s))^{\intercal})] + \boldsymbol{Q}_1(t(s), \boldsymbol{\delta})^{\intercal}\hat{\boldsymbol{\mu}}(s) + \boldsymbol{Q}_0(t(s), \boldsymbol{\delta})\}.$$

Then, after this time scaling transformation, Problem 2 is equivalent to

Problem 3. Given the dynamical system (24), (25) and (27), find a feasible parameter $\tilde{\boldsymbol{\delta}} \in \mathbb{K} \times \boldsymbol{\Upsilon} \times \boldsymbol{\Gamma}$ such that the cost function (28) is minimized subject to the constraints (26, 29).

Remark 1. Note that our formulation also holds for time varying control matrices $\boldsymbol{K} = \boldsymbol{K}(t)$, $\hat{\boldsymbol{K}} = \hat{\boldsymbol{K}}(t)$, $t > 0$. In this case, Problems 3 and 2 corresponding to Problem 1, as described above, are to be considered as deterministic optimal control problems with controls $\boldsymbol{K}(t)$ and $\hat{\boldsymbol{K}}(t)$ rather than deterministic optimal parameter selection problems with constant matrices $\boldsymbol{K}$ and $\hat{\boldsymbol{K}}$.

5 Gradient Formulae

Problem 3 is a constrained optimal parameter selection problem, where the state covariance matrix is not a vector. To solve this problem via the optimal

control software MISER3.3, we need to rewrite the dynamical system with the state in the form of vector.

Let $\boldsymbol{z}(s)$ be a vector consisting of $t(s)$, $\hat{\boldsymbol{\mu}}(s)$ and the independent components of $\hat{\boldsymbol{\Psi}}(s)$. That is,

$$\boldsymbol{z}(s) = [t(s), \hat{\boldsymbol{\mu}}^{\intercal}(s), \hat{\Psi}_{11}(s), \ldots, \hat{\Psi}_{1,n+p}(s), \hat{\Psi}_{22}(s), \ldots, \hat{\Psi}_{2,n+p}(s), \ldots, \hat{\Psi}_{n+p,n+p}(s)]^{\intercal}. \tag{30}$$

Let $\boldsymbol{f}$ be the corresponding vector obtained from the right-hand sides of (24a), (26a) and (27a). Furthermore, let Φ_i, $\mathcal{L}_i$, $i = 1, \ldots, n_0$, be obtained from $\hat{\Phi}_i$, $\hat{\mathcal{L}}_i$, $k = 1, \ldots, n_0$, respectively, with $t(s)$, $\hat{\boldsymbol{\mu}}(s)$ and $\hat{\boldsymbol{\Psi}}(s)$ replaced appropriately by $\boldsymbol{z}(s)$.

Then, Problem 3 is equivalent to

Problem 4. Given the dynamical system

$$\frac{d\boldsymbol{z}(s)}{ds} = \boldsymbol{f}(s, \boldsymbol{z}(s), \tilde{\boldsymbol{\delta}}), \tag{31a}$$

$$\boldsymbol{z}(0) = \boldsymbol{z}^0, \tag{31b}$$

$$\boldsymbol{z}(i^+) = \boldsymbol{\phi}^i(\boldsymbol{z}(i^-), \tilde{\boldsymbol{\delta}}), \qquad i = 1, \ldots, m, \tag{31c}$$

where $\boldsymbol{z}^0$ and $\boldsymbol{\phi}^i$ are obtained from (24b), (26b), (27b) and (26c), (27c), respectively, find a feasible parameter $\tilde{\boldsymbol{\delta}} \in \mathbb{K} \times \boldsymbol{\Upsilon} \times \boldsymbol{\Gamma}$, such that the cost function

$$\hat{g}_0(\tilde{\boldsymbol{\delta}}) = \Phi_0(\boldsymbol{z}(m+1 \mid \tilde{\boldsymbol{\delta}}), \tilde{\boldsymbol{\delta}}) + \sum_{k=1}^{m+1} \int_{k-1}^{k} \mathcal{L}_0(s, \boldsymbol{z}(s \mid \tilde{\boldsymbol{\delta}}), \tilde{\boldsymbol{\delta}}) ds \tag{32}$$

is minimized subject to the constraints (4.3) and

$$\hat{g}_i(\tilde{\boldsymbol{\delta}}) = \Phi_i(\boldsymbol{z}(m+1 \mid \tilde{\boldsymbol{\delta}}), \tilde{\boldsymbol{\delta}}) + \sum_{j=1}^{m+1} \int_{j-1}^{j} \mathcal{L}_i(s, \boldsymbol{z}(s \mid \tilde{\boldsymbol{\delta}}), \tilde{\boldsymbol{\delta}}) ds \le 0, \quad i = 1, \ldots, n_0. \tag{33}$$

To solve Problem 4 as a mathematical programming problem, we need the gradients of the cost function and the constraint functions. They can be obtained by using similar idea as that given for Theorem 5.2.1 of [16]. Details of these gradients are presented below in the following theorem.

Theorem 4. *The gradient of the cost function (32) and the constraints (33) with respect to* $\tilde{\boldsymbol{\delta}}$ *are given by*

$$\nabla_{\tilde{\boldsymbol{\delta}}} \hat{g}_i(\tilde{\boldsymbol{\delta}}) = \frac{\partial \Phi_i(\boldsymbol{z}(m+1), \tilde{\boldsymbol{\delta}})}{\partial \tilde{\boldsymbol{\delta}}} + \sum_{j=1}^{m} (\boldsymbol{\eta}^i(j^+))^{\intercal} \frac{\partial \boldsymbol{\phi}^i(\boldsymbol{z}(j^-), \tilde{\boldsymbol{\delta}})}{\partial \tilde{\boldsymbol{\delta}}} + \sum_{j=1}^{m+1} \int_{j-1}^{j} \frac{\partial H_i(s, \boldsymbol{z}, \boldsymbol{\eta}^i, \tilde{\boldsymbol{\delta}})}{\partial \tilde{\boldsymbol{\delta}}} ds, \qquad i = 0, 1, \ldots, n_0, \tag{34}$$

where the Hamiltonian H_i is defined by

$$H_i(s, \boldsymbol{z}, \boldsymbol{\eta}, \tilde{\boldsymbol{\delta}}) = \mathcal{L}_i(s, \boldsymbol{z}(s), \tilde{\boldsymbol{\delta}}) + (\boldsymbol{\eta}(s))^{\mathsf{T}} \boldsymbol{f}(s, \boldsymbol{z}(s), \tilde{\boldsymbol{\delta}}) \tag{35}$$

and $\boldsymbol{\eta}^i(s)$ is the co-state determined by the following differential equations:

$$\frac{d\boldsymbol{\eta}(s)}{ds} = -\left[\frac{\partial H_i(s, \boldsymbol{z}(s), \boldsymbol{\eta}(s), \tilde{\boldsymbol{\delta}})}{\partial \boldsymbol{z}}\right]^{\mathsf{T}}, \tag{36a}$$

with terminal condition

$$\boldsymbol{\eta}(m+1) = \left[\frac{\partial \boldsymbol{\Phi}_i(\boldsymbol{z}(m+1), \tilde{\boldsymbol{\delta}})}{\partial \boldsymbol{z}}\right]^{\mathsf{T}} \tag{36b}$$

and jump conditions

$$\boldsymbol{\eta}(j^-) = \left[\frac{\partial \boldsymbol{\phi}^i(\boldsymbol{z}(j^-), \tilde{\boldsymbol{\delta}})}{\partial \boldsymbol{z}}\right]^{\mathsf{T}} \boldsymbol{\eta}(j^+). \tag{36c}$$

Then, we use the following algorithm to calculate the gradients of the cost function and the constraint functions.

Algorithm 1

1. *For each given $\tilde{\boldsymbol{\delta}} \in \mathbb{K} \times \boldsymbol{\Upsilon} \times \boldsymbol{\Gamma}$, compute the solution $\boldsymbol{z}(\cdot|\tilde{\boldsymbol{\delta}})$ of the system (31a), (31b), (31c) by solving the differential equation (31a) forward in time from $s = 0$ to $s = m + 1$ with the initial condition (31b) and jump conditions (31c).*
2. *Compute the co-state solution $\boldsymbol{\eta}(\cdot|\tilde{\boldsymbol{\delta}})$ by solving the co-state differential equation (36a) backward in time from $s = m+1$ to $s = 0$ with the terminal condition (36b) and jump conditions (36c).*
3. *Apply Theorem 4 to compute the gradients of the cost function and the constraint functions.*

With the gradient given in Algorithm 1, we can apply a gradient-based method to solve Problem 4. In this chapter, we use the optimal control software package MISER3.3 (see [9]), which is based on sequential quadratic programming (SQP) routine, to solve Problem 4.

6 Example

In this section, we will give an example to find a vector $\boldsymbol{\delta} \in \mathbb{K} \times \boldsymbol{\mathcal{T}} \times \boldsymbol{\Gamma}$ such that the process $\boldsymbol{x}(t)$ of the dynamical system (1a), (1b), (1c) is closest to a given deterministic trajectory $\hat{\boldsymbol{x}}(t)$ while the uncertainty of the corresponding dynamical system is within a given acceptable limit. The cost function is given by

$$g_0(\boldsymbol{\delta}) = \psi(\boldsymbol{\gamma}) + \mathcal{E}\left\{\int_0^T [\boldsymbol{x}(t \mid \boldsymbol{\delta}) - \hat{\boldsymbol{x}}(t)]^\intercal [\boldsymbol{x}(t \mid \boldsymbol{\delta}) - \hat{\boldsymbol{x}}(t)] dt\right\}. \tag{37}$$

It can be simplified as

$$\begin{aligned} g_0(\boldsymbol{\delta}) &= \psi(\boldsymbol{\gamma}) + \int_0^T \mathcal{E}\{(\boldsymbol{x}(t \mid \boldsymbol{\delta}))^\intercal \boldsymbol{x}(t \mid \boldsymbol{\delta}) - 2(\hat{\boldsymbol{x}}(t))^\intercal \boldsymbol{x}(t \mid \boldsymbol{\delta}) + (\hat{\boldsymbol{x}}(t))^\intercal \hat{\boldsymbol{x}}(t)\} dt \\ &= \psi(\boldsymbol{\gamma}) + \int_0^T \mathcal{E}\{(\boldsymbol{\xi}(t \mid \boldsymbol{\delta}))^\intercal \boldsymbol{M} \boldsymbol{\xi}(t \mid \boldsymbol{\delta}) - 2(\hat{\boldsymbol{\xi}}(t))^\intercal \boldsymbol{\xi}(t \mid \boldsymbol{\delta}) + (\hat{\boldsymbol{x}}(t))^\intercal \hat{\boldsymbol{x}}(t)\} dt \\ &= \psi(\boldsymbol{\gamma}) + \int_0^T \{\mathrm{trace}[\boldsymbol{M}(\boldsymbol{\Psi}(t \mid \boldsymbol{\delta}) + (\boldsymbol{\mu}(t \mid \boldsymbol{\delta}))^\intercal \boldsymbol{\mu}(t \mid \boldsymbol{\delta}))] \\ &\quad - 2(\hat{\boldsymbol{\xi}}(t))^\intercal \boldsymbol{\mu}(t \mid \boldsymbol{\delta}) + (\hat{\boldsymbol{x}}(t))^\intercal \hat{\boldsymbol{x}}(t)\} dt, \end{aligned} \tag{38}$$

where $\hat{\boldsymbol{\xi}}(t) = [\hat{\boldsymbol{x}}(t) \;\; \boldsymbol{0}]^\intercal$ and $\boldsymbol{M} \in \mathbb{R}^{(n+p)\times(n+p)}$ is given by

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{I}_{n\times n} & 0 \\ 0 & 0 \end{bmatrix}$$

and $\boldsymbol{I}_{n\times n}$ is the identity matrix in $\mathbb{R}^{n\times n}$.

Let $\bar{\boldsymbol{x}}(t \mid \boldsymbol{\delta}) = \mathcal{E}\{\boldsymbol{x}(t \mid \boldsymbol{\delta})\}$. The constraint is given by

$$g_1(\boldsymbol{\delta}) = \mathcal{E}\left\{\int_0^T (\boldsymbol{x}(t \mid \boldsymbol{\delta}) - \bar{\boldsymbol{x}}(t \mid \boldsymbol{\delta}))^\intercal (\boldsymbol{x}(t \mid \boldsymbol{\delta}) - \bar{\boldsymbol{x}}(t \mid \boldsymbol{\delta})) dt\right\} - \varepsilon \le 0, \tag{39}$$

where ε is a positive constant corresponding to some acceptable level of uncertainty. Similar to (37), (39) can be simplified as

$$\begin{aligned} g_1(\boldsymbol{\delta}) &= \int_0^T \mathrm{trace}\{\boldsymbol{M}\boldsymbol{\Psi}(t \mid \boldsymbol{\delta})\} dt - \varepsilon \\ &= \int_0^T \mathrm{trace}\{\Psi_{11}(t \mid \boldsymbol{\delta}) + \Psi_{22}(t \mid \boldsymbol{\delta})\} dt - \varepsilon \le 0. \end{aligned} \tag{40}$$

We consider the dynamic system (1a), (1b), (1c) defined on $(0, 1]$ with the coefficients given by

$$\boldsymbol{A}(t) = \begin{pmatrix} 0.8 & 0.5 \\ 0.2 & -0.6 \end{pmatrix}, \; \boldsymbol{B}(t) = \begin{pmatrix} 1.2 & -0.8 \\ 0.8 & -1.2 \end{pmatrix}, \; \boldsymbol{D}(t) = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}.$$

The mean and the covariance matrix of the initial state are, respectively,

$$\bar{\boldsymbol{x}}^0 = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \boldsymbol{P}^0 = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.16 \end{pmatrix}.$$

Suppose that there are two switchings and the coefficients of the two jump functions are

$$\boldsymbol{J}^i = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}, \boldsymbol{P}^i = \begin{pmatrix} 0.04 & 0 \\ 0 & 0.04 \end{pmatrix}, \quad \forall i = 1, 2.$$

The coefficients of the observation system (2a), (2b) are given by

$$\boldsymbol{G}(t) = \begin{pmatrix} 1 & 1.2 \end{pmatrix}, \quad \boldsymbol{D}^0(t) = 0.5$$

and the feedback control is given by (3) with $\boldsymbol{C}(t) = (\boldsymbol{B}(t))^{-1}$.

The system and observation system are subject to constant-sized random shocks $\boldsymbol{N}(t)$ and $\boldsymbol{N}^0(t)$, with their mean intensity given by

$$\lambda_1(t) = \lambda_2(t) = \lambda^0(t) = 1.$$

The vectors $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ are constrained by

$$\mathbb{K} = \{\boldsymbol{\kappa} = [K_1, K_2, \hat{K}_1, \hat{K}_2]^\intercal \in \mathbb{R}^4 : -5 \le K_i, \hat{K}_i \le 5,\ i = 1, 2\},$$
$$\boldsymbol{\Gamma} = \{\boldsymbol{\gamma} = [\gamma_1^1, \gamma_2^1, \gamma_1^2, \gamma_2^2]^\intercal \in \mathbb{R}^4 : -5 \le \gamma_j^i \le 5,\ i, j = 1, 2\}.$$

The cost function is given by (37) with the target trajectory given by $\hat{\boldsymbol{x}}(t) = 1$ and the penalty function given by

$$\psi(\boldsymbol{\gamma}) = \sum_{i=1}^{2} \frac{1}{2} (\boldsymbol{\gamma}^i)^\intercal \boldsymbol{\gamma}^i$$

and the constraint is given by (39) with $\varepsilon = 0.25$.

We apply the optimal control software package MISER3.3 ([9]) to solve this problem. The optimal solutions obtained are

$$\begin{aligned} \boldsymbol{K}^* &= (-0.26791 \ \ -1.18258)^\intercal, \\ \hat{\boldsymbol{K}}^* &= (-0.33051 \ \ 0.03585)^\intercal, \\ \boldsymbol{\tau}^* &= (0.19023 \ \ 0.65155)^\intercal, \\ \boldsymbol{\gamma}^{1*} &= (0.36662 \ \ 0.38707)^\intercal, \\ \boldsymbol{\gamma}^{2*} &= (0.28008 \ \ 0.27691)^\intercal. \end{aligned}$$

The minimum objective value is $g_0^* = 0.68380$.

For simulation, we have obtained 100 sample paths of $\boldsymbol{\xi}$ in Matlab. The results are illustrated in Figs. 1, 2, and 3. These sample paths are computed from equations (6) using the optimal parameters. Each sample path has a stochastic jump at time τ_1 and τ_2. Since the system is driven by Poisson processes, one would expect discontinuities in the sample paths for the state

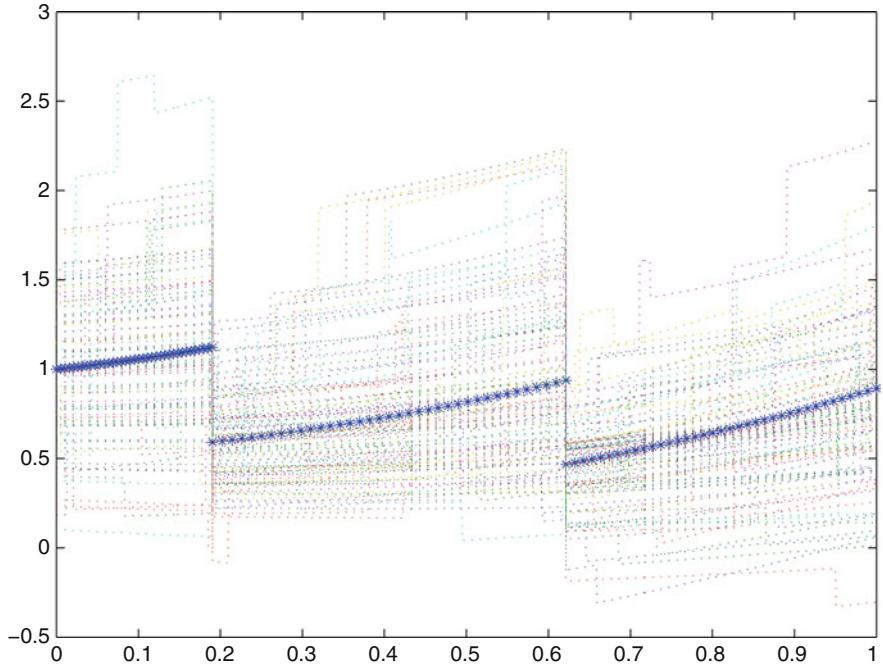


Fig. 1. * line: $\mu_1(t)$; *dotted line*: 100 sample paths of $x_1(t)$

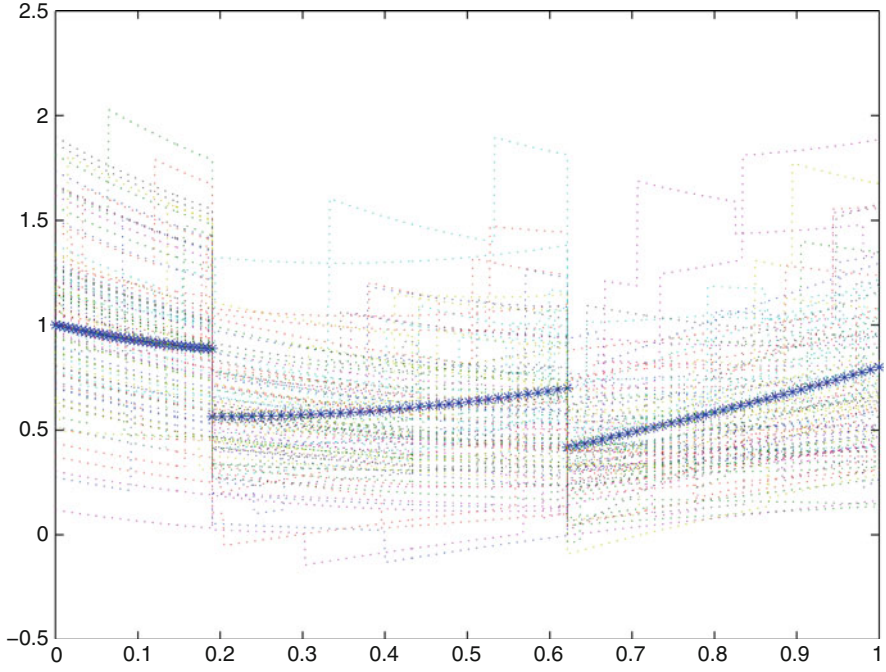


Fig. 2. * line: $\mu_2(t)$; *dotted line*: 100 sample paths of $x_2(t)$

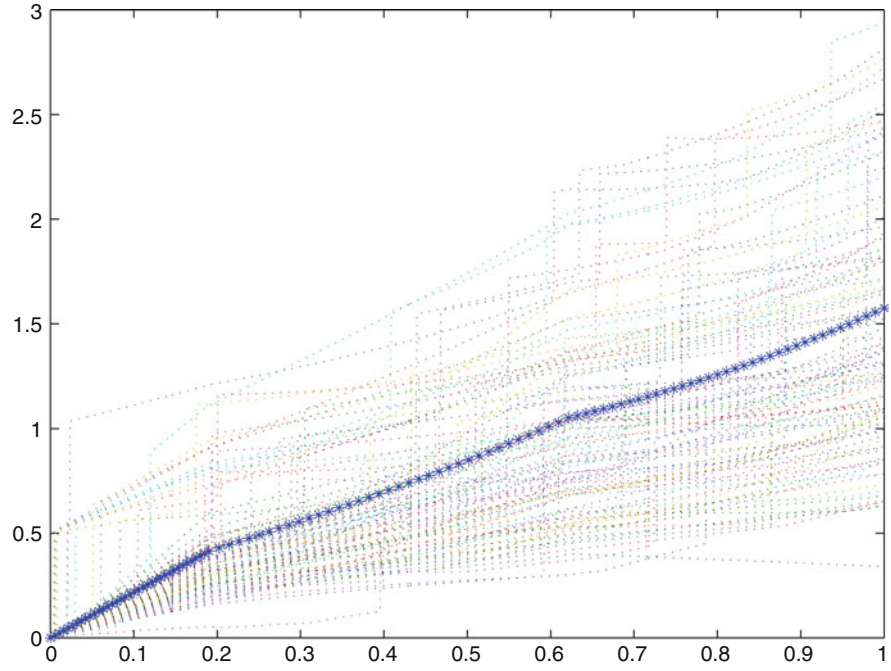


Fig. 3. * line: $\mu_3(t)$; *dotted line*: 100 sample paths of $y(t)$

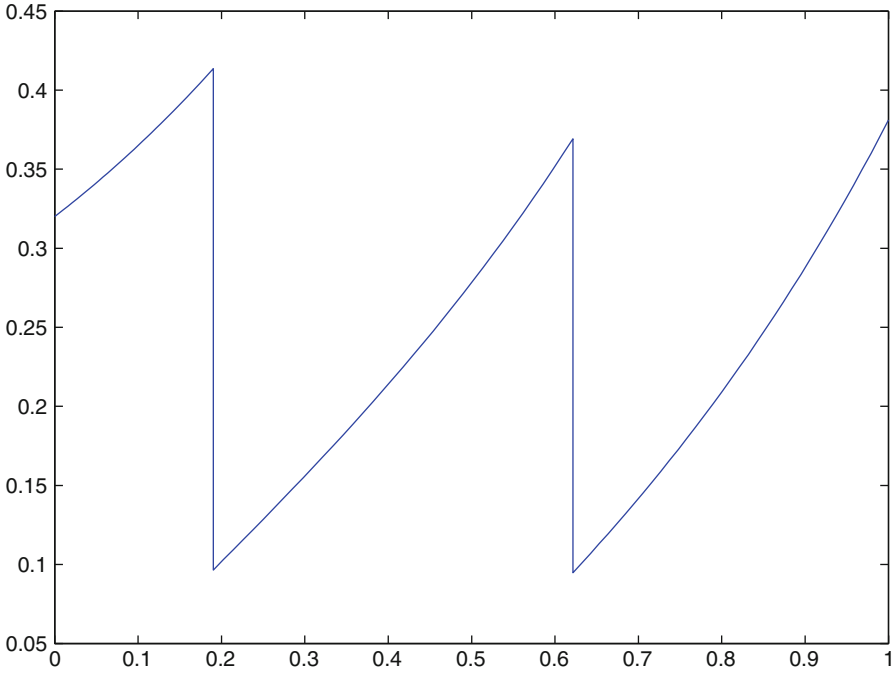


Fig. 4. $\Psi_{11}(t) + \Psi_{22}(t)$, $t \in [0, 1]$

variables. In fact, besides at time τ_1 and τ_2, the sharp corners in Figs. 1, 2, and 3, correspond to the time points at which the Poisson jumps take place.

From Figs. 1 and 2, we see the deviation of $\boldsymbol{x}(t)$ to the target trajectory $\hat{\boldsymbol{x}}(t) = 1$. This is because the constraint (40) must be satisfied. To see how this works, we plot the figure of $\Psi_{11}(t) + \Psi_{22}(t)$, which is the function integrated in (40), in Fig. 4. We can see from this figure that the occurrence of each jump tends to reduce the fluctuation, aiming to ensure the satisfaction of the constraint.

7 Conclusion

In this chapter, a class of optimal feedback control problems involving a stochastic impulsive dynamical system is considered. We have shown that this stochastic optimal impulsive parameter selection problem is equivalent to a deterministic impulsive optimal parameter selection problem. A numerical method was developed for solving this equivalent constrained deterministic impulsive optimal parameter selection problem. From the numerical study through solving a numerical example, we see that the solution method is effective.

Acknowledgment

This project is supported by a research grant from the Australian Research Council and Chongqing Key Laboratory of Operations Research and System Engineering.

References

1. Ahmed, N.U., Georgenas, N.D.: On optimal parameter selection. IEEE Trans. Automat. Contr. 18, 313–314 (1973)
2. Dolezal, J.: On the solution of optimal control problems involving parameters and general boundary conditions. Kybernetika 17, 71–81 (1981)
3. Feng, Z.G., Teo, K.L., Ahmed, N.U., Zhao, Y., Yan, W.Y.: Optimal fusion of sensor data for Kalman filtering. Discrete Contin. Dyn. Syst., 14, 483–503 (2006)
4. Feng, Z.G., Teo, K.L., Ahmed, N.U., Zhao, Y., Yan, W.Y.: Optimal fusion of sensor data for discrete Kalman filtering. Dyna. Syst. Appl. 16, 393–406 (2007)
5. Feng, Z.G., Teo, K.L., Rehbock, V.: Hybrid method for a general optimal sensor scheduling problem in discrete time. Automatica 44, 1295–1303 (2008)
6. Feng, Z.G., Teo, K.L., Rehbock, V.: Optimal sensor scheduling in continuous time. Dyn. Syst. Appl. 17, 331–350 (2008)

7. Feng, Z.G., Teo, K.L., Zhao, Y.: Branch and bound method for sensor scheduling in discrete time. J. Ind. Manage. Optim. 1, 499–512 (2005)
8. Goh, C.J., Teo, K.L.: On constrained stochastic optimal parameter selection problems. Bull. Aust. Math. Soc. 41, 393–405 (1990)
9. Jennings, L.S., Fisher, M.E., Teo, K.L., Goh, C.J.: MISER3, Optimal Control Software, Version 3.0, Theory and User Manual. On the website (2004) http://www.cado.uwa.edu.au/miser
10. Jennings, L.S., Teo, K.L.: A computational algorithm for functional inequality constrained optimization problems. Automatica 26, 371–375 (1990)
11. Liu, C.M., Feng, Z.G., Teo, K.L.: On a class of stochastic impulsive optimal parameter selection problems. Int. J. Innov. Comput., Info. Control 5, 1043–1054 (2009)
12. Lee, H.W.J., Teo, K.L., Rehbock, V., Jennings, L.S.: Control parametrization enhancing technique for time optimal control problems. Dyn Syst Appl. 6, 243–262 (1997)
13. Situ, R.: Theory of Stochastic Differential Equations with Jumps and Applications: Mathematical and Analytical Techniques with Applications to Engineering. Springer, New York (2005)
14. Teo, K.L., Ahmed, N.U.: Optimal feedback control for a class of stochastic systems. Internat. J. Systems Sci. 5, 357–365 (1974)
15. Teo, K.L., Ahmed, N.U., Fisher, M.E.: Optimal feedback control for linear stochastic systems driven by counting processes. Eng. Optim. 15, 1–16 (1989)
16. Teo, K.L., Goh, C.J., Wong, K.H.: A Unified Computational Approach to Optimal Control Problems. Longman Scientific and Technical (1991)

Analysis of Differential Inclusions: Feedback Control Method

Vyacheslav Maksimov

Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia
maksimov@imm.uran.ru

Summary. In this chapter, controlled differential inclusions in a Hilbert space containing subdifferentials of convex functions are considered. The following three problems are studied: the problem of prescribed motion realization, the problem of robust control, and the problem of input dynamical reconstruction. Solution algorithms that are stable with respect to informational noises and computational errors are presented. The algorithms are based on the method of feedback control. They adaptively take into account inaccurate measurements of phase trajectories and are regularized in the following sense: the more precise is incoming information, the better is the algorithm's output.

Key words: differential inclusions, feedback control, reconstruction

1 Introduction

In the recent time, a part of mathematical control theory, namely, the theory of control of distributed systems, has been intensively developed. To a considerable degree, this is stimulated by the fact that a rather wide set of problems of mathematical physics and mechanics are described by distributed systems. At present, there exists a number of monographs devoted to control problems for dynamical systems in Hilbert or Banach spaces [1, 2, 9].

In all these works, the emphasis is on problems of program control in the case when all system's parameters are precisely specified and not subject to changes. However, investigation of control problems for systems with uncontrollable disturbances (problems of robust control) is also natural. Similar problems are poorly investigated. This is connected with the fact that

This work was supported in part by the Russian Foundation for Basic Research (project 09-01-00378), the Russian Fund for Humanities 08-02-00315, the Program for Support of the Leading Scientific Schools of Russia, the Program of Presidium of RAS (project 09-P-1-1007), and the Ural-Siberian interdisciplinary project 09-C-1-1010.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_14,

the well-known maximum principle is not applicable to solving them. In the early 1970s, N.N. Krasovskii suggested an effective approach to solving robust (guaranteed) control problems. This approach is based on the formalism of positional strategies. Its essence consists in reduction of the robust control problem to two problems:

(i) the problem of choosing an auxiliary-controlled system M (hereinafter, it is called a model);
(ii) the problem of choosing some rule for synchronous controlling of both model and real systems.

The systematic description of the theory of guaranteed positional control for dynamical systems described by ordinary differential equations is given in [6, 7]. This theory for systems with distributed parameters is presented in [10, 12].

Then it was revealed that the approach developed in [6, 7] is useful for solving dynamical inverse problems (problems of dynamical reconstruction), see, for example, [4, 5, 8, 11, 13, 14] (here we mention only monographs and reviews). The goal of the present work is to illustrate possibilities of the approach in question for investigating some control and reconstruction problems for systems described by differential inclusions containing subdifferentials of convex functions. These systems have been rather actively studied in the resent years [1, 3]; this is caused, in particular, by the fact that variational inequalities are often reduced to inclusions of similar kind.

Let a dynamical system be described by the parabolic inclusion:

$$\dot{x}(t) + \partial\varphi(x(t)) \ni Bu(t) - Cv(t) + f(t), \quad t \in T = [t_0, \vartheta]. \tag{1}$$

Here $H = H^*$ is a real Hilbert space with a norm $|\cdot|_H$ and a scalar product $(\cdot;\cdot)_H$, $f(\cdot) \in L_2(T;H)$ is a given function, $\varphi : H \to \overline{\mathbb{R}} = \{r \in \mathbb{R} : -\infty < r \leq +\infty\}$ is a lower semicontinuous convex function, $\partial\varphi$ is the subdifferential of φ. Let $x(t_0) = x_0 \in D(\varphi) = \{x \in H : \varphi(x) < +\infty\}$ be an initial state. Let $(U, |\cdot|_U)$ and $(V, |\cdot|_V)$ be uniformly convex Banach spaces; $B \in \mathcal{L}(U;H)$, $C \in \mathcal{L}(V;H)$ be linear continuous operators. It is known that there exists (for any $\{u(\cdot), v(\cdot)\} \in L_2(T;U) \times L_2(T;V)$) a unique solution $x(\cdot) = x(\cdot; t_0, x_0, u(\cdot), v(\cdot))$ of inclusion (1) with the following properties [1, 3]:

$$x(\cdot) \in W(T), \quad x(t) \in D(\varphi) \quad \forall t \in T, \quad t \to \varphi(x(t)) \in AC(T).$$

Here $AC(T)$ is the set of absolutely continuous functions $z(\cdot)\colon T \to \mathbb{R}$, $W(T) = \{z(\cdot) \in L_2(T;H)\colon \dot{z}(\cdot) \in L_2(T;H)\}$; the derivative $\dot{z}(\cdot)$ is understood in the sense of distributions.

The chapter is devoted to three problems: the problem of prescribed motion realization (Problem 1), the problem of robust control (Problem 2), and the problem of input dynamical reconstruction (Problem 3). Let us give the contensive formulation of these problems and describe the approach to their solution.

Let a uniform net

$$\Delta = \{\tau_i\}_{i=0}^{m}, \quad \tau_i = \tau_{i-1} + \delta, \quad \tau_0 = t_0, \quad \tau_m = \vartheta$$

with a diameter $\delta = \delta(\Delta) = \tau_i - \tau_{i-1}$ be fixed on a given time interval T. Let a solution of inclusion (1), namely $x(\cdot)$, be unknown. At moments $\tau_i \in \Delta$ the phase states $x(\tau_i)$ are inaccurately measured. Results of measurements $\xi_i^h \in H$, $i \in [0 : m-1]$, satisfy the inequalities

$$|\xi_i^h - x(\tau_i)|_H \le h. \tag{2}$$

Here, $h \in (0, 1)$ is a level of informational noise.

Let us consider the following problem.

Problem 1. Assume that $v = v(t) \equiv 0$, $t \in T$, in the right-hand part of inclusion (1). A number $\varepsilon > 0$ is given. There is some prescribed motion $x^*(\cdot)$; it is a solution of the inclusion

$$\begin{aligned} \dot{x}^*(t) + \partial\varphi(x^*(t)) &= Bu^*(t) + f(t), \quad t \in T, \\ x^*(t_0) &= x_0. \end{aligned} \tag{3}$$

Both the solution $x^*(\cdot)$ and the function $u^*(\cdot)$ are unknown. It is only known that $u^*(t) \in D_*$ for a. a. (almost all) $t \in T$, where $D_* \subset U$ is a given bounded and closed set. At the moments $\tau_i \in \Delta$ the states $x^*(\tau_i)$ as well as $x(\tau_i)$ are (inaccurately) measured. Results of measurements, elements $\psi_i^h \in H$, $i \in [0 : m-1]$, satisfy the inequalities

$$|\psi_i^h - x^*(\tau_i)|_H \le h.$$

The problem of prescribed motion realization consists in designing an algorithm forming (by the feedback principle) a control $u = u(\tau_i, \xi_i^h, \psi_i^h)$, $t \in \delta_i = [\tau_i, \tau_{i+1})$, $i \in [0 : m-1]$, such that the solution of inclusion (1) remains within the ε-neighborhood of the solution $x^*(\cdot)$ of inclusion (3) for all $t \in T$, i.e.,

$$\sup_{t \in T} |x(t) - x^*(t)|_H \le \varepsilon.$$

Let the following quality criterion be given:

$$I(x(\cdot), u(\cdot)) = \sigma(x(\vartheta)) + \int_{t_0}^{\vartheta} \chi(t, x(t), u(t))\, dt,$$

where $\sigma : H \to \mathbb{R}$ and $\chi : T \times H \times U \to \mathbb{R}$ are given functions satisfying the local Lipschitz conditions. A prescribed value of the criterion, number I_*, is fixed.

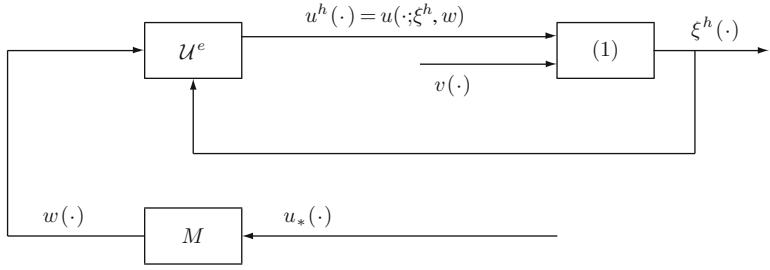
The next consists in the following.

Problem 2. It is required to construct an algorithm of feedback control $u = u(\tau_i, \xi_i^h)$, $t \in \delta_i = [\tau_i, \tau_{i+1})$, $i \in [0 : m-1]$, of inclusion (1) providing fulfillment of the following condition. Whatever a value $\varepsilon > 0$ and a disturbance $v(\cdot)$ ($v(t) \in Q$, $t \in T$) may be, one can indicate (explicitly) numbers $h_* > 0$ and $\delta_* > 0$ such that the inequality $|I(x(\cdot), u(\cdot)) - I_*| \le \varepsilon$ is fulfilled.

Problem 3 is as follows.

Problem 3. Let the control $u = u(t)$, $t \in T$, is equal to zero in inclusion (1). It is required to design a dynamical algorithm of reconstruction of an unknown input $v = v(\cdot)$ in the "real-time" mode.

The scheme of an algorithm for solving the problem of robust control is given in the figure below [6, 7].



In the beginning, an auxiliary system M (called a model) is introduced. The model has an input $u_*(\cdot)$ and an output $w(\cdot)$. The process of synchronous feedback control of inclusion (1) and M is organized on the interval T. This process is decomposed into $(m-1)$ identical steps. At the ith step carried out during the time interval $\delta_i = [\tau_i, \tau_{i+1})$, the following actions are fulfilled. First, at the time moment τ_i, according to some chosen rule $\mathcal{U}^e$, the element

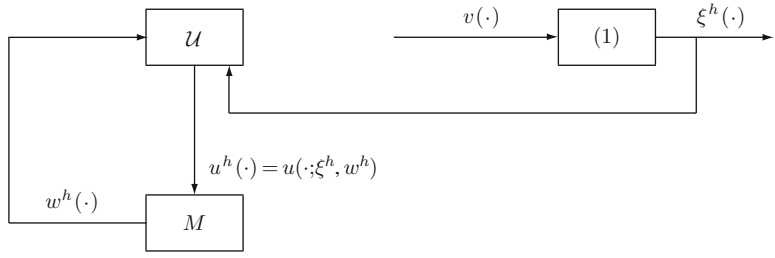
$$u_i = \mathcal{U}^e(\tau_i, p_i)$$

is calculated. Here the symbol p_i denotes some value called a position which includes a pair (ξ_i^h, w_i), $w_i = w(\tau_i)$. Then (till the moment τ_{i+1}) the control $u(t) = u_i$, $\tau_i \le t < \tau_{i+1}$, is fed to the input of inclusion (1). The values ξ_{i+1}^h and $w_{i+1} = w(\tau_{i+1})$ are treated as algorithm's output at the ith step.

An analogous scheme is applicable to solving the problem of prescribed motion realization. In this case, inclusion (3) plays the role of the model.

The scheme of algorithm for solving the problem of reconstruction is shown in the following figure [8, 11, 13].

In this case, an auxiliary system M (a model) is also introduced. The model has an input $u^h(\cdot)$ and an output $w^h(\cdot)$. The problem of reconstruction is replaced by the problem of designing an algorithm of feedback control of the model. This algorithm is identified with some function $\mathcal{U}$ which is chosen in such a way that the control $u^h(\cdot)$ approximates the unknown disturbance $v(\cdot)$: $u^h(t) = u_i^h = \mathcal{U}(\tau_i, p_i)$, $\quad t \in \delta_i$, where $p_i = (\xi_i^h, w^h(\tau_i))$.



2 Statement of the Problems

Before we present rigorous formulations of the problems in question, let us give some definitions. Thereinafter, we denote by $u_{a,b}(\cdot)$ a function $u(t)$, $t \in [a,b]$, considered as a whole. Any strongly measurable functions $u(\cdot) : T \to P$ and $v(\cdot) : T \to Q$ are called an open-loop control and a disturbance, respectively. The sets of all open-loop controls and disturbances are denoted by the symbols $P_T(\cdot)$ and $Q_T(\cdot)$. The symbol $P_{a,b}(\cdot)$ stands for restriction of the set $P_T(\cdot)$ onto the segment $[a,b] \subset T$. A unique solution of inclusion (1) with the properties $x(t_*) = x_*$, $x(\cdot) = x(\cdot; t_*, x_*, u_{t_*,\vartheta}(\cdot), v_{t_*,\vartheta}(\cdot)) \in W([t_*,\vartheta])$, $x(t) \in D(\varphi)$ $\forall t \in [t_*,\vartheta]$, $t \to \varphi(x(t)) \in AC([t_*,\vartheta])$ is called a motion of system (1) starting from a position $(t_*, x_*) \in T \times D(\varphi)$ and corresponding to a control $u_{t,\vartheta}(\cdot) \in P_{t,\vartheta}(\cdot)$ and a disturbance $v_{t,\vartheta}(\cdot) \in Q_{t,\vartheta}(\cdot)$. If $u(t) = 0$ for $t \in [t_*,\vartheta]$ (or $v(t) = 0$ for $t \in [t_*,\vartheta]$), then we write $x(\cdot) = x(\cdot; t_*, x_*, v_{t_*,\vartheta}(\cdot))$ (or $x(\cdot) = x(\cdot; t_*, x_*, u_{t_*,\vartheta}(\cdot))$).

The symbol $\mathcal{P}$ denotes some set called the set of "positions." Each problem has its own set of positions. The sense of $\mathcal{P}$ will be clarified for each specific problem. Any possible function (multifunction)

$$\mathcal{U} : T \times \mathcal{P} \to P \tag{4}$$

is said to be a feedback strategy. Feedback strategies correct controls at discrete time moments given by some partition of the interval T.

Hereinafter, denote a phase trajectory of a model M by the symbol $w(\cdot)$ (or $w^h(\cdot)$).

Consider the problem of prescribed motion realization (Problem 1). In this case, the model M is described by inclusion (3), i.e.,

$$w(\cdot) = x^*(\cdot; t_0, x_0, u^*_{t_0,\vartheta}(\cdot)).$$

The set of positions $\mathcal{P}$ is $H \times H$, i.e., $p_i = (\xi^h_i, \psi^h_i)$, $|\xi^h_i - x(\tau_i)|_H \le h$, $|\psi^h_i - x^*(\tau_i)|_H \le h$. A solution $x(\cdot)$ of inclusion (1) starting from an initial state (t_*, x_*) and corresponding to a piecewise constant control $u^h(\cdot)$ (formed by the feedback principle

$$u^h(t) = u^h_i \in \mathcal{U}(\tau_i, p_i) \in U, \quad t \in [\tau_i, \tau_{i+1}), \quad i \in [i(t_*) : m-1], \quad p_i \in \mathcal{P}, \tag{5}$$

$$i(t_*) = \min\{i : \tau_i > t_*\}, \quad u^h(t) = u^h_* \in \mathcal{U}(t_*, x_*, x_*) \ \text{ for } t \in [t_*, \tau_{i(t_*)})\big)$$

and to a disturbance $v_{t_*,\vartheta}(\cdot) \in Q_{t_*,\vartheta}(\cdot)$ is called an (h,Δ,w)-motion $x^h_{\Delta,w}(\cdot; t_*, x_*, \mathcal{U})$ generated by the positional strategy $\mathcal{U}$ on partition Δ. Thus, when we write $x^h_{\Delta,w}(\cdot)$, we mean a solution of inclusion (1) constructed by the feedback principle. The set of all (h,Δ,w)-motions is denoted by $X_h(t_*, x_*, \mathcal{U}, \Delta, w)$. Problem 1 consists in constructing a feedback strategy $\mathcal{U}$ (4) with the following properties: whatever a value $\varepsilon > 0$ may be, one can specify (explicitly) numbers $h_* > 0$ and $\delta_* > 0$ such that the inequalities

$$\rho(x^h_{\Delta,w}(\cdot), x^*(\cdot)) \le \varepsilon \qquad \forall x^h_{\Delta,w}(\cdot) \in X_h(t_0, x_0, \mathcal{U}, \Delta, w) \tag{6}$$

are fulfilled uniformly with respect to all measurements ξ^h_i with properties (2), if $h \le h_*$ and $\delta = \delta(\Delta) \le \delta_*$.

Here $\rho(x(\cdot), x^*(\cdot)) = \sup_{t\in T} |x(t) - x^*(t)|_H$.

Let us pass to the problem of robust control (Problem 2). Consider the following ordinary differential equation:

$$\dot g(t) = \chi(t, x(t), u(t)), \quad g(t_0) = 0. \tag{7}$$

Introducing this new variable g, we reduce the robust control problem of Bolza type to a control problem with a terminal quality criterion of the form $I = \sigma(x(\vartheta)) + g(\vartheta)$. In this case, the controlled system consists of inclusion (1) in the Hilbert space H and ordinary differential equation (7).

Let a model be described by the inclusion

$$\dot w_1(t) + \partial\varphi(w_1(t)) \ni u_1(t) + f(t), \quad w_1(t) \in H, \quad w_1(t_0) = x_0, \tag{8}$$

and the ordinary differential equation

$$\dot w_2(t) = u_2(t), \quad w_2(t) \in \mathbb{R}, \quad w_2(t_0) = 0. \tag{9}$$

The model is, in essence, a "copy" of system (1), (7): inclusion (8) and equation (9) correspond to inclusion (1) and equation (7), respectively. However, in contrast to "real" system (1), (7), the model does not contain a disturbance.

In this case, the phase state of the model at some moment t is the pair $w(t) = \{w_1(t), w_2(t)\} \in H \times \mathbb{R}$. The set of positions $\mathcal{P}$ is $(H \times \mathbb{R})^2$. A pair $\{x(\cdot), g(\cdot)\}$, where $x(\cdot)$ is a solution of inclusion (1) starting from an initial state (t_*, x_*) and $g(\cdot)$ is a solution of (7) starting from an initial state (t_*, g_*) corresponding to a piecewise constant control $u^h(\cdot)$ (formed by the feedback principle)

$$\begin{aligned} u^h(t) &= u_i \in \mathcal{U}^e(\tau_i, p_i), \quad t \in [\tau_i, \tau_{i+1}), \quad i \in [i(t_*) : m-1], \\ p_i &= (\xi^h_i, \tilde\psi^h_i, w(\tau_i)), \quad |\xi^h_i - x(\tau_i)|_H \le h, \quad |\tilde\psi^h_i - g(\tau_i)| \le h, \\ i(t_*) &= \min\{i : \tau_i > t_*\}, \quad u^h(t) = u_* \in \mathcal{U}^e(t_*, x_*, g_*, x_*, g_*) \\ &\qquad \text{for } t \in [t_*, \tau_{i(t_*)})) \end{aligned} \tag{10}$$

and to a disturbance $v_{t_*,\vartheta}(\cdot) \in Q_{t_*,\vartheta}(\cdot)$ is called an (h,Δ,χ)-motion

$$z_\Delta^h(\cdot) = \{x_\Delta^h(\cdot; t_*, x_*, \mathcal{U}^e, v_{t_*,\vartheta}(\cdot)), g_\Delta^h(\cdot; t_*, g_*, \mathcal{U}^e, v_{t_*,\vartheta}(\cdot))\}$$

generated by the positional strategy $\mathcal{U}^e$: $T \times \mathcal{P} \to P$ on the partition Δ. The set of all (h, Δ, χ)-motions is denoted by $Z_h^\chi(t_*, x_*, g_*, \mathcal{U}^e, \Delta)$. It is clear that the set $Z_h^\varphi(t_*, x_*, g_*, \mathcal{U}^e, \Delta)$ is not empty for $(t_*, x_*, g_*) \in T \times D(\varphi) \times \mathbb{R}$.

Problem 2 consists in the following. A prescribed value of the criterion, number I_*, is fixed. It is necessary to construct a positional strategy $\mathcal{U}^e$: $T \times \mathcal{P} \to P$ with the following properties: whatever a value $\varepsilon > 0$ and a disturbance $v_T(\cdot) \in Q_T(\cdot)$ may be, one can indicate (explicitly) numbers $h_* > 0$ and $\delta_* > 0$ such that the inequalities

$$|I(x_{\Delta T}^h(\cdot), u_T^h(\cdot)) - I_*| \le \varepsilon \tag{11}$$

are fulfilled uniformly with respect to all measurements ξ_i^h with properties (2) and measurements $\tilde{\psi}_i$ with properties $|\tilde{\psi}_i - g_\Delta^h(\tau_i)| \le h$, if $h \le h_*$ and $\delta = \delta(\Delta) \le \delta_*$. Here $\{x_\Delta^h(\cdot), g_\Delta^h(\cdot)\} \in Z_h^\chi(t_0, x_0, 0, \mathcal{U}^e, \Delta)$, $x_\Delta^h(\cdot) = x(\cdot; t_0, x_0, \mathcal{U}^e(\cdot), v_T(\cdot))$, $g_\Delta^h(\cdot) = p(\cdot; t_0, 0, \mathcal{U}^e(\cdot), v_T(\cdot))$, the control $u^h(\cdot)$ is defined by (10).

Let us turn to the problem of reconstruction. In this case, a disturbance $v(\cdot)$ to be reconstructed is an element of the space $L_2(T; V)$. Inclusion (1) has the form

$$\dot{x}(t) + \partial\varphi(x(t)) \ni -Cv(t) + f(t), \quad t \in T. \tag{12}$$

Its solution generated by a disturbance $v(\cdot) \in L_2(T; V)$ is denoted by the symbol $x(\cdot) = x(\cdot; t_0, x_0, v(\cdot))$. A model M is described by the inclusion

$$\dot{w}^h(t) + \partial\varphi(w^h(t)) \ni -Cu^h(t) + f(t), \quad t \in T, \quad w^h(t_0) = x_0. \tag{13}$$

A control in model (13) is defined by the rule

$$u^h(t) = u_i^h = \mathcal{U}(\tau_i, p_i) \in V, \quad t \in \delta_i = [\tau_i, \tau_{i+1}), \tag{14}$$

where $\mathcal{P} = H \times H$, $p_i = (\xi_i^h, w^h(\tau_i))$. The element ξ_i^h, being a result of measuring the phase state of inclusion (12) at the moment τ_i, satisfies inequality (2). Thus, the control $u^h(\cdot)$ in model (14) is formed by the feedback principle. This means that at each moment $t \in \delta_i$ ($i \in [0 : m-1]$) the element u_i^h is calculated by the position p_i. Then the constant control of the form (14) is fed to the input of the model during the time interval δ_i. The procedure for forming the control $u^h(\cdot)$ stops at the moment ϑ.

Let $v_*(\cdot) = v_*(\cdot; x(\cdot))$ be a minimal $L_2(T; V)$-norm element of the set $V_*(x(\cdot))$ of all functions $v(\cdot) \in L_2(T; V)$ generating a solution $x(\cdot)$: $V_*(x(\cdot)) = \{\tilde{v}(\cdot) \in L_2(T; V)\colon x(\cdot) = x(\cdot; t_0, x_0, \tilde{v}(\cdot))\}$.

The problem of dynamical reconstruction (Problem 3) consists in constructing a feedback strategy $\mathcal{U} : T \times \mathcal{P} \to V$ such that the control $u^h(\cdot)$ defined by (14) possesses the property

$$u^h(\cdot) \to v_*(\cdot; x(\cdot)) \text{ in } L_2(T; V) \quad \text{as} \quad h \to 0.$$

The sequence of actions (the solving algorithm), which are necessary for reconstruction of the function $v_*(\cdot)$, is described in Section 5. Note that the reconstruction procedure is performed synchronously with the operation of system (12). Namely, before an arbitrary moment $t \in [t_0, \vartheta)$, the trajectory $x(\tau), \tau \in [t_0, t]$, is realized. Up to the same moment, the control $u^h(\tau), \tau \in [t_0, t)$ being an approximation to $v_*(\tau), \tau \in [t_0, t)$, is calculated. Up to any subsequent moment $t_1 \in (t, \vartheta]$, the new "part" of the control $u^h(\tau), \tau \in [t, t_1)$ is calculated.

3 The Algorithm for Solving Problem 1

Let us describe the procedure of forming an (h, Δ, w)-motion $x^h_{\Delta,w}(\cdot; t_0, x_0, \mathcal{U})$ generated by a fixed partition Δ and a strategy $\mathcal{U}$ of the form

$$\mathcal{U}(t, x, w) = \arg\min\{(x - w, Bu)_H : u \in P\}, \tag{15}$$

i.e., we describe the algorithm for solving Problem 1.

We take an arbitrary element $u^h_0 \in P$ on the interval $[t_0, \tau_1)$. The (h, Δ, w)-motion $\{x^h_{\Delta,w}(\cdot; t_0, x_0, u^h_0)\}_{t_0,\tau_1}$ is realized under the action of the control $u(t) = u^h_0$, $t \in [t_0, \tau_1)$. At the moment $t = \tau_1$ we determine u^h_1 from the condition

$$u^h_1 \in \mathcal{U}(\tau_1, p_1),\ p_1 = (\xi^h_1, \psi^h_1),\ |\xi^h_1 - x^h_{\Delta,w}(\tau_1)|_H \le h,\ |\psi^h_1 - x^*(\tau_1)|_H \le h.$$

Then we compute the realization of the (h, Δ, w)-motion $\{x^h_{\Delta,w}(\cdot; \tau_1, x^h_{\Delta,w}(\tau_1), u^h_1)\}_{\tau_1,\tau_2}$. Let the (h, Δ, w)-motion $x^h_{\Delta,w}(\cdot)$ be defined on the interval $[t_0, \tau_i]$ $(\tau_i = \tau_{i,h})$. At the moment $t = \tau_i$ we choose

$$u^h_i \in \mathcal{U}(\tau_1, p_i),\ p_i = (\xi^h_i, \psi^h_i),\ |\xi^h_i - x^h_{\Delta,w}(\tau_i)|_H \le h,\ |\psi^h_i - x^*(\tau_i)|_H \le h.$$

As the result of the action of the control $u^h(t) = u^h_i$, $t \in [\tau_i, \tau_{i+1})$, $i \in [0 : m-1]$, the (h, Δ, w)-motion of system (1) $\{x^h_{\Delta,w}(\cdot; \tau_i, x^h_{\Delta,w}(\tau_i), u^h_i)\}_{\tau_i,\tau_{i+1}}$ is realized on the interval $[\tau_i, \tau_{i+1}]$. The described above procedure of forming the (h, Δ, w)-motion stops at the moment ϑ.

Theorem 1. *Let $D_* = P$. Then the positional strategy $\mathcal{U}$ defined by (15) solves Problem 1.*

The proof of Theorem 1 is performed by the scheme of the proof of Theorem 2 from the next section.

4 The Algorithm for Solving Problem 2

To solve Problem 2, we use ideas from [6, 7], namely, the method of a priori stable sets. In our case, this method consists in the following. At first, a

trajectory of model (8), (9), $w(\cdot) = \{w_1(\cdot), w_2(\cdot)\}$, possessing the property $\sigma(w_1(\vartheta)) + w_2(\vartheta) = I_*$ is constructed in a special way. Then a feedback strategy $\mathcal{U} = \mathcal{U}^e$ providing tracing the prescribed trajectory of the model by the trajectory of real system (1), (7) is constructed. This means that the (h, Δ, χ)-motion $z^h_\Delta(\cdot) = \{x^h_\Delta(\cdot), g^h_\Delta(\cdot)\}$ formed by the feedback principle (see (10)) by means of the strategy above remains at a "small" neighborhood of the trajectory $w(\cdot)$ during the whole interval T. This property of the (h, Δ, χ)-motion allows us to conclude that the chosen strategy solves the robust control problem. Let us pass to the realization of this scheme.

Let

$$\Phi(t, x, u, v) = \{Bu - Cv, \chi(t, x, u)\},$$

$$\Phi_u(t, x, v) = \bigcup_{u \in P} \Phi(t, x, u, v), \quad H_*(t; x) = \bigcap_{v \in Q} \Phi_u(t, x, v),$$

$$H_*(\cdot; x) = \{u(\cdot) \in L_2(T; H \times \mathbb{R}) : u(t) \in H_*(t; x) \quad \text{for a. a. } t \in T\}.$$

The following condition is fulfilled.

Condition 1. There exists an open-loop control $u_*(\cdot) = \{u_1(\cdot), u_2(\cdot)\}$, $u_*(t) \in H_*(t; w_1(t))$ for a.a. $t \in T$, such that $I_* = \sigma(w_1(\vartheta)) + w_2(\vartheta)$.

Let, for example, a closed set $D \subset H$ be such that $BP = CQ + D$ and $\chi = \chi(t, u)$. Here we use the following notation: $BP = \{Bu\colon u \in P\}$, $CQ = \{Cv\colon v \in Q\}$, $CQ + D = \{u\colon u = u_1 + u_2,\ u_1 \in CQ,\ u_2 \in D\}$. Then

$$H_*(t; x) = H_*(t) = D \times \left\{ \bigcup_{u \in P} \chi(t, u) \right\} \subset H \times \mathbb{R}.$$

Let us describe the procedure of forming the (h, Δ, χ)-motion $z^h_\Delta(\cdot) = \{x^h_\Delta(\cdot), g^h_\Delta(\cdot)\}$ corresponding to a fixed partition Δ and a strategy $\mathcal{U}^e$ of the form:

$$\begin{aligned} \mathcal{U}^e(t, x, p, w) = \{u^e \in P : (x - w_1(t), Bu^e)_H + (p - w_2(t))\chi(t, x, u^e) \\ \le \inf_{u \in P}[(x - w_1(t), Bu)_H + (p - w_2(t))\chi(t, x, u)] + h\}. \quad (16) \end{aligned}$$

The algorithm for solving Problem 2 is as follows. Before the start of algorithm's work, we fix a value $h \in (0, 1)$ and a partition $\Delta = \{\tau_i\}_{i=0}^m$ with a diameter $\delta = \delta(\Delta)$. The work of the algorithm is decomposed into $m - 1$ identical steps. We assume that

$$u^h(t) = u_0 \in \mathcal{U}^e(t_0, x, p, w(t_0)) = P$$

on the interval $[t_0, \tau_1)$. Under the action of this control as well as of an unknown disturbance $v_{t_0,\tau_1}(\cdot)$, the (h, Δ, χ)-motion $\{z^h_\Delta(\cdot)\}_{t_0,\tau_1} = \{x^h_\Delta(\cdot; t_0, x_0, \mathcal{U}^e, v_{t_0,\tau_1}(\cdot)), g^h_\Delta(\cdot, t_0, 0, \mathcal{U}^e, v_{t_0,\tau_t}(\cdot))\}_{t_0,\tau_1}$ is realized. At the moment $t = \tau_1$ we determine u_1 from the condition

$$u_1 \in \mathcal{U}^e(\tau_1, \xi_1^h, \tilde{\psi}_1^h, w(\tau_i)), \quad |\xi_1^h - x_\Delta^h(\tau_1)|_H \le h, \quad |\tilde{\psi}_1^h - g_\Delta^h(\tau_1)| \le h,$$

i.e., $u^h(t) = u_1$ for $t \in [\tau_1, \tau_2)$. Then we calculate the realization of the (h, Δ, χ)-motion

$$\{z_\Delta^h(\cdot)\}_{\tau_1,\tau_2} = \{x_\Delta^h(\cdot; \tau_1, x_\Delta^h(\tau_1), \mathcal{U}, v_{\tau_1,\tau_2}(\cdot)), g_\Delta^h(\cdot; \tau_1, g_\Delta^h(\tau_1), \mathcal{U}, v_{\tau_1,\tau_2}(\cdot))\}_{\tau_1,\tau_2}.$$

Let the (h, Δ, χ)-motion $z_\Delta^h(\cdot)$ be defined in the interval $[t_0, \tau_i]$. At the moment $t = \tau_i$ we assume that

$$u_i \in \mathcal{U}^e(\tau_i, \xi_i^h, \tilde{\psi}_i^h, w(\tau_i)), \quad |\xi_i^h - x_\Delta^h(\tau_i)|_H \le h, \quad |\tilde{\psi}_i^h - g_\Delta^h(\tau_i)| \le h,$$

i.e., $u^h(t) = u_i$ for $t \in [\tau_i, \tau_{i+1})$. As the result of the action of this control and of an unknown disturbance $v_{\tau_i,\tau_{i+1}}(\cdot)$, the (h, Δ, χ)-motion

$$\begin{aligned}\{z_\Delta^h(\cdot)\}_{\tau_i,\tau_{i+1}} &= \{x_\Delta^h(\cdot; \tau_i, x_\Delta^h(\tau_i), \mathcal{U}, v_{\tau_i,\tau_{i+1}}(\cdot)), g_\Delta^h(\cdot; \tau_i, g_\Delta^h(\tau_i), \\ &\quad \mathcal{U}, v_{\tau_i,\tau_{i+1}}(\cdot))\}_{\tau_i,\tau_{i+1}}\end{aligned}$$

is realized on the interval $[\tau_i, \tau_{i+1}]$. The described above procedure of forming the (h, Δ, χ)-motion stops at the moment ϑ.

Theorem 2. *Let condition 1 be fulfilled. Then the strategy $\mathcal{U}^e(t, x, p, w)$ of the form (16) solves Problem 2.*

Proof. We can give the following scheme of the proof. Let a partition $\Delta = \{\tau_i\}_{i=0}^m$ of the interval T with a diameter $\delta(\Delta) = \delta$ and a value of the level of informational noise h be fixed. We estimate the evolution of the function

$$\mu(t) = \frac{1}{2}|x_\Delta^h(t) - w_1(t)|_H^2 + \frac{1}{2}|g_\Delta^h(t) - w_2(t)|^2.$$

Introduce the functional $l(y(\cdot)) : W(T) \to \mathbb{R}$,

$$l(y(\cdot)) = |y(\cdot)|_{C(T;H)} + |\dot{y}(\cdot)|_{L_2(T;H)}.$$

One can prove in a standard way [1, 3] that there exists a number K_* such that, for any $x_0 \in D(\varphi)$, $u_T(\cdot) \in P_T(\cdot)$, $v_T(\cdot) \in Q_T(\cdot)$, $x(\cdot) = x(\cdot; t_0, x_0, u_T(\cdot), v_T(\cdot))$, the inequality

$$l(x(\cdot)) \le K_*(1 + \varphi^{1/2}(x_0) + |u(\cdot)|_{L_2(T;U)} + |v(\cdot)|_{L_2(T;V)}) \tag{17}$$

is true. It is easily seen that for a. a. $t \in [\tau_i, \tau_{i+1})$, $i \ge 1$, the inequality

$$\begin{aligned}\frac{d}{dt}\mu(t) &\le (Bu_i - Cv(t) - u_1(t), x_\Delta^h(t) - w_1(t))_H \\ &\quad + (\chi(t, x_\Delta^h(t), u_i) - u_2(t))(g_\Delta^h(t) - w_2(t))\end{aligned} \tag{18}$$

holds. Here

$$u_i \in \mathcal{U}^e(\tau_i, \xi_i^h, \tilde{\psi}_i^h, w(\tau_i)), \qquad |\xi_i^h - x_\Delta^h(\tau_i)|_H \le h, \quad |\tilde{\psi}_i^h - g_\Delta^h(\tau_i)| \le h, \tag{19}$$

$v_{\tau_i,\tau_{i+1}}(\cdot)$ is an unknown realization of disturbance, the strategy $\mathcal{U}^e$ is determined from (16). It follows from (17),(18), and (19) and the local Lipschitz property of the function $\chi(\cdot)$ that

$$\begin{aligned}\frac{d}{dt}\mu(t) \le{}& (Bu_i - Cv(t) - u_1(t), \xi_i^h - w_1(\tau_i))_H \\ &+ (\chi(\tau_i, \xi_i^h, u_i) - u_2(t))(\tilde{\psi}_i^h - w_2(\tau_i)) \\ &+ k_1 \left(h + \int_{\tau_i}^{t} \{|\dot{x}_\Delta^h(\tau)|_H + |\dot{w}_1(\tau)|_H + |\dot{g}_\Delta^h(\tau)| + |\dot{w}_2(\tau)|\}\, d\tau \right),\end{aligned} \tag{20}$$

$$t \in \delta_i = [\tau_i, \tau_{i+1}),$$

and constant k_1 can be explicitly written. Let

$$s_i = \{\xi_i^h - w_1(\tau_i), \tilde{\psi}_i^h - w_2(\tau_i)\}.$$

Then the sum of the two first terms in the right-hand part of inequality (20) can be written in the form of the scalar product

$$\Big(s_i, \Phi(\tau_i, \xi_i^h, u_i, v(t)) - u_*(t)\Big)_{H\times\mathbb{R}},$$

where

$$\Phi(\tau_i, \xi_i^h, u_i, v(t)) = \{Bu_i - Cv(t), \chi(\tau_i, \xi_i^h, u_i)\}, \quad u_* = \{u_1, u_2\}.$$

The symbol $(\cdot,\cdot)_{H\times\mathbb{R}}$ denotes the scalar product in space $H \times \mathbb{R}$. Let us define elements v_i^e from the conditions

$$\inf_{u\in P}(s_i, \Phi(\tau_i, \xi_i^h, u, v_i^e))_{H\times\mathbb{R}} \ge \sup_{v\in Q}\inf_{u\in P}(s_i, \Phi(\tau_i, \xi_i^h, u, v))_{H\times\mathbb{R}} - h. \tag{21}$$

It is obvious (see Condition 1) that

$$u_*(t) \in H(t, w_1(t)) \subset \bigcup_{u\in P} \Phi(t, u, w_1(t), v_i^e) \quad \text{for a. a.} \quad t \in [\tau_i, \tau_{i+1}).$$

Then there exists a control $u^{(1)}(t) \in P$, $t \in \delta_i$, such that

$$\Phi(t, w_1(t), u^{(1)}(t), v_i^e) = u_*(t) \quad \text{for a. a.} \quad t \in [\tau_i, \tau_{i+1}]. \tag{22}$$

Using the rule of definition of the strategy $\mathcal{U}^e$, we deduce that

$$\begin{aligned}(s_i, \Phi(\tau_i, \xi_i^h, u_i, v(t)))_{H\times\mathbb{R}} &\le \sup_{v\in Q}(s_i, \Phi(\tau_i, \xi_i^h, u_i, v))_{H\times\mathbb{R}} \\ &\le \inf_{u\in P}\sup_{v\in Q}(s_i, \Phi(\tau_i, \xi_i^h, u, v))_{H\times\mathbb{R}} + h.\end{aligned} \tag{23}$$

In turn, from (21) we have

$$\sup_{v\in Q}\inf_{u\in P}(s_i,\Phi(\tau_i,\xi_i^h,u,v))_{H\times\mathbb{R}}\le\inf_{u\in P}(s_i,\Phi(\tau_i,\xi_i^h,u,v_i^e))_{H\times\mathbb{R}}+h. \tag{24}$$

Moreover, it is evident that the equality

$$\inf_{u\in P}\sup_{v\in Q}(s_i,\Phi(\tau_i,\xi_i^h,u,v))_{H\times\mathbb{R}}=\sup_{v\in Q}\inf_{u\in P}(s_i,\Phi(\tau_i,\xi_i^h,u,v))_{H\times\mathbb{R}} \tag{25}$$

is valid. From (23), (24), and (25) we have

$$\begin{aligned}(s_i,\Phi(\tau_i,\xi_i^h,u_i,v(t)))_{H\times\mathbb{R}}&\le\inf_{u\in P}(s_i,\Phi(\tau_i,\xi_i^h,u,v_i^e))_{H\times\mathbb{R}}+2h \qquad (26)\\ &\le(s_i,\Phi(t,\xi_i^h,u^{(1)}(t),v_i^e))_{H\times\mathbb{R}}+2h+L(t-\tau_i).\end{aligned}$$

Here L is a Lipschitz constant of the function $\chi(\cdot)$. In this case, it follows from (22), (26) that for $t\in\delta_i$

$$(s_i^*,\Phi(\tau_i,\xi_i^h,u_i,v(t))-u_*(t))_{H\times\mathbb{R}}\le 2h+L(t-\tau_i)+L|\xi_i^h-w_1(t)|_H. \tag{27}$$

We derive from inequalities (20), (27)

$$\begin{aligned}\mu(t)\le\mu(\tau_i)+k_2\delta\Bigg(h+\delta+\int_{\tau_i}^{\tau_{i+1}}\Big\{&|\dot x_\Delta^h(\tau)|_H+|\dot g_\Delta^h(\tau)| \qquad (28)\\ &+|\dot w_1(\tau)|_H+|\dot w_2(\tau)|+|x_\Delta^h(\tau)-w_1(\tau)|_H\Big\}\,d\tau\Bigg),\quad t\in\delta_i.\end{aligned}$$

Since

$$\mu(t_0)=0,\quad \mu(\tau_1)\le k_2(h+\delta^{1/2}),$$

$$\int_{\tau_i}^{\tau_{i+1}}|x_\Delta^h(\tau)-w_1(\tau)|_H\,d\tau\le 0,5\left(\delta+\int_{\tau_i}^{\tau_{i+1}}|x^h(\tau)-w_1(\tau)|_H^2\,d\tau\right),$$

by (28) we have

$$\begin{aligned}\mu(t)\le{}&k_2(h+\delta^{1/2})\\ &+k_3\delta\left(1+h(\vartheta-t_0)/\delta+\int_{t_0}^{t}\{|\dot x_\Delta^h(\tau)|_H+|\dot w_1(\tau)|_H+|\dot g_\Delta^h(\tau)|+|\dot w_2(\tau)|\}\,d\tau\right)\\ &+k_4\delta\int_{t_0}^{t}|x_\Delta^h(\tau)-w_1(\tau)|_H^2\,d\tau,\quad t\in T.\end{aligned}$$

Here constants k_j, $j=1,\dots,4$, do not depend on h,δ and can be explicitly written. From (17) and the last inequality it follows that for any $\gamma>0$ one can find numbers $h_1>0$ and $\delta_1>0$ such that inequality $\mu(t)\le\gamma$ is fulfilled for all $h\in(0,h_1)$ and $\delta\in(0,\delta_1)$. The conclusion of the theorem follows from this inequality. The theorem is proved.

5 The Algorithm for Solving Problem 3

In this section, we consider inclusion (12) with some unknown $v(\cdot)$. We assume that $U = V$ is a Hilbert space with a scalar product $(\cdot,\cdot)_U$ and a norm $|\cdot|_U$. Constructions described below are based on the approach developed in [4, 5, 8, 11, 13, 14].

Let a family of partitions

$$\Delta_h = \{\tau_{i,h}\}_{i,h=0}^{m_h}, \quad \tau_{i,h} = \tau_{i-1,h} + \delta, \quad \tau_{0,h} = t_0, \quad \tau_{m_h,h} = \vartheta, \tag{29}$$

and a function $\alpha(h) : (0,1) \to \mathbb{R}^+$ be fixed. Let the following condition be fulfilled:

$$\begin{gathered} \alpha(h) \to 0, \quad \delta(h) \to 0, \quad h\delta^{-1}(h) \le \text{const}, \\ (\delta^{1/2}(h) + h)\alpha^{-1}(h) \to 0 \quad \text{as} \quad h \to 0. \end{gathered} \tag{30}$$

A positional strategy $\mathcal{U} : T \times H \times H \to V$ is defined by the rule

$$\mathcal{U}(\tau_i, p_i) = \alpha^{-1}(h)C^*(\xi_i^h - w^h(\tau_i)), \tag{31}$$

where $p_i = (\xi_i^h, w^h(\tau_i))$ is a position for $t \in \delta_i = [\tau_i, \tau_{i+1})$, $\tau_i = \tau_{i,h}$, $w^h(\cdot)$ is a solution of inclusion (13) with $u^h(\cdot)$ defined by (14), (31).

Let us describe the algorithm for solving Problem 3. The work of the algorithm corresponds to the following scheme. First, before the moment t_0, a partition $\Delta = \Delta_h = \{\tau_i\}_{i=0}^m$, $\tau_i = \tau_{i,h}$, of the interval T is chosen and fixed. The work of the algorithm is decomposed into $m-1$ identical steps. At the ith step carried out during the time interval $[\tau_i, \tau_{i+1})$, the following sequence of actions is fulfilled. The output $x(\tau_i)$ is inaccurately measured, i.e., some value $\xi_i^h \in H$ with properties (2) is calculated. Then the model control is determined by (14), (31) and after that we form the new part of the model trajectory $w^h(t)$, $t \in (\tau_i, \tau_{i+1}]$ instead of $w^h_{t_0,\tau_i}(\cdot)$ (memory correction). The procedure stops at the time moment ϑ.

As it follows from results of the works cited above (see, for example, [11, Theorem 1.2.1]), the convergence $u^h(\cdot) \to v_*(\cdot)$ in $L_2(T;V)$ as $h \to 0$ takes place if the model control $u^h(\cdot)$ possesses the following properties:

$$\sup_{t\in T} |x(t) - w^h(t)|_H \le \mu_1(h),$$

$$|u^h(\cdot)|^2_{L_2(T;U)} \le |v_*(\cdot)|_{L_2(T;U)} + \mu_2(h),$$

where $\mu_1(h) \to 0+$, $\mu_2(h) \to 0+$ as $h \to 0+$. From the proof of Theorem 3 (see (42), (43)) we conclude that the control $u^h(\cdot)$ formed by the strategy $\mathcal{U}$ (31) possesses these properties.

Theorem 3. *Let condition (30) be fulfilled. Then the positional strategy $\mathcal{U}$ of the form (14), (31) solves Problem 3.*

Proof. Let us estimate the variation of the Lyapunov functional

$$\varepsilon_h(t) = |\mu_h(t)|_H^2 + \alpha(h) \int_{t_0}^{t} \{|u^h(\tau)|_U^2 - |v_*(\tau)|_U^2\}\, d\tau,$$

where $\mu_h(t) = w^h(t) - x(t)$, $w^h(\cdot)$ is the solution of inclusion (13). It is easily seen that

$$\begin{aligned}\dot{\varepsilon}_h(t) &= (\mu_h(t), \dot{\mu}_h(t))_H + \alpha(h)\{|u^h(t)|_U - |v_*(t)|_U\} \\ &\le \big(\mu_h(t), C(v_*(t) - u^h(t))\big)_H + \alpha(h)\{|u^h(t)|_U - |v_*(t)|_U\}.\end{aligned}$$

In virtue of (2), we have

$$\begin{aligned}(C(v_*(t) - u^h(t)), \mu_h(t))_H &\le (C(v_*(t) - u^h(t)), \xi_i^h - w^h(\tau_i))_H \\ &+ c_1\{|v_*(t)|_U + |u^h(t)|_U\} \left(h + \int_{\tau_i}^{t} \{|\dot{x}(\tau)|_H + |\dot{w}^h(\tau)|_H\}\, d\tau \right), \quad t \in \delta_i.\end{aligned} \tag{32}$$

Note (see (14), (31)) that

$$u^h(t) = \arg\min\{\alpha|u|_U^2 - 2(C^*(\xi_i^h - w^h(\tau_i)), u)_U : u \in U\}, \quad t \in \delta_i. \tag{33}$$

Therefore, by (33), we obtain for $t \in \delta_i$

$$\begin{aligned}\varepsilon_h(t) &\le \varepsilon_h(\tau_i) \\ &+ c_2 \left(h^2 + \delta \int_{\tau_i}^{t} \left(|v_*(\tau)|_U^2 + |u^h(\tau)|_U^2 + |\dot{x}(\tau)|_H^2 + |\dot{w}^h(\tau)|_H \right) d\tau \right).\end{aligned} \tag{34}$$

Similarly to (17) we have

$$\int_{t_0}^{\vartheta} |\dot{w}^h(\tau)|_H^2\, d\tau \le K_1(1 + \varphi(x_0) + |u^h(\cdot)|_{L_2(T;U)}^2). \tag{35}$$

By summing the right-hand and left-hand parts of inequality (34) over i, we have

$$\begin{aligned}\varepsilon_h(t) &\le \varepsilon_h(t_0) + c_4 h^2 \delta^{-1} + c_3 \delta \left(1 + \int_{t_0}^{t} \{|v_*(\tau)|_U^2 + |u^h(\tau)|_U^2\} d\tau \right), \\ &\le \varepsilon_h(t_0) + c_4 h^2 \delta^{-1} + c_5 \delta + c_3 \delta^2 \sum_{j=0}^{i(t)} |u_j^h|_U^2,\end{aligned} \tag{36}$$

where the symbol $i(t)$ denotes the integer part of a number t. Besides, by the rule of definition of u_i^h (see (14), (31)), we have

$$|u_i^h|_U^2 \le 2b^2(\mu_i^h + h^2)\alpha^{-2}(h), \tag{37}$$

where $b = |C^*|_{L(U;H)}$, $\mu_i^h = |\mu^h(\tau_i)|_H^2$. From (36), (37) and the inequality $h\delta^{-1}(h) \le \text{const}$, we derive the estimate

$$\mu_i^h \le \varepsilon_h(t_0) + c_4h^2\delta^{-1} + c_5\delta + \alpha|v_*(\cdot)|_{L_2(T;U)}^2 + c_3\delta^2\sum_{j=0}^{i-1} 2b^2(\mu_j^h + h^2)\alpha^{-2}$$

$$\le c_6(h + \delta + \alpha) + c_7\delta^2\alpha^{-2}\sum_{j=0}^{i-1}\mu_j^h.$$

Taking into account the Gronwall inequality and the inequality

$$\delta(h)\alpha^{-2}(h) \le C, \tag{38}$$

we conclude that

$$\mu_i^h \le c_6(h + \delta + \alpha)\exp\{c_7(\vartheta - t_0)\delta\alpha^{-2}\} \le c_8(h + \delta + \alpha). \tag{39}$$

Summing the left-hand part of inequality (37) over i, we obtain from (39)

$$\delta^2\sum_{j=0}^{m_h-1}|u_j^h|_U^2 \le 2\delta^2b^2\sum_{j=0}^{m_h-1}(\mu_i^h + h^2)\alpha^{-2} \le c_9\delta\alpha^{-2}(\alpha + h + \delta). \tag{40}$$

Using (36) and (40), we can derive the estimation

$$\varepsilon_h(t) \le c_{10}(h + \delta + \delta\alpha^{-1} + \delta^2\alpha^{-2} + h\delta\alpha^{-2}) \le c_{11}(h + \delta\alpha^{-1}). \tag{41}$$

Therefore,

$$|u^h(\cdot)|_{L_2(T;U)}^2 \le |v_*(\cdot)|_{L_2(T;U)}^2 + c_{11}(h + \delta^{1/2})\alpha^{-1}, \tag{42}$$

$$|\mu_h(t)|_H^2 \le c_{12}(h + \delta\alpha^{-1} + \alpha). \tag{43}$$

The validity of the theorem follows from (42), (43) and Theorem 2.1 [11].

Let us adduce an estimate of the algorithm's convergence rate. Let the following condition be fulfilled.

Condition 2. The function φ is differentiable and operator $\Phi x = \operatorname{grad}\phi(x)$: $H \to H$ is Lipshitz.

Then the following theorem takes place.

Theorem 4. *Let $U = H$, C be the identity operator and $v_*(\cdot)$ be a function of bounded variation. Then the estimate*

$$|v_*(\cdot) - u^h(\cdot)|_{L_2(T;H)} \le K\{[h + \delta(h)\alpha^{-1}(h) + \alpha(h)]^{1/2} + \alpha^{-1}(h)(h + \delta^{1/2}(h))\}$$

is valid.

Proof. In this case, inclusion (1) can be rewritten in the form of parabolic equation :

$$\dot{x}(t) + \Phi x(t) = -Cv(t) + f(t).$$

Due to the Lipschitz property of mapping Φ, we have

$$\left| \int_{t_1}^{t_2} C(v_*(t) - u^h(t))\, dt \right|_H$$

$$\leq |\mu_h(t_2) - \mu_h(t_1)|_H + K_1 \int_{t_1}^{t_2} |\mu_h(\tau)|_H \, d\tau \quad \text{for any } t_1, t_2 \in [t_0, \vartheta],\ t_1 < t_2.$$

From this inequality and estimation (43), we get

$$\left| \int_{t_1}^{t_2} C(v_*(t) - u^h(t))\, dt \right|_H \leq K_2 \{h + \delta(h)\alpha^{-1}(h) + \alpha(h)\}^{1/2}. \tag{44}$$

The following lemma is known.

Lemma 1. (Osipov and Kryazhimskii [13] and Maksimov [11]) *Let* $(X, |\cdot|_X)$ *be a Hilbert space,* $u(\cdot) \in L_\infty(T;\ X)$*,* $v(\cdot)$ *be a function of bounded variation,*

$$\left| \int_{t_0}^{t} u(\tau) d\tau \right|_X \leq \varepsilon, \quad |v(t)|_X \leq K \quad \forall t \in T.$$

Then

$$\left| \int_{t_0}^{t} (u(\tau), v(\tau))_X \, d\tau \right| \leq \varepsilon(K + \operatorname{var}_X(T;\ v(\cdot))).$$

Here symbol $\operatorname{var}_X(T;\ v(\cdot))$ means the total variation of function $t \to v(t) \in X$ over the interval T. Taking into account this lemma, from (42, 44), we can conclude that

$$|v_*(\cdot) - u^h(\cdot)|^2_{L_2(T;H)}$$

$$\leq 2|v_*(\cdot)|^2_{L_2(T;H)} - 2 \int_{t_0}^{\vartheta} (v_*(t), u^h(t))_H \, dt + K_3\alpha^{-1}(h)(h + \delta^{1/2}(h))$$

$$\leq K_4\{h + \delta(h)\alpha^{-1}(h) + \alpha(h)\}^{1/2} + K_3\alpha^{-1}(h)(h + \delta^{1/2}(h)).$$

The theorem is proved.

6 Conclusion

In this chapter, differential inclusions containing subdifferentials of convex functions are investigated. The method of auxiliary models controlled by the feedback principle is developed for such inclusions. On the base of this method, algorithms for solving some reconstruction and control problems are designed.

References

1. Barbu, V.: Optimal Control of Variational Inequalities. Research Notes in Mathematics, Pitman Advanced Publishing Program, London (1984)
2. Bensoussan, A., Da Prato, G., Delfour, M.C., Mitter, S.K.: Representation and Control of Infinite Dimensional Systems. Vol. 1. Birkhäuser, Boston, MA (1992)
3. Brezis, H.: Propriétés régularisantes de certains semi-groupes non linéaires. Israel J. Math. 9, 4, 513–534 (1971)
4. Favini, A., Maksimov, V.I., Pandolfi, L.: A deconvolution problem related to a singular system. J. Mat. Anal. Applic. 292, 1, 60–72 (2004)
5. Keesman, K.J., Maksimov, V.I.: On feedback identification of unknown characteristics: a bioreactor case study. International Journal of Control, 81, 1, 134–145 (2008)
6. Krasovskii, N.N.: Controlling of a Dynamical System. Nauka, Moscow (1985) (in Russian)
7. Krasovskii, N.N., Subbotin, A.I.: Game-Theoretical Control Problems. Springer, Berlin (1988)
8. Kryazhimskii, A.V., Osipov, Yu.S.: Modeling of a control in a dynamic system. Eng. Cybern 21, 2, 38–47 (1983) (in Russian)
9. Lions, J.L.: Contrôle des systémes distribués singuliers. Bordas, Paris (1983). English transl.: Control of distributed singular systems. Gauthier–Villars (1985)
10. Maksimov, V.: Feedback minimax control for parabolic variational inequality. C. R. Acad. Sci., Paris. Series II b, 328, 1, 105–108 (2000)
11. Maksimov, V.I.: Dynamical Inverse Problems of Distributed Systems. VSP, Boston (2002)
12. Osipov, Yu.S.: On a positional control in parabolic systems. Prikl. Math. Mechan. 41, 2, 23–27 (1977) (in Russian)
13. Osipov, Yu.S., Kryazhimskii, A.V.: Inverse Problems for Ordinary Differential Equatios: Dynamical Solutios. Gordon and Breach, London (1995)
14. Osipov, Yu.S., Kryazhimskii, A.V., Maksimov, V.I.: Dynamical inverse problems for parabolic systems. Differ. Equa., 36, 5, 579–597 (2000) (in Russian)

A Game Theoretic Algorithm to Solve Riccati and Hamilton–Jacobi–Bellman–Isaacs (HJBI) Equations in H_∞ Control

Brian D. O. Anderson[1], Yantao Feng[2], and Weitian Chen[3]

[1] Research School of Information Sciences and Engineering, the Australian National University, Canberra ACT 0200, Australia; National ICT Australia, Tower A, 7 London Circuit, Canberra ACT 2601, Australia
brian.anderson@anu.edu.au
[2] Research School of Information Sciences and Engineering, the Australian National University, Canberra ACT 0200, Australia; National ICT Australia, Tower A, 7 London Circuit, Canberra ACT 2601, Australia
alex.feng@anu.edu.au
[3] Research School of Information Sciences and Engineering, the Australian National University, Canberra ACT 0200, Australia
weitian.chen@anu.edu.au

Summary. In this chapter, we propose a new algorithm to solve Riccati equations and certain Hamilton–Jacobi–Bellman–Isaacs (HJBI) equations arising in H_∞ control. The need for the algorithm is motivated by the existence of H_∞ problems for which standard Riccati solvers break down, but which can be handled by the algorithm. By using our algorithm, we replace the problem of solving H_∞ Riccati equations or HJBI equations by the problem of solving a sequence of H_2 Riccati equations or Hamilton–Jacobi–Bellman (HJB) equations. The algorithms have some advantages such as a simple initialization, local quadratic rate of convergence, and a natural game theoretic interpretation. Some numerical examples are given to demonstrate advantages of our algorithm.

Key words: Riccati, HJBI, iterative, game theoretic, convergence

1 Introduction

This chapter addresses computational issues in H_∞ control, in particular advancing algorithms for solving H_∞ Riccati equations and their generalization and Hamilton–Jacobi–Bellman–Isaacs (HJBI) equations. Though algorithms are not especially well developed for HJBI equation solution, there are certainly standard software packages allowing H_∞ Riccati equation solution, e.g.,

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_15,

RICPACK (see [5]) and MATLAB, and it is natural to ask why another algorithm should be needed, at least for this class of equation. Therefore, we spend some time describing the motivation for this work.

The motivation actually goes back to the problem of solving H_2 Riccati equations. Again, well-established software tools exist, for example, LAPACK and BLAS (see [4]). However, there exist examples of H_2 Riccati equations where these tools break down, as we review later. One technique which can remain viable in such situations is the recursive algorithm of Kleinman (see [24]). The Kleinman algorithm replaces the task of solving the Riccati equation directly by the task of solving a recursive sequence of Lyapunov equations. Now just as there exist examples where standard H_2 Riccati solvers break down, so is this true with standard H_∞ Riccati solvers. It is natural then to ask, Can the problem be fixed by the extension of the algorithm of Kleinman to this situation? The immediate answer is no (see Example 2 in the Appendix of [27] for a demonstration of this). However, as this chapter sets out, there is a fix, motivated by the Kleinman algorithm and in some ways constituting a significant extension of the Kleinman algorithm. The relevant ideas were originally presented in [26, 27].

To the extent that an H_2 Riccati equation is effectively a particular example of a Hamilton–Jacobi–Bellman (HJB) equation, and an H_∞ Riccati equation a particular example of a HJBI equation, it is natural to seek generalizations of the Riccati algorithms. Such generalizations may not just be useful in a few situations where numerical problems arise, but might be generally useful, given the limited development to this point of standard packages for solving HJB and HJBI equations. Indeed, there is an old generalization of the Kleinman algorithm, which replaces solution of the HJB nonlinear partial differential equation by the recursive solution of a sequence of linear partial differential equations (see [29]). There is, however, no corresponding algorithm for HJBI equations, and this chapter's second main contribution is to offer such an algorithm.

Apart from the ability of the algorithms we present to solve problems that may defeat conventional solvers, we note their following specific advantages: (1) a simple initialization; (2) local quadratic rate of convergence; (3) a natural game theoretic interpretation; (4) high numerical stability and reliability.

We shall now provide a high-level description of the algorithms. First, we replace the problem of solving an H_∞ Riccati equation by the problem of solving a sequence of H_2 Riccati equations. By using our algorithm, we transfer an H_∞ problem into a sequence of optimal control problems; by doing so, we indeed transfer a difficult problem into a sequence of less difficult problems. Since any single H_2 Riccati equation can be solved using the Kleinman algorithm, i.e., by solving an iterative sequence of Lyapunov equations, it is also apparent that an H_∞ equation can, if desired, be solved by using a nested double iteration of Lyapunov equations.

Second, and by way of generalization, we replace the problem of solving an HJBI equation by the problem of solving a sequence of Hamilton–Jacobi–

Bellman (HJB) equations. As for the Riccati case, an HJBI equation can be solved using a nested double iteration of linear partial differential equations. Whether one uses the single or double iteration is of course optional.

We shall now present more details on the approach for Riccati equations. Consider the following algebraic Riccati equation (ARE) in the variable P:

$$0 = A^T P + PA + PRP + Q, \tag{1}$$

where A, Q, R are real $n \times n$ matrices with Q and R symmetric. Here, $(\cdot)^T$ denotes the transpose of $(\cdot)$. Associated with this Riccati equation is a $2n \times 2n$ Hamiltonian matrix

$$H := \begin{pmatrix} A & R \\ -Q & -A^T \end{pmatrix}.$$

Generally speaking, existing methods to solve AREs can be divided into two categories:

1. Direct: solutions of ARE (1) can be constructed via computation of an n- dimensional invariant subspace of the Hamiltonian matrix H (for example, using the Schur algorithm in [28]).
2. Iterative: a sequence of matrices which converge to the unique stabilizing solution of special classes of the ARE (1) is constructed (for example, using the Kleinman algorithm in [24]).

Several different direct methods to solve the ARE (1) are given in [1, 5, 11, 15, 25, 28, 31, 33, 35]. However, compared with iterative methods to solve ARE (1), direct methods present computational disadvantages in some situations. For example, in Example 6 in [28], the solution to an H_2 ARE obtained by the Schur algorithm in [28] is inaccurate but the iterative solution obtained by the Kleinman algorithm in [24] is accurate to 13 digits in just two iterations.

Traditionally, in H_2 control, one needs to solve AREs with $Q \geq 0$ and $R \leq 0$. In H_∞ control, one needs to solve AREs with positive semidefinite Q and sign indefinite R. Although the Kleinman algorithm in [24] has been shown to have many advantages such as convergence for any suitable initial condition and a local quadratic rate of convergence [24], these advantages are strictly restricted to AREs arising in H_2 control where R in (1) must be negative semidefinite. It is not difficult to adjust the Kleinman method (which is effectively an implementation of an equation solver using Newton's method) to also handle the separate case where $R \geq 0$, but still sign-indefinite R cannot be handled. So the question naturally arises, "Can one extend the Kleinman algorithm in [24] to solve AREs with a sign indefinite quadratic term, as those that arise in H_∞ control?" The answer is that an iterative algorithm with very simple initialization to solve such a class of AREs will be given in this chapter, but the algorithm cannot be obtained by simply permitting indefinite R to occur in the Kleinman algorithm.

In the Kleinman algorithm, when a suitable initial condition is chosen and some necessary assumptions hold, it is proved that a series of Lyapunov

equations can be recursively constructed at each iteration, and positive semidefinite solutions of these Lyapunov equations converge to the stabilizing solution of the corresponding H_2 ARE. In our proposed algorithm, an ARE with a sign indefinite quadratic term is replaced by a sequence of H_2 AREs (each of which could be solved by the Kleinman algorithm if desired, though this need not happen), and the solution of the original ARE with a sign indefinite quadratic term is obtained by recursively solving these H_2 AREs.

Besides the Kleinman algorithm, there are some other iterative methods to solve AREs [1–3, 11–16, 22, 25, 32, 33, 35], some of which exhibit quadratic convergence. Among iterative methods to solve the ARE (1), Newton-type algorithms are typical and widely used [1, 11–16, 25, 33, 35]. In fact, Newton's method can be used to solve more than just symmetric AREs like (1). It can also be used to solve non-symmetric AREs where Q and R in (1) are not necessarily symmetric [20, 21]. However, besides Newton-type algorithms, there are other iterative algorithms again to solve AREs with a sign indefinite quadratic term, for example, the matrix sign function method (see [15, 33, 35]). However, there are also disadvantages when the matrix sign function method is used to solve AREs, for example, when the eigenvalues of the corresponding Hamiltonian matrix of a given ARE are close to the imaginary axis, this method will perform poorly or even fail.

As noted above, in the work presented in this chapter, we reduce the problem of solving a generic Riccati equation with a sign indefinite quadratic term to one of generating successive iterations of solutions of conventional H_2 AREs with a negative semidefinite quadratic term (each of which is then amenable to the Kleinman algorithm). Consequently, we are reducing a Riccati equation that has no straightforwardly initialized iterative scheme for its solution to a number of successive iterations of Riccati equations, each of which can (if desired) be solved by an existing iterative scheme (e.g., the Kleinman algorithm).

Although linear optimal control theory, as well as linear H_∞ control theory, has been well developed in the past decades, matters become more complicated when a nonlinear control system is considered. For example, in nonlinear optimal control, HJB equations may need to be solved to obtain an optimal control law. However, HJB equations are first-order, nonlinear partial differential equations that have been proven to be impossible to solve in general and are often very difficult to solve even for specific nonlinear systems. Since these equations are difficult to solve analytically, there has been much research directed toward approximating their solutions. For example, the technique of successive approximation in policy space [8–10] can be used to approximate the solutions of HJB equations iteratively. In fact, it can be shown (see [29]) that the technique of policy space iteration can be used to replace a nonlinear HJB partial differential equation by a sequence of linear partial differential equations. Also, in some sense, the iterative procedure to solve HJB equations in [29] is a generalization of the Kleinman algorithm in [24], since both of them obtain solutions by constructing a sequence of monotonic functions

or matrices while the algorithm in [29] can be used in more general cases than just the LQ problem.

In nonlinear H_∞ control, given a disturbance attenuation level $\gamma > 0$, in order to solve the H_∞ suboptimal control problem, one needs to solve Hamilton–Jacobi–Bellman–Isaacs (HJBI) equations. It is clear that HJBI equations are generally more difficult to solve than HJB equations, since the disturbance inputs are additionally reflected in HJBI equations. Recall that the Riccati equation algorithm to be presented will reduce an ARE with an indefinite quadratic term to a sequence of AREs with a negative semidefinite quadratic term, which are more easily solved by an existing algorithm (e.g., the Kleinman algorithm). If we regard HJB equations as the general version of AREs with a negative semidefinite quadratic term and HJBI equations as the general version of AREs with an indefinite quadratic term, then the question arising here is, "Can we approximate the solution of an HJBI equation by obtaining the solutions of a sequence of HJB equations and thereby extend the recursive H_∞ Riccati algorithm to nonlinear control systems?" In this chapter, we will answer this question to some degree, that is, we extend the Riccati algorithm for a specific class of nonlinear control systems and develop an iterative procedure to solve a broad class of HJBI equations associated with the nonlinear H_∞ control problem. It is important to note that others have made a direct attack on HJBI equations using single and double iterations, but their methods do not allow a simple initialization of the algorithms, which is a severe disadvantage (see [36] and [7]). To implement the algorithms in [36] and [7], one has to choose a stabilizing control law achieving the prescribed attenuation level, which is not always straightforward to obtain.

Besides the advantages mentioned above, our algorithm can be expected to have a higher accuracy and numerical stability than existing algorithms to solve HJBI equations since our algorithm in the linear time-invariant case (i.e., solving H_∞ algebraic Riccati equations) has shown higher accuracy and numerical reliability, see Example 2 in Section 3.5 for a demonstration of this.

The notation is as follows: $\mathbb{R}$ denotes the set of the real numbers; $\mathbb{R}^+$ denotes the set of the nonnegative numbers; $(\cdot)^T$ denotes the transpose of a vector or a matrix; $\overline{\sigma}(\cdot)$ denotes the maximum singular value of a matrix; $\mathbb{Z}$ denotes the set of integers with $\mathbb{Z}_{\geq a}$ denoting the set of integers greater or equal to $a \in \mathbb{R}$; $\mathbb{R}^n$ denotes an n-dimensional Euclidean space; $\mathbb{S}^{n\times n}$ denotes the set of n-dimensional real symmetric matrices. Let $X \in \mathbb{R}^{n\times n}$ be a real matrix, then $X \geq 0$ means that X is positive semidefinite; Let $X,\ Y \in \mathbb{R}^{n\times n}$ be two positive semidefinite matrices, then $X \geq Y$ means that the matrix $X - Y$ is positive semidefinite (i.e., $X - Y \geq 0$). Let $P_k \in \mathbb{R}^{n\times n}$ be a matrix sequence for $k \in \mathbb{Z}_{\geq 0}$, if $P_k \geq 0$ and $P_{k+1} \geq P_k$ for all $k \in \mathbb{Z}_{\geq 0}$, then the sequence P_k is called monotonically non-decreasing.

For a given control system, denote the state space by $\mathbb{X} \subseteq \mathbb{R}^n$, the set of control input values by $\mathbb{U} \subseteq \mathbb{R}^m$, the set of disturbance input values by $\mathbb{W} \subseteq \mathbb{R}^q$, and the set of output values by $\mathbb{Y} \subseteq \mathbb{R}^p$. Moreover, define $\mathbb{X}_0$ as a neighborhood of the origin in $\mathbb{R}^n$, $\mathbb{U}_0$ as a neighborhood of the origin in $\mathbb{R}^m$,

$\mathbb{W}_0$ as a neighborhood of the origin in $\mathbb{R}^q$, and $\mathbb{Y}_0$ as a neighborhood of the origin in $\mathbb{R}^p$. Define the function space $\mathcal{X}_0$ as follows:

$$\mathcal{X}_0 = \left\{ x : \mathbb{R}^+ \to \mathbb{X}_0 \;\middle|\; \int_{t_0}^{t_1} \|x(t)\|^2 dt < \infty \quad \forall t_0, t_1 \in \mathbb{R}^+ \right\}.$$

Function spaces $\mathcal{U}_0$, $\mathcal{W}_0$, and $\mathcal{Y}_0$ are defined similarly as $\mathcal{X}_0$.

A matrix is said to be *Hurwitz* if all of its eigenvalues have negative real parts.

2 Solving the LQ Problem by the Kleinman Algorithm

As noted in the previous section, for the linear time-invariant case of our algorithm, we replace the problem of solving an H_∞ Riccati equation by the problem of solving a sequence of H_2 Riccati equations; then each of these H_2 Riccati equations can be solved by the Kleinman algorithm. The Kleinman algorithm, originally used to solve the LQ problem, will be reviewed in this section. By using the Kleinman algorithm, we can replace the problem of solving an H_2 Riccati equation by the problem of solving a sequence of Lyapunov equations; then each Lyapunov equation can be solved by existing numerical algorithms. By doing so, we transfer a nonlinear matrix equation (an H_2 Riccati equation) into a sequence of linear equations (Lyapunov equations).

Consider a continuous-time linear system described by

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned}$$

with a cost functional defined as

$$J = \int_0^\infty \left(x^T C^T C x + u^T u \right) dt.$$

In the LQ problem, it is well known that the feedback control law that minimizes the value of the cost J is

$$u = -B^T K x,$$

with K solving

$$0 = A^T K + KA - KBB^T K + C^T C. \tag{2}$$

If $(A,\ B)$ is stabilizable and $(C,\ A)$ is detectable, it can be shown that (2) has a unique stabilizing solution. In such a situation, the Kleinman algorithm can be used to solve (2). The Kleinman algorithm to solve (2) is given as follows:

1. choose an initial stabilizing state feedback gain L_0;
2. $A_0 := A - BL_0$;

3. obtain V_0 by solving the Lyapunov equation

$$0 = A_0^T V_0 + V_0 A_0 + L_0^T L_0 + C^T C;$$

4. $L_k := B^T V_{k-1}, \quad k = 1, 2, \ldots;$
5. $A_k := A - BL_k, \quad k = 1, 2, \ldots;$
6. obtain V_k by solving the following Lyapunov equation:

$$0 = A_k^T V_k + V_k A_k + L_k^T L_k + C^T C, \quad k = 1, 2, \ldots.$$

It can be proved that the sequence V_k is monotonically non-increasing and converges to the stabilizing solution K of (2).

In H_2 control, typically, the Kleinman algorithm for solving H_2-type AREs with a negative semidefinite quadratic term is well suited as a "second iterative stage" refinement to achieve the prescribed accuracy for the stabilizing solution of the ARE. For example, if an approximate stabilizing solution is known (e.g., one observes using the Schur method), and this is stabilizing, then one to two iterations are sufficient to achieve the limiting accuracy (because of the guaranteed final quadratic rate of convergence of a typical Newton algorithm). Now as noted, the Kleinman algorithm reduces a quadratic (Riccati) equation (with a negative semidefinite quadratic term) to several successive iterations of linear (Lyapunov) equations; the complexity of solving algebraic Lyapunov and Riccati equations with sound numerical methods (e.g., Schur form-based reductions) is $O(n^3)$ for both. When Schur form-based reductions are used to solve Lyapunov equations, the computation for such (Schur form-based) reductions needs about $25n^3$ flops (see [18]), where 1 flop equals 1 addition/subtraction or 1 multiplication/division. About $7n^3$ flops are necessary to solve the reduced equation and to compute the solution. The basic method is described in [6]. For the solutions of AREs, the Schur approach of Laub [28] requires $240n^3$ flops of which $25(2n)^3$ flops are required to reduce a $2n \times 2n$ Hamiltonian matrix to real Schur form and the rest accounts for the computation of the eigenvalues and solving a linear equation of order n (i.e., $\frac{5}{3}n^3$ flops). Consequently, both Riccati and Lyapunov equations require $O(n^3)$ computations. Hence the advantage of iterative schemes such as the Kleinman algorithm (which will require several Lyapunov equations to be solved, typically) is not always the speed of computation, but rather it is the numerical reliability of the computations to reach the prescribed accuracy of a solution.

3 Solving H_∞ Riccati Equations

This section includes five subsections: (1) the summarizing theorem; (2) an algorithm to solve H_∞ Riccati equations; (3) an examination of the rate of convergence of the algorithm; (4) a game theoretic interpretation of the algorithm, (5) numerical examples.

3.1 The Summarizing Theorem

In this subsection, we will restrict attention to the unique stabilizing solution Π for the following ARE:

$$0 = \Pi A + A^T\Pi - \Pi(B_2B_2^T - B_1B_1^T)\Pi + C^TC, \tag{3}$$

where A, B_1, B_2, C are real matrices with compatible dimensions. Note that stabilizing solutions to AREs are always unique (see [37]) when they exist, but for AREs with a sign indefinite quadratic term, the unique stabilizing solution Π may not always be positive semidefinite. Since our interest arises from AREs used for H_∞ control, in this case, in order to obtain an H_∞ controller, we need to solve AREs with a sign indefinite quadratic term and the stabilizing solutions of these AREs are also required to be positive semidefinite if such an H_∞ controller exists. So we focus on a unique stabilizing solution to (3) that happens to be also positive semidefinite when this exists. The algorithm we will propose in Section 3.2 has two aspects: (1). Check the existence of the unique stabilizing solution, which is also positive semidefinite, of (3) and (2). Construct the unique stabilizing solution, which is also positive semidefinite, of (3) if such a solution exists.

Motivated by the right-hand side of (3), we define a function F which will be used in our summarizing theorem:

$$\begin{aligned} F:\ \mathbb{R}^{n\times n} &\longrightarrow \mathbb{R}^{n\times n} \\ P &\longmapsto PA + A^TP - P(B_2B_2^T - B_1B_1^T)P + C^TC. \end{aligned} \tag{4}$$

In this section, we set up the summarizing theorem by constructing two positive semidefinite matrix series P_k and Z_k, and we also prove that the series P_k is monotonically non-decreasing and converges to the unique stabilizing solution Π (which is also positive semidefinite) of ARE (3) if such a solution exists.

Theorem 1. (The summarizing theorem) *Let A, B_1, B_2, C be real matrices with compatible dimensions. Suppose that (C, A) has no unobservable modes on the $j\omega$-axis and (A, B_2) is stabilizable, define $F : \mathbb{R}^{n\times n} \longrightarrow \mathbb{R}^{n\times n}$ as in (4). Suppose there exists a stabilizing solution Π, which is also positive semidefinite, of ARE (3).*
Then

(I) *two square matrix series Z_k and P_k can be defined for all $k \in \mathbb{Z}_{\geq 0}$ recursively as follows*:

$$P_0 = 0, \tag{5}$$

$$A_k = A + B_1B_1^TP_k - B_2B_2^TP_k, \tag{6}$$

$Z_k \geq 0$ is the unique stabilizing solution of

$$0 = Z_kA_k + A_k^TZ_k - Z_kB_2B_2^TZ_k + F(P_k), \tag{7}$$

$$P_{k+1} = P_k + Z_k; \tag{8}$$

(II) *the two series* P_k *and* Z_k *in part* (I) *have the following properties*:
(1) $(A + B_1 B_1^T P_k, B_2)$ *is stabilizable* $\forall k \in \mathbb{Z}_{\geq 0}$.
(2) $F(P_{k+1}) = Z_k B_1 B_1^T Z_k \ \forall k \in \mathbb{Z}_{\geq 0}$.
(3) $A + B_1 B_1^T P_k - B_2 B_2^T P_{k+1}$ *is Hurwitz* $\forall k \in \mathbb{Z}_{\geq 0}$.
(4) $\Pi \geq P_{k+1} \geq P_k \geq 0 \ \forall k \in \mathbb{Z}_{\geq 0}$;

(III) *the limit*

$$P_\infty := \lim_{k \to \infty} P_k$$

exists with $P_\infty \geq 0$. *Furthermore,* $P_\infty = \Pi$ *is the unique stabilizing solution of ARE* (3)*, which is also positive semidefinite.*

Proof. See [26, 27]. □

The following corollary gives a condition under which there does *not* exist a stabilizing solution $\Pi \geq 0$ to $F(\Pi) = 0$. This is useful for terminating the recursion in finite iterations. If there does not exist a stabilizing solution $\Pi \geq 0$ to (3), there are two possible situations in Theorem *1*: (1) The stabilizability condition (II1) fails at some iteration; and (2) The sequence P_k in Theorem 1 diverges to infinity.

Corollary 1. *Let* A, B_1, B_2, C *be real matrices with compatible dimensions. Suppose that* (C, A) *has no unobservable modes on the* $j\omega$*-axis and* (A, B_2) *is stabilizable, and let* $\{P_k\}$ *and* $F : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}^{n \times n}$ *be defined as in Theorem 1. If* $\exists k \in \mathbb{Z}_{\geq 0}$ *such that* $(A + B_1 B_1^T P_k, B_2)$ *is not stabilizable, then there does not exist a stabilizing solution* $\Pi \geq 0$ *to* $F(\Pi) = 0$.

Proof. Restatement of Theorem 1, implication (II1). □

3.2 Algorithm

Let A, B_1, B_2, C be real matrices with compatible dimensions and $\Delta > 0$ be a specified tolerance. Suppose that (C, A) has no unobservable modes on the $j\omega$-axis and (A, B_2) is stabilizable. Then an iterative algorithm for finding the positive semidefinite stabilizing solution of (3), when it exists, is given as follows:

1. Let $P_0 = 0$ and $k = 0$.
2. Set $A_k = A + B_1 B_1^T P_k - B_2 B_2^T P_k$.
3. Construct (for example, using the Kleinman algorithm in [24], though this is not necessary) the unique real symmetric stabilizing solution $Z_k \geq 0$ which satisfies

$$0 = Z_k A_k + A_k^T Z_k - Z_k B_2 B_2^T Z_k + F(P_k), \tag{9}$$

   where $F : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}^{n \times n}$ is defined in (4).
4. Set $P_{k+1} = P_k + Z_k$.
5. If $\overline{\sigma}(B_1^T Z_k)^2 < \Delta$, then set $\Pi = P_{k+1}$ and exit. Otherwise, go to step 6.

6. If $(A+B_1B_1^TP_{k+1}, B_2)$ is stabilizable, then increment k by 1 and go back to step 2. Otherwise, exit as there does not exist a real symmetric stabilizing solution $\Pi \geq 0$ satisfying $F(\Pi) = 0$.

From Corollary 1 we see that if the stabilizability condition in step 6 fails at some $k \in \mathbb{Z}_{\geq 0}$, then there does not exist a stabilizing solution $\Pi \geq 0$ to $F(\Pi) = 0$ and the algorithm should terminate (as required by step 6). But when this stabilizability condition is satisfied $\forall k \in \mathbb{Z}_{\geq 0}$, construction of the series P_k and Z_k is always possible and either P_k converges to Π (which is captured by step 5) or P_k just diverges to infinity, which again means that there does not exist a stabilizing solution $\Pi \geq 0$ to $F(\Pi) = 0$.

Remark 1. It is worth pointing out that when the Kleinman algorithm is used to solve (9), how to stop the Kleinman iteration can be an important issue. In fact, a simple criterion to stop the Kleinman iteration is to compute the residue of (9). When the residue of the right-hand side terms of (9) is small enough, the Kleinman iteration should be stopped.

Remark 2. For a numerical method to check the stabilizability of a matrix pair in step 6, one can refer to the staircase algorithm in [34]. In the staircase algorithm, some SVDs are performed to formulate a staircase form of the matrix pair, then the stabilizability of the matrix pair is checked based on the staircase form. Since the computational complexity of SVD is $O(n^3)$ (see [18]), the computational complexity of the staircase algorithm is also $O(n^3)$.

3.3 Rate of Convergence

The following theorem states that the local rate of convergence of the algorithm given in Section 3.2 is quadratic.

Theorem 2. *Given the suppositions of Theorem* 1*, and two series* P_k*,* Z_k *as defined in Theorem* 1 *Part* I*, then there exists a* $\theta > 0$ *such that the rate of convergence of the series* P_k *is quadratic in the region* $\|P_k - \Pi\| < \theta$*.*

Proof. We prove the rate of convergence of P_k by proving the rate of convergence of Z_k. Let $\acute{A}_k = A + B_1B_1^TP_k - B_2B_2^TP_{k+1}$ and $\tilde{A} = A + B_1B_1^T\Pi - B_2B_2^T\Pi$. Note that $\acute{A}_k$ is Hurwitz (see Theorem 1 Part II) and $\tilde{A}$ is Hurwitz since Π is the stabilizing solution of (3) (see [39]), and let $\tilde{Y}$ and $\acute{Y}_k$ be uniquely defined by

$$0 = \tilde{Y}\tilde{A} + \tilde{A}^T\tilde{Y} + I, \tag{10}$$

$$0 = \acute{Y}_k\acute{A}_k + \acute{A}_k^T\acute{Y}_k + I, \tag{11}$$

where I is the identity matrix with appropriate dimensions. The matrices $\tilde{Y}$ and $\acute{Y}_k$ are positive definite because of the stability properties of $\acute{A}_k$ and $\tilde{A}$. Since $\tilde{Y}$ and $\acute{Y}_k$ are uniquely defined and $\lim_{k\to\infty} \acute{A}_k = \tilde{A}$, then $\lim_{k\to\infty} \acute{Y}_k = \tilde{Y}$, and thus for any small $\gamma > 0$, $\exists K_1 \in \mathbb{Z}_{\geq 0}$ such that

$$\overline{\sigma}(\acute{Y}_k - \tilde{Y}) \leq \gamma \quad \forall k \geq K_1.$$

This implies that

$$\overline{\sigma}(\acute{Y}_k) \leq \overline{\sigma}(\tilde{Y}) + \gamma \quad \forall k \geq K_1. \tag{12}$$

Now, define a monotonically non-increasing sequence ε_k by

$$\varepsilon_k = \sup_{m \geq k} \overline{\sigma}(Z_m).$$

From (7) and Theorem 1, we have $\forall k \in \mathbb{Z}_{\geq 1}$:

$$0 = Z_k A_k + A_k^T Z_k - Z_k B_2 B_2^T Z_k + Z_{k-1} B_1 B_1^T Z_{k-1}, \tag{13}$$

which can be equivalently rewritten as follows:

$$0 = Z_k \acute{A}_k + \acute{A}_k^T Z_k + Z_k B_2 B_2^T Z_k + Z_{k-1} B_1 B_1^T Z_{k-1}. \tag{14}$$

Now, there exists $\eta > 0$ (e.g., $\eta = 4\max\{\overline{\sigma}(B_1)^2, \overline{\sigma}(B_2)^2\}$), independent of k, such that

$$Z_k B_2 B_2^T Z_k + Z_{k-1} B_1 B_1^T Z_{k-1} \leq \eta \varepsilon_{k-1}^2 I \quad \forall k \in \mathbb{Z}_{\geq 1}. \tag{15}$$

Multiplying $(\eta \varepsilon_{k-1}^2)$ on each side of (11), we obtain

$$0 = (\eta \varepsilon_{k-1}^2) \acute{Y}_k \acute{A}_k + (\eta \varepsilon_{k-1}^2) \acute{A}_k^T \acute{Y}_k + (\eta \varepsilon_{k-1}^2) I. \tag{16}$$

Then subtracting (14) from (16), we obtain

$$\begin{aligned} 0 = {} & (\eta \varepsilon_{k-1}^2 \acute{Y}_k - Z_k) \acute{A}_k + \acute{A}_k (\eta \varepsilon_{k-1}^2 \acute{Y}_k - Z_k) + \eta \varepsilon_{k-1}^2 I \\ & - (Z_k B_2 B_2^T Z_k + Z_{k-1} B_1 B_1^T Z_{k-1}). \end{aligned} \tag{17}$$

Now note that $\acute{A}_k$ is Hurwitz and the inequality (15) holds, then by (17) we have (see Lemma 3.18 in [39])

$$\eta \varepsilon_{k-1}^2 \acute{Y}_k \geq Z_k. \tag{18}$$

Since (18) holds, $\overline{\sigma}(Z_k) \leq \eta \varepsilon_{k-1}^2 \overline{\sigma}(\acute{Y}_k)$ (see [18]). Hence $\forall k \geq K_1 \geq 1$,

$$\begin{aligned} \varepsilon_k = \sup_{m \geq k} \overline{\sigma}(Z_m) &\leq \sup_{m \geq k} \left[\eta \varepsilon_{m-1}^2 \overline{\sigma}(\acute{Y}_m) \right] \\ &\leq \eta(\overline{\sigma}(\tilde{Y}) + \gamma) \sup_{m \geq k} \varepsilon_{m-1}^2 = \eta(\overline{\sigma}(\tilde{Y}) + \gamma) \varepsilon_{k-1}^2. \end{aligned}$$

Now let $M := \eta(\overline{\sigma}(\tilde{Y}) + \gamma)$ and define $\delta_k := \frac{M}{c} \varepsilon_k$ for $0 < c < 1$, then

$$\delta_k = \frac{M}{c} \varepsilon_k \leq \frac{M^2}{c} \varepsilon_{k-1}^2 = c \frac{M^2}{c^2} \varepsilon_{k-1}^2 = c \delta_{k-1}^2.$$

Thus, $\forall k \geq K_1 \geq 1$ such that $\delta_{k-1} < 1$, we obtain a quadratic rate of convergence, which concludes the proof. □

3.4 Game Theoretic Interpretation of the Algorithm

In this subsection, for the purpose of motivation, interpretation, and further research, a game theoretic interpretation of the algorithm will be given. At the same time, we will also note that this interpretation is closely linked to an optimal control concept of approximation in policy space [8–10]. In this section, we will show that a game theory performance index can be approximated by a series of successive optimal control cost functions. At each iteration, the optimal policy is found to minimize the corresponding cost functions. With the increment of each iteration, the optimal policies approach the final optimum and the saddle point of the cost functions approaches the saddle point of the game theory performance index. In fact, it can be shown (see [29]) that the technique of policy space iteration can be used to replace nonlinear Hamilton–Jacobi partial differential equations by a sequence of linear partial differential equations (even when the approximations and the optimal feedback law are nonlinear functions of the state). This is important because it is difficult [29] to solve Hamilton–Jacobi equations directly to obtain optimal feedback control in many cases.

Consider the dynamical system

$$\dot{x} = Ax + B_1 w + B_2 u \tag{19}$$

with the game theory performance index

$$J(x_0, u, w) = \int_0^\infty \left(u^T u + x^T C^T C x - w^T w\right) dt, \tag{20}$$

where x_0 denotes the initial state of the system, x is the state vector, u denotes the control input, w denotes the disturbance input, and A, B_1, B_2, C are given real matrices with compatible dimensions and appropriate stabilizability/detectability conditions. In this game, u minimizes the cost function J while w maximizes it. It is well known [19] that the optimal control law and the worst case disturbance (a saddle point of $J(x_0, u, w)$) are given by

$$u_{\text{optimal}} = -B_2^T \Pi x, \tag{21}$$

$$w_{\text{worst}} = B_1^T \Pi x, \tag{22}$$

where $\Pi \geq 0$ is the unique stabilizing solution to (3). See [19] for more details on such game theory problems.

Let us now propose a heuristic induction which gives a game theoretic interpretation to our proposed algorithm. Suppose that at iteration k we have a trial control law $u_k = -B_2^T P_k x$ with P_k defined as in Theorem 1 Part I. Then we set $w_k = B_1^T P_k x$. Note that this is NOT the worst case w_k corresponding to $u_k = -B_2^T P_k x$ (unless $P_k = \Pi$), but it is a strategy we wish to impose since it will connect the heuristic ideas of this section to the earlier algorithm. The choice is also motivated by what happens at the optimum, as $P_k \longrightarrow \Pi$ when

$k \longrightarrow \infty$. With this choice of w fixed, we now wish to find a new optimal control, i.e., u now has to minimize the following LQ cost function:

$$J_k(x_0, u) = \int_0^\infty \left(u^T u + x^T C^T C x - x^T P_k B_1 B_1^T P_k x\right) dt, \tag{23}$$

subject to

$$\dot{x} = (A + B_1 B_1^T P_k)x + B_2 u, \tag{24}$$

where $w_k = B_1^T P_k x$ has been substituted in (19) and (20) to yield the above problem. We first consider the following equation:

$$0 = W_k A_k + A_k^T W_k - W_k B_2 B_2^T W_k + F(P_k), \tag{25}$$

where $A_k = A + B_1 B_1^T P_k - B_2 B_2^T P_k$. Since this is the same equation as (9), we conclude that W_k satisfies the same equation as Z_k. Now let $\Lambda_{k+1} = P_k + W_k$, then existence of W_k is equivalent to existence of Λ_{k+1}. Necessary and sufficient conditions for the existence of W_k are the following: (A_k, B_2) is stabilizable and $(F(P_k), A_k)$ has no unobservable modes on the jw-axis. These conditions were analyzed in the proof of Theorem 1 and were shown to be fulfilled via the existence of the stabilizing solution $\Pi \geq 0$ to (3). Under appropriate conditions, the LQ problem defined by (23) and (24) has an optimal solution for u given by

$$u_{k+1} = -B_2^T \Lambda_{k+1} x,$$

where Λ_{k+1} is the unique stabilizing solution to

$$\begin{aligned} 0 =& \Lambda_{k+1}(A + B_1 B_1^T P_k) + (A + B_1 B_1^T P_k)^T \Lambda_{k+1} \\ & - \Lambda_{k+1} B_2 B_2^T \Lambda_{k+1} + (C^T C - P_k B_1 B_1^T P_k). \end{aligned} \tag{26}$$

We will now show that Λ_{k+1} actually equals P_{k+1} as defined in Theorem 1 Part I. To do this, we will equivalently show that $\Lambda_{k+1} - P_k$ equals Z_k. Using $\Lambda_{k+1} = P_k + W_k$ in (26), we have

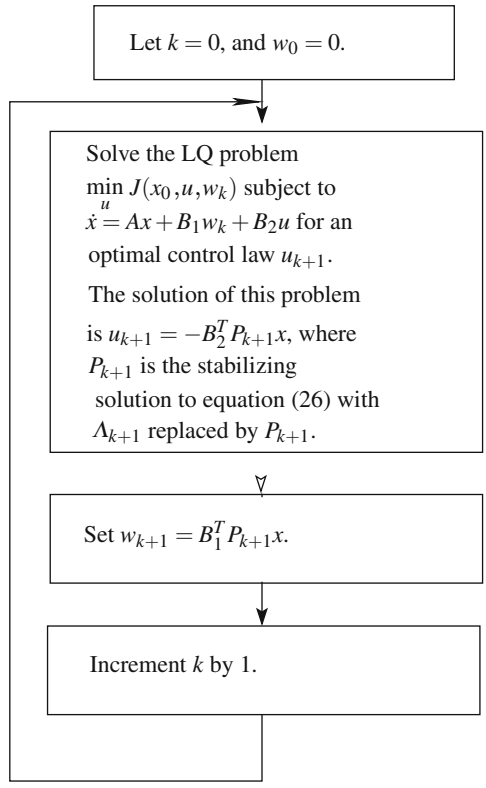
$$\begin{aligned} 0 =& P_k(A + B_1 B_1^T P_k) + W_k(A + B_1 B_1^T P_k) \\ & + (A + B_1 B_1^T P_k)^T P_k + (A + B_1 B_1^T P_k)^T W_k \\ & - P_k B_2 B_2^T P_k - W_k B_2 B_2^T P_k - P_k B_2 B_2^T W_k \\ & - W_k B_2 B_2^T W_k + C^T C - P_k B_1 B_1^T P_k. \end{aligned} \tag{27}$$

The terms above independent of W_k are

$$P_k A + A^T P_k + P_k(B_1 B_1^T - B_2 B_2^T)P_k + C^T C,$$

which are equal to $F(P_k)$. Then (27) reduces to (25). Now note that $\Lambda_{k+1} = P_k + W_k$ is a stabilizing solution to (26), meaning that $A + B_1 B_1^T P_k - B_2 B_2^T (W_k + P_k)$ is Hurwitz. But $A + B_1 B_1^T P_k - B_2 B_2^T (W_k + P_k) = (A_k - B_2 B_2^T W_k)$. Hence W_k also makes $A_k - B_2 B_2^T W_k$ Hurwitz. Since Z_k is the

unique stabilizing solution to (9) and since W_k satisfies the same equation and is also stabilizing, we conclude that $W_k = Z_k$ (see [37]), thereby in turn giving $\Lambda_{k+1} = P_{k+1}$. Since the optimal control law that minimizes cost function (23) subject to constraint (24) is $u_{k+1} = -B_2^T P_{k+1} x$, we now set (according to our game plan) $w_{k+1} = B_1^T P_{k+1} x$ and proceed in this manner as outlined in the following chart:



This heuristic game plan converges to the optimal control (21) and the worst case disturbance (22), thereby giving a game theoretic interpretation to the proposed algorithm.

3.5 Numerical Examples

In this subsection, three examples are given. Example 1 provides a random test to compare our algorithm with the MATLAB command CARE and shows that our algorithm has good efficiency and accuracy when compared with CARE. Example 2 shows that our algorithm still works well when some other approaches (such as the MATLAB command CARE, the Schur method of [28],

and the matrix sign function method of [33]) do not work. Example 3 demonstrates that there in fact does not exist a stabilizing solution $\Pi \geq 0$ to (3) when the stabilizability condition in step 6 of the algorithm is not satisfied, and the sequence P_k diverges quadratically in such a situation.

Example 1

In this example, to show the efficiency and accuracy of our algorithm compared with the MATLAB command CARE, we have a random test including 200 samples (100 examples for the specified tolerance $\Delta = 0.1$ and 100 examples for the specified tolerance $\Delta = 0.01$). Note that the MATLAB command CARE always works well in this example (i.e., the residue for the solutions of AREs obtained by using CARE is always small).

The test procedure is as follows:

1. consider a state-space representation for a dynamic system
$$\dot{x} = Ax + Bu,$$
$$y = Cx + Du,$$
where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times(p+r)}$, $C \in \mathbb{R}^{m\times n}$, and u, x, y are input, state, output, respectively;
2. set the example number $i = 0$;
3. choose n, m, p, r randomly and uniformly among the integers from 1 to 100;
4. generate a random system by using MATLAB command sysrand with n, m, p, r obtained in step 3 and obtain A, B, C, D by MATLAB command unpck;
5. partition the matrix $B = [B_1 \quad B_2]$, where $B_1 \in \mathbb{R}^{n\times p}$ and $B_2 \in \mathbb{R}^{n\times r}$;
6. try MATLAB command CARE to solve the corresponding ARE with $E = I_n, S = 0, B = [B_1 \quad B_2], Q = C^TC, R = \begin{pmatrix} -I_p & 0 \\ 0 & I_r \end{pmatrix}$, where I_n, I_p, and I_r are identity matrices with dimensions n, p, r, respectively. If there does not exist a stabilizing solution to the ARE, go back to step 3;
7. use our algorithm to solve this ARE. For our algorithm, the iteration will be stopped when $\overline{\sigma}(B_1^T Z_k)^2 < \Delta = 0.1$, where Z_k is the matrix sequence defined in Theorem 1 Part I;
8. let the solution of this ARE obtained by our algorithm be X_1 and the solution of this ARE obtained by the MATLAB command CARE be X_2;
9. set $i = i + 1$, and let $T_i = \frac{\|X_1 - X_2\|}{\|X_2\|}$;
10. repeat steps 3–9 until $i = 100$;
11. replace $\Delta = 0.1$ in step 7 by $\Delta = 0.01$, set $i = 0$, and repeat steps 3–10.

For each random example in Tables 1 and 2, "Iterations" indicates the number of necessary iterations to obtain the specified tolerance Δ in step 7; "$O(T_i)$" is the order of magnitude of T_i (defined in step 9) or the order of magnitude

of the normalized comparison error between the stabilizing solutions obtained by our algorithm and the stabilizing solutions obtained by the MATLAB command CARE; "Number of examples" means the number of random examples that required "Iterations" number to converge. From Tables 1 and 2, we can conclude that our proposed algorithm works well in most cases with good efficiency (only three to five iterations in most examples) and accuracy when compared with the MATLAB command CARE.

Table 1. Illustration of iteration count of our algorithm and accuracy comparison with CARE for 100 random examples ($\overline{\sigma}(B_1^T Z_k)^2 < \Delta = 0.1$)

Iterations / $O(T_1)$	2	3	4	5	6	7	8
10^{-4}	1	5	4		1	1	
10^{-5}		7	6	7	1	2	3
10^{-6}		13	10	3	2		
10^{-7}		5	3	3	1	1	
10^{-8}		1	4	7			
10^{-9}		4	3				
10^{-10}			2				
10^{-11}							
10^{-12}							
10^{-13}							
Number of examples	1	35	32	20	5	4	3

A summary of the results in Tables 1 and 2 is as follows:

1. In both Tables 1 and 2, our proposed algorithm converges in ONLY three to five iterations in most examples.
2. When the prescribed tolerance is $\Delta = 0.1$, we have $T_i < 10^{-3}$ for each random example in Table 1; when the prescribed tolerance is $\Delta = 0.01$, we have $T_i < 10^{-5}$ for each random example in Table 2.

Example 2

The following example illustrates that the proposed algorithm works well when other traditional methods fail. This example is a slight modification of Example 6 in [28]. Choose the matrices $A \in \mathbb{R}^{21\times 21}, B_1 \in \mathbb{R}^{21\times 1}, B_2 \in \mathbb{R}^{21\times 1}, C \in \mathbb{R}^{21\times 21}$ in (3) as follows:

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ & \ddots & \ddots & \bigcirc & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ & \bigcirc & & \ddots & 1 \\ 0 & & \cdots & & 0 \end{pmatrix}, \; B_1 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \delta \\ 0 \end{pmatrix}, \; B_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

Table 2. Illustration of iteration count of our algorithm and accuracy comparison with CARE for 100 random examples ($\overline{\sigma}(B_1^T Z_k)^2 < \Delta = 0.01$)

Iterations / $O(T_1)$	2	3	4	5	6	7	8	9	10	11
10^{-4}										
10^{-5}										
10^{-6}		1	2	1	2	1				
10^{-7}	1	6	7	5	2	1	1		2	1
10^{-8}	1	6	2	4	1	1	2	1		
10^{-9}		2	11	1	1	1				
10^{-10}		1	5	2	1	2	1			
10^{-11}			1	4	1	2		1		1
10^{-12}			3		1					
10^{-13}			4	1	2					
Number of examples	2	16	35	18	11	8	4	2	2	2

$$C = \text{diag}\{1, 0, \cdots, 0\},$$

where $\delta = 10^{-2}$ is the introduced modification. In this example, the Schur method in [28] does not produce an accurate result, similarly to the MATLAB command CARE. The algorithm proposed by Kleinman in [24] cannot be used because the term $(B_2 B_2^T - B_1 B_1^T)$ in the Riccati equation (3) is not positive semidefinite. However, the algorithm proposed in Section 3.2 easily gives the solution with the specified accuracy as will be shown next.

First we attempt the Schur method in [28] with this example. Let F be defined as in (4), and S_1 be the solution to (3) by using this Schur method. We evaluate the accuracy of the solution S_1 by calculating $\rho[F(S_1)]$. The smaller $\rho[F(S_1)]$ is, the closer S_1 is to the correct solution. After calculation, we obtain $\rho[F(S_1)] = 1.9802 \times 10^3$ which is far too large. Thus, we can conclude that the Schur method in [28] fails to give an accurate solution in this example. Similarly, let S_2 be the solution obtained by the MATLAB command CARE. For this solution, we can obtain $\rho[F(S_2)] = 1.9811 \times 10^3$ which again is too large. So we conclude that MATLAB command CARE also fails to give a solution in this example. If we were to try to refine the very coarse solution obtained by the Schur method in [28] using Kleinman's method in [24], this too fails as this algorithm diverges with each iteration (as expected). This can be shown as follows: let X_k with $k \in \mathbb{Z}_{\geq 1}$ denote the iterative series produced by the Kleinman algorithm, then we obtain $\rho[F(X_1)] = 5.7083 \times 10^2$, $\rho[F(X_2)] = 5.9959 \times 10^2$, ... , $\rho[F(X_{20})] = 8.2965 \times 10^7$, ... , $\rho[F(X_{100})] = 6.6206 \times 10^9$. If we use the matrix sign function method in [33] to solve this ARE and let Q be the corresponding solution, we obtain $\rho[F(Q)] = 1.235 \times 10^8$ which is again far too large.

However, when we use our proposed algorithm, we note that a unique stabilizing solution $P_4 > 0$ to (3) can be found with limiting accuracy after only four iterations with $\rho[F(P_4)] = \overline{\sigma}(B_1^T Z_3)^2 = 2.9205 \times 10^{-5}$.

Example 3

The following example shows that if $(A+B_1B_1^TP_{k+1}, B_2)$ is not stabilizable in step 5 of the algorithm, then there does not exist a stabilizing solution $\Pi \geq 0$ to (3). Choose

$$A = \begin{pmatrix} 1 & 100 \\ 1 & 0 \end{pmatrix},\ B_1 = \begin{pmatrix} 10 \\ 0 \end{pmatrix},\ B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ C = \begin{pmatrix} 1\ 0 \end{pmatrix}.$$

We note that (C, A) has no unobservable modes on the $j\omega$-axis and (A, B_2) is stabilizable. When we run our algorithm, we find that $(A + B_1B_1^TP_1, B_2)$ is not stabilizable after one iteration since

$$P_1 = \begin{pmatrix} 0.3517 & 2.6442 \\ 2.6442 & 22.9964 \end{pmatrix}.$$

This is consistent with the fact that there does not exist a stabilizing solution $\Pi \geq 0$ to (3). In fact, we can find the unique stabilizing solution

$$\Pi = \begin{pmatrix} -0.5744 & -4.9147 \\ -4.9147 & -37.8481 \end{pmatrix},$$

which is clearly not positive semidefinite. Meanwhile, we find that the sequence P_k diverges quadratically on noting that $\rho[F(P_1)] = 7.1156 \times 10^2,\ \rho[F(P_2)] = 5.9180 \times 10^3, \ldots,\ \rho[F(P_{16})] = 4.7489 \times 10^{15}$.

4 Solving an HJB Equation by a Sequence of Linear PDEs

As we indicated in the introduction section, for our algorithm in the nonlinear case, we replace the problem of solving HJBI equations by the problem of solving a sequence of HJB equations. However, HJB equations are first-order, generally nonlinear partial differential equations and difficult to solve in general. In this section, we review the technique of [29]. By using the technique of [29], we can replace the problem of solving an HJB equation by the problem of solving a sequence of linear PDEs; by doing so, we transfer a difficult problem into a sequence of less difficult problems.

Consider a dynamical system represented by

$$\dot{x} = f(x, u, t), \quad x(t_0) = x_0, \tag{28}$$

where the n-vector x is the plant state, f is a continuously differentiable n-vector function, and $u(x,t) \in \mathbb{R}^n \times \mathbb{R}$ is an r-vector function defined on $\mathbb{R}^n \times \mathbb{R}$. The solution of (28) is denoted as $\phi_u(t) = \phi_u(t; x_0, t_0)$.

Let G be a closed subset of $\mathbb{R}^n \times \mathbb{R}$ to which all motions of the system (28) are restricted. Let the target set S be a closed subset of G. In this section, the function u is called an *admissible feedback control law* if

(1) it is continuously differentiable with values $u(x,t)$ belonging to a locally compact set $U_0 \in \mathbb{R}^r$ for all t;
(2) it has the property that when substituted into (28), any motion beginning in $G - S$ reaches S, or approaches S, in a uniform asymptotic manner without leaving G.

The class of functions satisfying the above properties is denoted by U_1. The terminal time $t_1 = t_1(x_0, t_0)$ is defined to be the time when the motion $(\phi_u(t), t)$ becomes a member of S, or, in the asymptotic case, $t_1 = \infty$. Note that in the finite time case, S itself might be simply $\mathbb{R}^n \times t_1$. (Indeed, we shall focus in what follows on this possibility).

In the following, we first consider the situation when t_1 is finite, then consider the situation when t_1 is infinite.

4.1 t_1 Is Finite

The system performance is evaluated by the functional

$$J(x_0, t_0, t_1; u) = \lambda[\phi_u(t_1), t_1] + \int_{t_0}^{t_1} L[\phi_u(\alpha), u(\phi_u(\alpha), \alpha), \alpha] d\alpha, \tag{29}$$

where L and λ are nonnegative scalar, continuously differentiable functions. It is assumed that L is strictly convex in u.

We define

$$V^0(x_0, t_0, t_1) = \inf_{u \in U_1} J(x_0, t_0, t_1; u). \tag{30}$$

Let H be defined as

$$H(x, p, t, u) = \langle f(x, u, t), p \rangle + L(x, u, t), \tag{31}$$

where p is an n-vector, u is an r-vector, and $<,>$ denotes the inner product. Assume that H has a unique absolute minimum for each x, p, and t with respect to the values $u \in U_0$, and let the associated location of minimum be denoted as $c(x, p, t)$. Assuming that c is a continuously differentiable function of x, p, and t, we define the Hamiltonian as

$$H^0(x, p, t) = H(x, p, t, c(x, p, t)) = \min_{u \in U_0} H(x, p, t, u). \tag{32}$$

The HJB equation we consider is

$$0 = V_t + H^0(x, V_x, t) \tag{33}$$

subject to the boundary condition

$$V(x,t_1)=\lambda(x,t_1),$$

where $V(x,t)$ is a scalar function defined on $\mathbb{R}^n\times\mathbb{R}$, $V_t=\frac{\partial V}{\partial t}$, and $V_x=\frac{\partial V}{\partial x}$. It can be shown [23] that if $V(x,t)$ is twice continuously differentiable in all arguments, if it satisfies (33) in G and the boundary condition $V(x,t_1)=\lambda(x,t_1)$ on S, and in addition if the function $u^0(x,t)=c(x,V_x,t)$ is admissible, then $V(x,t)=V^0(x,t)$. It is shown in [29] that for any optimal control problem described by (28), (29), and (30), we have

$$\frac{dJ}{dt}(\phi_u(t;x_0,t_0),t;u)=-L[\phi_u(t;x_0,t_0),u(\phi_u(t;x_0,t_0),t),t], \tag{34}$$

with $V(x,t)=\lambda(x,t)$ or, denoting $\phi_u(t;x_0,t_0)$ by x, we have

$$\dot{J}(x,t;u)=-L(x,u,t), \tag{35}$$

with $V(x,t)=\lambda(x,t)$. Given any optimal feedback control described by (28), (29) and (30), we define $\mathcal{V}$ as the set of all continuously differentiable functions $V:\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}$ such that $V(x,t_1)=\lambda(x,t_1)$ on S. Let $\mathcal{V}^0$ be the subset of $\mathcal{V}$ such that if $u(x,t)=c(x,V_x,t)$ then u is admissible.

The following theorem comes from [29], and it constructs a monotone nonincreasing function sequence which converges to the solution of (33). Meanwhile, it provides a technique of transferring an HJB equation into a sequence of linear PDEs.

Theorem 3. [29] *Define a mapping $T:\mathcal{V}_0\to\mathcal{V}$ for $V\in\mathcal{V}_0$ and $T(V)=J(x,t;u)$ with $u(x,t)=c(x,V_x,t)$. Suppose $T(V^n)\in\mathcal{V}^0$ for $n=1,2,3,\ldots$. If $V^1\in\mathcal{V}^0$ and $V^{n+1}=T(V^n)$, $n=1,2,3,\ldots$, then*

$$V(x,t)\le V^{n+1}(x,t)\le V^n(x,t)\le V^1(x,t), \tag{36}$$

where V is the solution of (33). For every sequence V^n, there exists a function V^* such that $V^n(x,t)\downarrow V^*(x,t)$ pointwise on G. If G is bounded, the convergence is uniform. If T is continuous in $\mathcal{V}_0\subset\mathcal{V}$ and $V^*\in\mathcal{V}^0$, then $V^*=V$. Furthermore, suppose V^1 is given, then the sequence V^n can be recursively obtained by solving the following sequence of linear PDEs:

$$0=\left\langle V_x^{n+1},f(x,c(x,V_x^n,t),t)\right\rangle+V_t^{n+1}+L(x,c(x,V_x^n,t),t),$$

with the boundary condition $V^n(x,t_1)=\lambda(x,t_1)$.

Proof. See [29]. □

4.2 t_1 Is Infinite

In this subsection, we assume that there exists a solution of HJB equation (33) which minimizes the cost function (29) when t_1 is infinite. In such a situation, it is clear that the scalar function L in (29) approaches zero when $t_1\to\infty$.

Evidently, for infinite time problems, key interest centers around stabilizing control laws. This motivates us to make the following assumption, which generalizes controllability/stabilizability and observability/detectability assumptions commonly made for H_2 problems.

Assumption A1 There exists a feedback law for (28) ensuring that the closed-loop system is uniformly asymptotically stable with associated finite performance index, and if a feedback control law ensures that the achieved performance with $t_1 = \infty$ and an arbitrary $x(t_0)$ is finite, then the associated closed-loop system is uniformly asymptotically stable.

The following theorem gives an infinite time version of Theorem 3.

Theorem 4. *Let the mapping $T : \mathcal{V}_0 \to \mathcal{V}$ be as defined in Theorem 3 and suppose V is a solution of (33) with infinite terminal time t_1. Suppose $T(V^n) \in \mathcal{V}^0$ for $n = 1, 2, 3, \ldots$. If $V^1 \in \mathcal{V}^0$ and $V^{n+1} = T(V^n)$, $n = 1, 2, 3, \ldots$, then*

$$V(x,t) \le V^{n+1}(x,t) \le V^n(x,t) \le V^1(x,t), \tag{37}$$

where V is the solution of (33). For every sequence V^n, there exists a function V^* such that $V^n(x,t) \downarrow V^*(x,t)$ pointwise on G. If T is continuous in $\mathcal{V}_0 \subset \mathcal{V}$ and $V^* \in \mathcal{V}^0$, then $V^* = V$. Furthermore, suppose a bounded V^1 is chosen such that the system $\dot{x} = f(x, c(x, V_x^1, t), t)$ is uniformly asymptotically stable, then the sequence V^n can be recursively obtained by solving the following sequence of linear PDEs:

$$0 = \left\langle V_x^{n+1}, f(x, c(x, V_x^n, t), t) \right\rangle + V_t^{n+1} + L(x, c(x, V_x^n, t), t),$$

with the boundary condition

$$\lim_{t_1 \to \infty} V^n(x, t_1) = \lim_{t_1 \to \infty} \lambda(x, t_1),$$

and the closed-loop system $\dot{x} = f(x, c(x, V_x^n, t), t)$ is uniformly asymptotically stable for $n = 1, 2, \ldots$.

Proof. To show the algorithm in [29] holds for infinite time t_1, we can first show that (35) holds for infinite t_1 by using the argument in [29]. Then by using similar argument in Lemma 2, Theorem 2, and Theorem 5 of [29], we can construct the sequence V^n satisfying (37). The convergence of V^n can be proved by using the argument in Lemma 3 and Theorem 7 in [29]. We now use an inductive argument to prove the uniform asymptotical stability of the closed-loop system $\dot{x} = f(x, c(x, V_x^n, t), t)$. It is clear that the closed-loop system $\dot{x} = f(x, c(x, V_x^1, t), t)$ is uniformly asymptotically stable. Now we assume that the closed-loop system $\dot{x} = f(x, c(x, V_x^n, t), t)$ is uniformly asymptotically stable for $n = k$, then we need to show that the closed-loop system $\dot{x} = f(x, c(x, V_x^n, t), t)$ is uniformly asymptotically stable for $n = k+1$. Note that $V^1(x,t)$ is finite, then by (37), $V^k(x,t)$ and $V^{k+1}(x,t)$ are both finite. Then by Assumption A1, we conclude that the closed-loop system $\dot{x} = f(x, c(x, V_x^{k+1}, t), t)$ is also uniformly asymptotically stable, which completes the proof. □

Based on the observations of the Kleinman algorithm and the algorithm in Theorems 3 and 4, we can find there are some similarities between them: both these two algorithms are used to solve equations arising in optimal control; in both these two algorithms, monotone non-increasing sequences are constructed to approximate the solutions of the desired equations (i.e., Riccati equations or HJB equations). In view of these similarities, we can reasonably suppose that the algorithm in [29] is a generalization of the Kleinman algorithm, and that it will have some similar features to those of the Kleinman algorithm, for example, a local quadratic rate of convergence. The second point is now under consideration and should be resolved in the near future.

5 Solving an HJBI Equation by a Sequence of HJB Equations

As we indicated in the introduction section, HJBI equations arise in nonlinear H_∞ control and they are first-order nonlinear partial differential equations and more difficult to solve than HJB equations. In this section, we present the recursive algorithm to solve HJBI equations. By using our algorithm, we replace the problem of solving an HJBI equation by the problem of solving a sequence of HJB equations; then by using the technique in Section 4, we can, if we wish, transfer each of these HJB equations to a sequence of linear PDEs. There are of course existing methods to solve HJBI equations, such as the method in [36] and the Galerkin approximation method in [7]. However, as we indicated in Section 1, there are some clear disadvantages for these algorithms. For example, when the methods in [36] and [7] are used to solve HJBI equations, it is difficult to choose an initial condition. Compared with the methods in [36] and [7], our algorithm has a very simple initial condition, viz, $V_0 = 0$. There are five subsections in this section: (1) preliminaries and definitions; (2) the summarizing theorem; (3) the algorithm; (4) rate of convergence and game theoretic interpretation of our algorithm; (5) a numerical example.

5.1 Preliminaries and Definitions

We work with the nonlinear control system $\Gamma : \mathcal{U}_0 \times \mathcal{W}_0 \to \mathcal{Y}_0$ given by the following equations:

$$x(0) = x_0, \tag{38}$$

$$\dot{x}(t) = f(x) + g_1(x)w(t) + g_2(x)u(t), \tag{39}$$

$$y(t) = h(x), \tag{40}$$

where $x \in \mathcal{X}_0$ is the state; $x_0 \in \mathbb{X}_0$ is the initial state; $u \in \mathcal{U}_0$ is the input; $w \in \mathcal{W}_0$ is the disturbance; $y \in \mathcal{Y}_0$ is the output. $f : \mathbb{X}_0 \to \mathbb{R}^n, g_1 : \mathbb{X}_0 \to \mathbb{R}^{n\times q}$,

$g_2 : \mathbb{X}_0 \to \mathbb{R}^{n\times m}$, and $h : \mathbb{X}_0 \to \mathbb{R}^p$ are smooth functions with $f(0) = 0$ and $h(0) = 0$. It is assumed further that f, g_1, g_2 are such that (39) has a unique solution for any $u \in \mathcal{U}_0$, $w \in \mathcal{W}_0$, and $x_0 \in \mathbb{X}_0$. Throughout Section 5, it is further assumed that the functions f, g_1, g_2, h defined in the system Γ can be represented in the following form:

$$f(x) = Ax + f_r(x), \tag{41}$$
$$g_1(x) = B_1 + g_{1r}(x), \tag{42}$$
$$g_2(x) = B_2 + g_{2r}(x), \tag{43}$$
$$h(x) = Cx + h_r(x), \tag{44}$$

where $x := x(t)$, A, B_1, B_2, C are real constant matrices with suitable dimensions and $f_r(x)$, $g_{1r}(x)$, $g_{2r}(x)$, $h_r(x)$ are higher order remainder terms in power expansions.

The steady-state HJBI equation associated with the system Γ we treat in Section 5 is

$$\begin{aligned} 0 &= 2\left(\frac{\partial \Pi(x)}{\partial x}\right)^T f(x) + \left(\frac{\partial \Pi(x)}{\partial x}\right)^T \big(g_1(x)g_1^T(x) \\ &\quad - g_2(x)g_2^T(x)\big)\left(\frac{\partial \Pi(x)}{\partial x}\right) + h^T(x)h(x), \qquad (45) \\ \Pi(0) &= 0 \end{aligned}$$

where f, g_1, g_2, h are real functions in the system Γ, $x \in \mathbb{X}_0$ is the state vector of the system Γ, and $\Pi : \mathbb{X}_0 \to \mathbb{R}^+$ is the unique local nonnegative stabilizing solution we seek. Here, a solution of (45) is called a local *stabilizing solution* if this solution is such that the closed loop of the system Γ is locally exponentially stable under the feedback inputs $u^* = -g_2^T(x)\frac{\partial \Pi(x)}{\partial x}$ and $w^* = g_1^T(x)\frac{\partial \Pi(x)}{\partial x}$. In such a situation, the vector field

$$\tilde{f}(x) = f(x) + g_1(x)g_1^T(x)\frac{\partial \Pi(x)}{\partial x} - g_2(x)g_2^T(x)\frac{\partial \Pi(x)}{\partial x} \tag{46}$$

is locally exponentially stable at the equilibrium point $x^* = 0$.

In the remainder of this subsection, we give some definitions which will be useful in our summarizing theorem.

We now define linear stabilizability of a matrix function pair by the stabilizability of the linear parts of this matrix function pair.

Definition 1. *[30] Let f, g_2 be the real functions defined in the system Γ and suppose* (41) *and* (43) *hold. The pair (f, g_2) is called linearly*[1] *stabilizable if there exists a matrix $\hat{D}$ such that $(A + B_2\hat{D})$ is a Hurwitz matrix.*

[1] There exists nonlinear systems which cannot be stabilized by linear controllers, for example, the nonlinear system $\dot{x} = \sqrt{x} + u$. This is the reason for the term linearly stabilizable, as opposite to "stabilizable."

Next, we define a function Θ, motivated by the right-hand side of the HJBI equation (45) that will be useful throughout Section 5.

Definition 2. *Let f, g_1, g_2, h be the real vector functions defined in the system Γ, and $x \in \mathbb{X}_0$ be the state value of Γ. Let $\mathbb{T}$ be the set which includes all smooth mappings from $\mathbb{X}_0$ to $\mathbb{R}$ and define $\Theta : \mathbb{T} \to \mathbb{T}$ as*

$$(\Theta(V))(x) = 2\left(\frac{\partial V(x)}{\partial x}\right)^T f(x) + \left(\frac{\partial V(x)}{\partial x}\right)^T \big(g_1(x)g_1^T(x) - g_2(x)g_2^T(x)\big)\left(\frac{\partial V(x)}{\partial x}\right) + h^T(x)h(x) \tag{47}$$

for all $V \in \mathbb{T}$, $x \in \mathbb{X}_0$.

We define two functions $\hat{f}_V$ and $\bar{f}_V$ which will be used to simplify the expressions in our main results and the HJB equations in our proposed algorithm (see (51) for example).

Definition 3. *Let f, g_1, g_2, h be the real vector functions defined in the system Γ. Let $\mathbb{T}, \Theta$ be defined as in Definition 2. Suppose there exists a local nonnegative stabilizing solution $\Pi \in \mathbb{T}$ to (45). Let $V \in \mathbb{T}$, let $\hat{f}_V : \mathbb{X}_0 \to \mathbb{R}$ be defined as*

$$\hat{f}_V(x) = f(x) + g_1(x)g_1^T(x)\frac{\partial V(x)}{\partial x} - g_2(x)g_2^T(x)\frac{\partial V(x)}{\partial x}$$

for all $x \in \mathbb{X}_0$, and let $\bar{f}_V : \mathbb{X}_0 \to \mathbb{R}$ be defined as

$$\bar{f}_V(x) = f(x) + g_1(x)g_1^T(x)\frac{\partial V(x)}{\partial x} - g_2(x)g_2^T(x)\frac{\partial \Pi(x)}{\partial x}$$

for all $x \in \mathbb{X}_0$.

5.2 The Summarizing Theorem

In this subsection, we set up the summarizing theorem by constructing two nonnegative function series $Z_k(x)$ and $V_k(x)$, and we also prove that $V_k(x)$ is monotonically increasing and converges to the unique local nonnegative stabilizing solution $\Pi(x)$ of (45) if such a solution exists.

Theorem 5. *Consider the system Γ, and let A, B_1, B_2, C be the real matrices appearing in (41), (42), (43), and (44), respectively. Let $x \in \mathbb{X}_0$ be the state of the system Γ. Define $\Theta : \mathbb{T} \to \mathbb{T}$ as in (47). Suppose (C, A) is detectable, (A, B_2) is stabilizable and there exists a stabilizing solution $\Pi \geq 0$ to (3). Let A_k, Z_k, and P_k be the matrix sequences appearing in Theorem 1. Then*

(I) *there exists a unique local nonnegative stabilizing solution $\Pi(x)$ to (45);*

(II) *two unique real function sequences* $Z_k(x)$ *and* $V_k(x)$ *for all* $k \in \mathbb{Z}_{\geq 0}$ *can be defined recursively as follows:*

$$V_0(x) \;=\; 0 \quad \forall x \in \mathbb{X}_0, \tag{48}$$

$Z_k(x)$ *is the unique local nonnegative stabilizing solution of*

$$\begin{aligned} 0 \;=\; & 2\hat{f}_{V_k}^T(x)\frac{\partial Z_k(x)}{\partial x} - \left(\frac{\partial Z_k(x)}{\partial x}\right)^T g_2(x)g_2^T(x)\frac{\partial Z_k(x)}{\partial x} + \\ & + \left(\Theta(V_k)\right)(x) \quad \forall x \in \mathbb{X}_0 \end{aligned} \tag{49}$$

with $0 = Z_k(0),\ 0 = \frac{\partial Z_k(x)}{\partial x}\Big|_{x=0}$, *and then*

$$V_{k+1} \;=\; V_k + Z_k; \tag{50}$$

(III) *the two series* $V_k(x)$ *and* $Z_k(x)$ *in part* (II) *have the following properties*:

(1) $(f(x)+g_1(x)g_1^T(x)\frac{\partial V_k(x)}{\partial x}, g_2(x))$ *is linearly stabilizable* $\forall k \in \mathbb{Z}_{\geq 0}\ \forall x \in \mathbb{X}_0$,

(2) $(\Theta(V_{k+1}))(x) = \left(\frac{\partial Z_k(x)}{\partial x}\right)^T g_1(x)g_1^T(x)\frac{\partial Z_k(x)}{\partial x}\ \forall k \in \mathbb{Z}_{\geq 0}\ \forall x \in \mathbb{X}_0$,

(3) $f(x) + g_1(x)g_1^T(x)\frac{\partial V_k(x)}{\partial x} - g_2(x)g_2^T(x)\frac{\partial V_{k+1}(x)}{\partial x}$ *is locally exponentially stable at the origin* $\forall k \in \mathbb{Z}_{\geq 0} \quad \forall x \in \mathbb{X}_0$,

(4) $\Pi(x) \geq V_{k+1}(x) \geq V_k(x) \geq 0\ \forall k \in \mathbb{Z}_{\geq 0}\ \forall x \in \mathbb{X}_0$,

(5) $Z_k(x) = \frac{1}{2}x^T Z_k x + O_0(x)\ \forall k \in \mathbb{Z}_{\geq 0}\ \forall x \in \mathbb{X}_0$,
$V_k(x) = \frac{1}{2}x^T P_k x + O_1(x)\ \forall k \in \mathbb{Z}_{\geq 0}\ \forall x \in \mathbb{X}_0$,

where Z_k *and* P_k *are the matrix sequences appearing in Theorem 1 and* $O_0(x)$ *and* $O_1(x)$ *are terms of higher order than quadratic.*

(IV). *For all* $x \in \mathbb{X}_0$, *the limit*

$$V_\infty(x) := \lim_{k\to\infty} V_k(x)$$

exists with $V_\infty(x) \geq 0$. *Furthermore,* $V_\infty = \Pi$ *is the unique local nonnegative stabilizing solution to* (45).

Proof. See [17]. □

From Theorem 5 (I) and (III1), we know that if there exists a local nonnegative stabilizing solution of (45), then (III1) holds. However, by Definition 1, we can check the linear stabilizability of a matrix function pair by checking its linear part. Hence we have the following corollary which gives a condition under which there does *not* exist a local stabilizing solution $\Pi(x) \geq 0$ to $\Theta(\Pi) = 0$. This is useful for terminating the recursion after a finite number of iterations.

Corollary 2. *Let* A, B_1, B_2, C *be the real matrices appearing in* (41), (42), (43), *and* (44). *Let* P_k *be the matrix sequence appearing in Theorem* 1. *Suppose*

that (C, A) is detectable and (A, B_2) is stabilizable. Let $x \in \mathbb{X}_0$ be the state of the system Γ. Define $\Theta : \mathbb{T} \to \mathbb{T}$ as in Definition 2. If $\exists k \in \mathbb{Z}_{\geq 0}$ such that $(A + B_1 B_1^T P_k, B_2)$ is not stabilizable, then there does not exist a local nonnegative stabilizing solution to $\Theta(\Pi) = 0$.

Proof. See [17]. □

Remark 3. It is worth pointing out that the sequence of HJB equations (49) is associated with a sequence of nonlinear optimal control problems; hence by using our algorithm, we have transferred a nonlinear H_∞ control problem into a sequence of nonlinear optimal control problems.

5.3 Algorithm

Let f, g_1, g_2, h be the real functions defined in the system Γ and suppose (41), (42), (43), and (44) hold. Let P_k be the matrix sequence appearing in Theorem 1. Let $\hat{f}_V$ be defined in Definition 3. Suppose (A, B_2) is stabilizable and (C, A) is detectable; an iterative algorithm for finding the local nonnegative stabilizing solution of (45) is given as follows:

1. Let $V_0 = 0$ and $k = 0$.
2. Construct (for example, using the algorithm in Theorem 4, though this is not necessary) the unique local nonnegative stabilizing solution $Z_k(x)$ which satisfies
$$0 = 2\hat{f}_{V_k}^T(x)\frac{\partial Z_k(x)}{\partial x} - \left(\frac{\partial Z_k(x)}{\partial x}\right)^T g_2(x) g_2^T(x)\frac{\partial Z_k(x)}{\partial x} + (\Theta(V_k))(x), \quad (51)$$
with $0 = Z_k(0)$, $0 = \frac{\partial Z_k(x)}{\partial x}\Big|_{x=0}$, where Θ is defined in Definition 2 and $\hat{f}$ is defined in Definition 3.
3. Set $V_{k+1}(x) = V_k(x) + Z_k(x)$.
4. Rewrite $Z_k(x) = \frac{1}{2}x^T Z_k x + O_0(x)$ (note that this is always possible from Theorem 5 if $Z_k(x)$ exists), where $O_0(x)$ are terms of higher order than quadratic and $Z_k \geq 0$ is the matrix sequence appearing in Theorem 1.
5. If $\overline{\sigma}(Z_k) < \mu$ where μ is a specified accuracy, then set $\Pi(x) = V_{k+1}(x)$ and exit. Otherwise, go to step 6.
6. If $(A + B_1 B_1^T P_k, B_2)$ is stabilizable, then increment k by 1 and go back to step 2. Otherwise, exit as there does not exist a local nonnegative stabilizing solution Π satisfying $\Theta(\Pi) = 0$.

From Corollary 2 we see that if the stabilizability condition in step 6 fails for some $k \in \mathbb{Z}_{\geq 0}$, then there does not exist a local nonnegative stabilizing solution Π to $\Theta(\Pi) = 0$ and the algorithm should terminate (as required by step 5). But when this stabilizability condition is satisfied $\forall\ k \in \mathbb{Z}_{\geq 0}$, construction of the series $V_k(x)$ and $Z_k(x)$ is always possible and either $V_k(x)$ converges to $\Pi(x)$ (which is captured by step 5) or $V_k(x)$ just diverges to infinity, which again means that there does not exist a stabilizing solution $\Pi(x) \geq 0$ to (45).

5.4 Rate of Convergence and Game Theoretic Interpretation

It can be shown that our algorithm to solve HJBI equations has a local quadratic rate of convergence and a game theoretic interpretation. Due to space restrictions, we omit the two parts here, please see [17] for more details.

5.5 A Numerical Example

In this subsection, a numerical example will be given. The example provides a numerical comparison between the method of characteristics [38] and our algorithm to solve an HJBI equation arising in nonlinear systems, and it shows that our proposed algorithm converges faster than the method of characteristics for this particular example.

Example 1. In [38], the method of characteristics is used to solve HJBI equations recursively. The following example comes from [36], and it illustrates the proposed algorithm outperforming the method of characteristics in [38] when solving HJBI equations. The comparison is possible because in this particular case, we are able to obtain the exact solution of the HJBI equation. The scalar system is given by

$$\dot{x}(t) = u(t) + xw(t) \tag{52}$$

with output $y(t) = x$. For this example, we have $f(x) = 0,\ g_1(x) = x,\ g_2(x) = 1,\ h(x) = 1,\ A = 0,\ B_1 = 0,\ B_2 = 1,\ C = 1$ and it is clear that (A, B_2) is stabilizable and (C, A) is detectable. Now the steady-state HJBI equation becomes

$$x^2 - \left(\frac{\partial \Pi(x)}{\partial x}\right)^2 (1 - x^2) = 0, \tag{53}$$

with $\Pi(0) = 0$. We have (without any approximation)

$$\frac{\partial \Pi(x)}{\partial x} = \pm \frac{x}{\sqrt{1 - x^2}}\ \forall x \in (-1, 1), \quad \Pi(0) = 0. \tag{54}$$

a. Exact solution

Since $\Pi(0) = 0$ and we seek the solution for which $\Pi(x) \geq 0$ in a neighborhood of the origin, we have

$$\frac{\partial \Pi(x)}{\partial x} = \frac{x}{\sqrt{1 - x^2}} \tag{55}$$

for $-1 < x < 1$. Now the closed-loop saddle point solution for the system (52) is $u^*(x) = -\frac{x}{\sqrt{1-x^2}}, w^*(x) = \frac{x^2}{\sqrt{1-x^2}}$ and the closed loop of the system (52) under the saddle point inputs u^* and w^* is

$$\dot{x} = -x\sqrt{1 - x^2} \tag{56}$$

for $-1 < x < 1$. Then it is clear that $x^* = 0$ is a local stable equilibrium point for the system (56). We approximate the value of $\Pi(x)$ by approximating the value of $\frac{\partial \Pi(x)}{\partial x}$. From (55), we know that the value of $\Pi(x)$ is symmetric about the origin. In view of this, we only approximate the value of $\frac{\partial \Pi(x)}{\partial x}$ for $0 \leq x < 1$ in the following. The exact solution of $\frac{\partial \Pi(x)}{\partial x}$ in (53) can be approximated by both our algorithm and the method of characteristics in [38].

b. Our algorithm

To approximate $\frac{\partial \Pi(x)}{\partial x}$ in (53), we carry out our proposed algorithm from Section 5.3. For convenience, we denote $(\cdot)_{k,x} = \frac{\partial (\cdot)_k}{\partial x}$ in the following for $k = 0, 1, 2, 3$. After a straightforward computation, we obtain the first three approximations $V_{1,x}, V_{2,x}, V_{3,x}$ of $\frac{\partial \Pi(x)}{\partial x}$ in (53) as follows:

$$\begin{aligned} V_{1,x} &= Z_{0,x} = x, \\ Z_{1,x} &= x^3 - x + x\sqrt{x^4 - x^2 + 1}, \\ V_{2,x} &= x^3 + x\sqrt{x^4 - x^2 + 1}, \\ Z_{2,x} &= f_2 + \sqrt{f_2^2 + x^2 Z_{1,x}^2}, \\ V_{3,x} &= x^5 + x^3\sqrt{x^4 - x^2 + 1} + \sqrt{f_2^2 + x^2 Z_{1,x}^2}, \end{aligned}$$

where $f_2 = x^5 - x^3 + (x^3 - x)\sqrt{x^4 - x^2 + 1}$.

c. Algorithm of characteristics

If we use the method in [38] to approximate the local nonnegative stabilizing solution $\Pi(x)$ to the HJBI equation (5), the first three approximations $\overline{V}_{1,x}, \overline{V}_{2,x}, \overline{V}_{3,x}$ of $\frac{\partial \Pi(x)}{\partial x}$ in (53) are

$$\begin{aligned} \overline{V}_{1,x} &= x, \\ \overline{V}_{2,x} &= x + \frac{1}{2}x^3, \\ \overline{V}_{3,x} &= x + \frac{1}{2}x^3 + \frac{7}{16}x^5 + \frac{9}{80}x^7 + \frac{437}{53760}x^9. \end{aligned}$$

We plot these approximations together in Fig. 1 (we ignore the first approximations for both algorithms since they are identical) to compare their convergence to the "exact solution," which is given by (55).

From Fig. 1, we can see that our algorithm has better accuracy than the method of characteristics in [38], noting in particular the following points:

1. For both the second approximation and the third approximation, our algorithm is more accurate than the method in [38].
2. The second approximation (dotted line) of our algorithm is very close to the third approximation (dashed line) of the method in [38].
3. The third approximation of our algorithm (thin solid line) is very close to the exact solution (thick solid line).

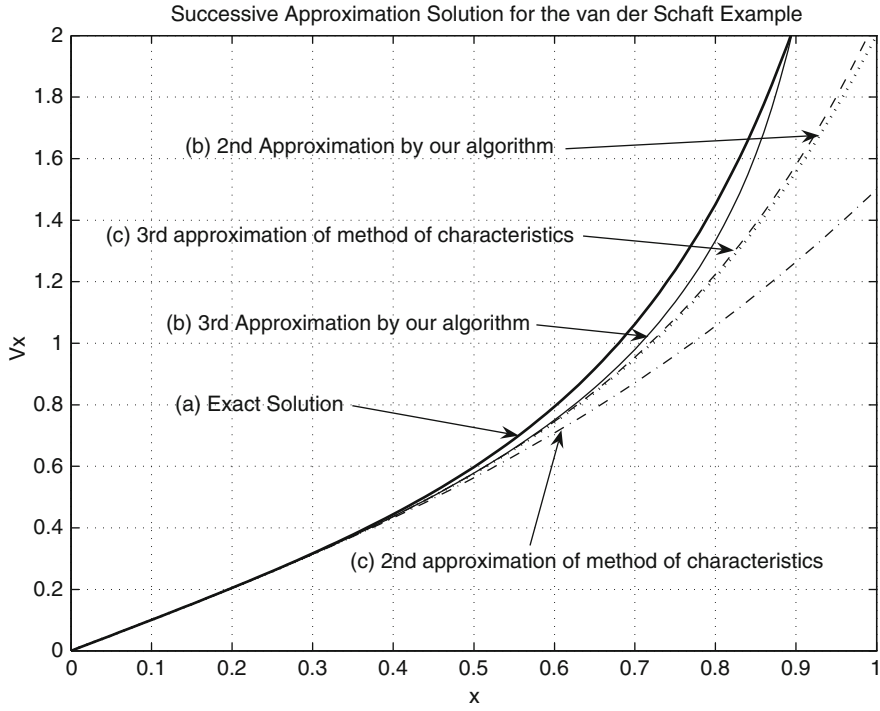


Fig. 1. Demonstration and comparison of algorithm

6 Conclusions

There have been two main thrusts in this chapter. First, motivated by the knowledge that standard Riccati equation solvers can encounter numerical problems on occasions, which can often be fixed for H_2 Riccati equations using a Kleinman algorithm, we developed a new algorithm to solve H_∞ Riccati equations. The algorithm is recursive, with each recursive step requiring solution of an H_2 Riccati equation (itself amenable to solution via the Kleinman algorithm). The algorithm has the following advantages: (1) a simple initialization; (2) local quadratic rate of convergence; (3) a natural game theoretic interpretation; (4) high numerical stability and reliability. Under some suitable assumptions, we can compute nonnegative stabilizing solutions of Riccati equations and HJBI equations recursively.

Second, motivated as much by the sparsity of solution methods for HJBI equations as by numerical problems arising from time to time in known solution procedures, we have illustrated how, for a class of HJBI equations, one can replace the problem of solving a single such equation by the problem of solving a sequence of HJB equations, each of which can be tackled using a sequence of linear partial differential equations.

The ideas presented may be valid in much broader game theoretic contexts than those considered here. One recent extension we have achieved is to the solution of H_∞ periodic Riccati differential equations (see [4]), and it is possible that our algorithm can be extended to solve H_∞ periodic HJBI equations.

Acknowledgment

This work has been supported in part by Australian Research Council Discovery Project Grant DP0664427.

References

1. Abou-Kandil, H., Freiling, G., Ionescu, V., Jank, G.: Matrix Riccati Equations in Control and Systems Theory, Birkhäuser, Basel, Switzerland (2003).
2. Anderson, B.D.O., Hitz, K.L., Diem, N.D.: Recursive algorithms for spectral factorization. IEEE Trans. Circuits Syst. CAS-21 (6), 742–750, November (1974).
3. Anderson, B.D.O.: Second order convergence algorithms for the steady-state Riccati equation. Int. J. Contr. 28(2), 295–306 (1978).
4. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users Guide: Third Edition, ser. Software Environments Tools. Philadelphia: SIAM, 1999.
5. Arnold, III W.F., Laub, A.J.: Generalized eigenproblem algorithms and software for algebraic Riccati equations, *Proceedings of the IEEE*, 72(12), 1746–1754 (1984).
6. Bartels, R.H., Stewart, W., Solution of the matrix $AX + XB = C$. Commun. ACM, 15, 820–826 (1972).
7. Beard, R.W., McLain, T.W.: Successive Galerkin approximation algorithm for nonlinear optimal and robust control. Int. J. Control, 71(5), 717–743 (1998).
8. Bellman, R.E.: Dynamic programming, Princeton, Princeton University Press, (1957).
9. Bellman, R.E., Kalaba, R.E.: Quasilinearization and Nonlinear Boundary-Value Problems, American Elsevier, New York (1965).
10. Bellman, R.E.: Introduction to the Mathematical Theory of Control Process, Academic, New York (1971).
11. Benner, P.: Computational methods for linear-quadratic optimization. Suppl. Rend. Circ. Mat. Palermo, Ser. II, 58, 21–56 (1999).
12. Benner, P., Hernández, V., Pastor, A.: On the Kleinman iteration for nonstabilizable systems. J. Math. Control Signals Syst. 16(1) 76–93 (2003).
13. Cherfi, L., Abou-Kandil, H., Bourles, H.: Iterative Method for General Algebraic Riccati Equation, *ACSE 05 conference*, 19–21 Dec. 2005, CICC, Cairo, Egypt.
14. Damm, T.: Rational Matrix Equations in Stochastic Control, volume 297 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, (2004).

15. Datta, B.N., Numerical Methods for Linear Control Systems, Elsevier New York, (2004).
16. Dieci, L.: Some numerical considerations and Newton's method revisited for solving algebraic Riccati equations. IEEE Trans. Automa. Control, 36, 608–616 (1991).
17. Feng, Y., Anderson, B.D.O., Rotkowitz, M.: A game theoretic algorithm to compute local stabilizing solutions to HJBI equations in nonlinear H_∞ control. Automatica, 45(4), 881–888 April (2009).
18. Golub, G.H., Van Loan, C.F.: Matrix Computations, The John Hopkins University Press, Baltimore and London (1996).
19. Green, M., Limebeer, D.J.N.: Linear Robust Control, Prentice Hall, Englewood Cliffs, NJ (1995).
20. Guo, C.H.: A note on the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation. Linear Algebra Appl. 357, 299–302 (2002).
21. Guo, C.H.: Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models. J. Comput Appl Math, 192, 353–373 (2006).
22. Hitz, K.L., Anderson, B.D.O.: An iterative method of computing the limiting solution of the matrix Riccati differential equation. Proc. IEE, 119(9), 1402–1406, September (1972).
23. Kalman, R.E.: The Theory of Optimal Control and the Calculus of Variations. R. Bellman (ed.), Mathematical Optimization Techniques, University of California Press, Berkeley, 309–330 (1963), Chap. 16.
24. Kleinman, D.L.: On an iterative technique for Riccati equation computation. IEEE Trans. Automat. Control 13, 114–115 (1968).
25. Lancaster, P., Rodman, L.: The Algebraic Riccati Equation, Oxford University Press, Oxford, (1995).
26. Lanzon, A., Feng, Y., Anderson, B.D.O.: An iterative algorithm to solve Algebraic Riccati equations with an indefinite quadratic term. *Proceedings of European Control Conference 2007*, 3033–3039, Greece.
27. Lanzon, A., Feng, Y., Anderson, B.D.O., Rotkowitz, M.: Computing the positive stabilizing solution to algebraic riccati equations with an indefinite quadratic term via a recursive method. IEEE Trans. Automat. Contr. 53(10), 2280–2291, November (2008).
28. Laub, A.J.: A Schur method for solving algebraic Riccati equation. IEEE Trans. Automat. Control 24, 913–921 (1979).
29. Leake, R.J., Liu, R.: Construction of suboptimal control sequences. SIAM J. Control, 5(1) 54–63 (1967).
30. Lukes, D.L.: Optimal regulation of nonlinear dynamical systems. SIAM J. Control 7(1), 75–100 (1969).
31. Martensson, K.: On the matrix Riccati equation: Inf. Sci. 3, 17–49 (1971).
32. Mehrmann, V., Tan, E.: Defect correction methods for the solution of algebraic Riccati equations. IEEE Trans. Automat. Control 33, 695–698 (1988).
33. Mehrmann, V.: The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution, number 163 in Lecture Notes in Control and Information Sciences. Springer-Verlag, Heidelberg (July 1991).
34. Mehrmann, V.: Numerical Methods for Eigenvalue and Control Problems. Frontiers in Numerical Analysis. Springer New York, NY (2002).

35. Sima, V.: Algorithms for Linear-Quadratic Optimization, volume 200 in Pure and Applied Mathematics, Marcel Dekker, New York, (1996).
36. van der Schaft, A.J.: L_2-gain analysis of nonlinear systems and nonlinear systems and nonlinear state feedback H_∞ control. IEEE Trans. Automat. Control 37, 770–784 (1992).
37. Willems, J.C.: Least Squares Stationary Optimal Control and the Algebraic Riccati Equation. IEEE Trans. Automat. Control, AC-16(6), 621–634 (1971).
38. Wise, K.A., Sedwick, J.L.: Successive approximation solution of the HJI equation. *Proceeding IEEE Conference on Decision and Control*, 1387–1991 (1994).
39. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control, Prentice Hall, Upper Saddle River, NJ (1996).

Online Adaptive Optimal Control Based on Reinforcement Learning

Draguna Vrabie[1] and Frank Lewis[2]

[1]Automation and Robotics Research Institute, University of Texas at Arlington, 7300 Jack Newell Blvd. S., Fort Worth, TX 76118, USA
dvrabie@uta.edu
[2]Automation and Robotics Research Institute, University of Texas at Arlington, 7300 Jack Newell Blvd. S., Fort Worth, TX 76118, USA
lewis@uta.edu

Summary. In this chapter a new online direct adaptive scheme is presented which converges to the optimal state feedback control solution for affine in the inputs nonlinear systems. The optimal control solution is obtained in a direct fashion, without system identification. The optimal adaptive control algorithm is derived in a continuous-time framework. The algorithm is an online approach to policy iterations based on an adaptive critic structure to find an approximate solution to the state feedback, infinite-horizon, optimal control problem.

Key words: adaptive control, optimal control, dual control, dynamic programming

1 Introduction

Adaptive control names the class of techniques developed with the purpose of maintaining prescribed control performances for systems with slowly time-varying or uncertain parameters [21]. The controller is defined as a parameterized mapping, also addressed as control policy, between the system states (i.e., measured information from the system) and the control output signal. The adaptive mechanism is concerned with changing the parameters and/or parametric structure of the controller such that the closed-loop system maintains certain prescribed performances. From this perspective an adaptive control system is constructed based on a learning mechanism which takes place at the level of the controller. In the following we shall restrict the discussion to the case in which the structure of the controller is fixed and adaptation is obtained based on the variation of the controller parameters only.

Depending on whether information on the system dynamics is available and used in the learning process, adaptive controllers can be classified as

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_16,

1. *indirect*, the adaptive mechanism makes use of a model of the system to be controlled, and
2. *direct*, the adaptation process does not require a model of the system.

In the case of indirect adaptive controllers the parameter adaptation is based on the information which encodes a model of the system's behavior. In this case, a secondary learning process takes place having as result a parametric description of the system to be controlled, i.e., a mapping between the control action and the system states. Following the system model identification step, the indirect adaptive controllers make use of the model, under the certainty equivalence assumption, to determine the controller parameters that will satisfy the specified performances. An important class of indirect adaptive controllers emerges from the dual control theory [7]. Dual controllers have a twofold objective when computing the control signal: it must satisfy the control performance goal while at the same time must sufficiently excite the plant to improve system parameters estimation.

Direct adaptive controllers learn the mapping between the system states and the action signal, i.e., the control policy parameters, only based on an error signal which encodes the difference between the system's performance with the present control policy and the specified desired performances. In this case identification of a model of the system dynamics is not required, the learning taking place only at the controller level.

At the same time, one can differentiate between adaptive controllers based on the way in which the desired performances are specified. From this perspective the measure of performance can be given by a formal cost function of the sort encountered in the optimal control problem specification or as a tracking error-based cost in which case the performance is prescribed in terms of a desired closed-loop trajectory. In the first case the control problem can be referred to as an optimal adaptive control problem. Stabilizing adaptive controllers that are inverse optimal, with respect to some relevant cost not specified by the designer, have also been derived [16].

It is well known that solving the optimal control problem is generally difficult even in the presence of complete and correct knowledge of the system dynamics, as Bellman's dynamic programming approach suffers from the so-called curse of dimensionality [15]. Also, developing controllers which would satisfy optimality performances, while making use of an approximate model of the system, will have as result suboptimal solutions; and this is an even more important issue when dealing with plants that have slowly time-varying or uncertain dynamics. This problem motivated several advances in solving the optimal control problem using dual adaptive control techniques, surveyed in [8, 28] which would simultaneously improve the estimated system model parameters and improve on the suboptimal controller. Nonetheless the curse of dimensionality appeared also in this case together with another difficulty, posed by dual control theory, known as the exploration–exploitation dilemma. The dilemma consists in the conflict which appears at the level of the controller when action signals must be computed such that to fa-

vor the system identification process while at the same time to ensure the prescribed performances.

In order to adaptively solve optimal control problems a new methodology, namely reinforcement learning (RL), was developed in the computational intelligence community and then gradually adapted to fit the control engineering requirements. Reinforcement learning means finding a control policy, i.e., learning the parameters of a controller mapping between the system states and the control signal, such that to maximize a numerical reward signal [22]. It is important to note that the integral over time of the reward signal can be viewed as the value/cost function to be maximized/minimized in an optimal control framework. Reinforcement learning is defined by characterizing a learning problem which is in fact the adaptive optimal control problem. Thus, from a control engineering perspective, RL algorithms can be viewed as a class of adaptive controllers which solve the optimal control problem based on reward information which gives information on the performance of a given controller. Considering whether the system model is or is not required by a certain reinforcement learning algorithm one can classify the RL algorithms as direct or indirect.

In this chapter we will focus our attention on a class of reinforcement learning algorithms, namely policy iteration. The goal of the chapter is to present a new policy iteration algorithm which, without making use of complete knowledge of a system's dynamics, will learn to approximate, in an online fashion and with arbitrary small accuracy, the optimal control solution for a general nonlinear affine in the input continuous-time system.

Generally the solution of the optimal control problem can be obtained by directly solving the Hamilton–Jacobi–Bellman (HJB) equation [15] for the optimal cost and then using the result to calculate the optimal control policy (i.e., the feedback gain for linear systems). This approach has a great disadvantage given by the difficulty of solving the HJB equation. In order to solve the optimal control problem, the policy iteration method starts with the evaluation of the cost associated with an initial stabilizing control policy and then uses this information to obtain a new policy which will result in improved control performances.

Policy iteration algorithms are built on a two-step iteration: policy evaluation and policy improvement. The two steps are repeated until the policy improvement step no longer changes the actual policy, thus the optimal control policy is obtained. The algorithm can be viewed as a directed search for the optimal controller in the space of admissible control policies. Policy iteration algorithm was first formulated in [12]. For continuous state linear systems, policy iteration algorithms were developed in [5, 18, 24] to find the optimal linear quadratic regulator (LQR) [15]. Convergence guarantees were given in [10, 13, 14]. Even more, in [5] the policy iteration algorithm, formulated to solve the discrete-time LQR problem, used the so-called Q-functions [25, 26], and this resulted in the model-free feature of the algorithm. As the Q-function-based formulation has not been yet considered in a continuous-time framework, in [18] the model-free quality of the approach

was achieved either by evaluating online the infinite horizon cost associated with an admissible control policy or by using measurements of the state derivatives. The policy iteration algorithm developed in [24] is an online technique which solves the LQR problem along a single-state trajectory, using only partial knowledge about the system dynamics and without requiring measurements of the state derivative. The mathematical formulation of the policy iteration in [13] shows that the algorithm is in fact a two-step Newton method for which the policy evaluation step is equivalent with solving a Lyapunov equation. In the case of linear systems the Lyapunov equation can be solved exactly based on data collected online along a single, sufficiently exciting, state trajectory [24].

When the optimal control problem is formulated for continuous-time nonlinear systems the policy iteration approach to solve this problem is in fact the method of successive approximations developed in [14, 20]. This method iterates on a sequence of Lyapunov equations, also addressed as generalized HJB equations, which are somewhat easier to solve than the HJB equation. In [2, 3] the solution for these Lyapunov equations was obtained using the Galerkin spectral approximation method and in [1] they were solved, in the presence of saturation restrictions on the control input, using neural network approximator structures. Neural network-based structures for learning the optimal control solution via the HJB equation, namely adaptive critics, were first proposed in [17]. The use of neural network-type learning elements in control structures was motivated by the capability of such elements to approximate nonlinear maps and by the fact that in the past decades there have been developed a variety of algorithms which allow the online adaptation of such structures based on data acquired from a not completely known environment. Neural network-based adaptive critics and training algorithms were presented both in discrete-time [19] and continuous-time [9] frameworks.

It is now important to mention that the policy iteration methods developed in [1–3] as well as the inverse optimal controller in [16] are generally applied offline as they require complete knowledge on the dynamics of the system to be controlled. Due to their offline character imposed by the system model requirement these methods are not sensitive to changes in the system dynamics. The algorithm that we present in this chapter is a policy iteration algorithm which uses the Bellman optimality equation as a consistence relation when solving for the value associated with a given policy and not the regular, Hamiltonian-based, Lyapunov equation. The fact that in Bellman's equation the system dynamics does not explicitly appear is the major advantage which results in the model-free property of the proposed algorithm and grants its online implementation capability.

In the next section we briefly review the formulation of the continuous-time optimal control problem for nonlinear systems. The new online policy iteration algorithm is then presented followed by its neural network-based online implementation on an actor–critic structure. A numerical example is then given, followed by concluding remarks.

2 The Continuous-Time Optimal Control Problem

Consider the time-invariant affine in the input dynamical system given by

$$\dot{x}(t) = f(x(t)) + g(x(t))u(x(t)); \quad x(0) = x_0, \tag{1}$$

with $x(t) \in \mathrm{R}^n$, $f(x(t)) \in \mathrm{R}^n$, $g(x(t)) \in \mathrm{R}^{n\times m}$, and the input $u(t) \in U \subset \mathrm{R}^m$. We assume that the system is such that $f(0) = 0$, $f(x) + g(x)u$ is Lipschitz continuous on a set $\Omega \subseteq \mathrm{R}^n$ that contains the origin, and that the dynamical system is stabilizable on Ω, i.e., there exists a continuous control function $u(t) \in U$ such that the system is asymptotically stable on Ω.

Define the infinite horizon integral cost

$$V(x_0) = \int_0^\infty r(x(\tau), u(\tau))d\tau, \tag{2}$$

where $r(x, u) = Q(x) + u^T Ru$ with $Q(x)$ positive definite, i.e., $\forall x \neq 0, Q(x) > 0$ and $x = 0 \Rightarrow Q(x) = 0$, and $R \in \mathrm{R}^{m\times m}$ is a positive definite matrix.

Definition 1 (Admissible policy). *A control policy $\mu(x)$ is defined as admissible with respect to (2) on Ω, denoted by $\mu \in \Psi(\Omega)$, if $\mu(x)$ is continuous on Ω, $\mu(0) = 0$, $\mu(x)$ stabilizes (1) on Ω and $V(x_0)$ is finite $\forall x_0 \in \Omega$.*

For any admissible control policy $\mu \in \Psi(\Omega)$ if the associated cost function

$$V^\mu(x_0) = \int_0^\infty r(x(\tau), \mu(x(\tau)))d\tau \tag{3}$$

is C^1, then a infinitesimal version of (3) is

$$0 = r(x, \mu(x)) + V_x^{\mu T}(f(x) + g(x)\mu(x)),\ V^\mu(0) = 0, \tag{4}$$

where V_x^μ denotes the partial derivative of the value function V^μ with respect to x, as the value function does not depend explicitly on time. Equation (4) is a Lyapunov equation for nonlinear systems which, given the controller $\mu(x) \in \Psi(\Omega)$, can be solved for the value function $V^\mu(x)$ associated with it. Given that $\mu(x)$ is an admissible control policy, if $V^\mu(x)$ satisfies (4), with $r(x, \mu(x)) \geq 0$, then $V^\mu(x)$ is a Lyapunov function for the system (1) with control policy $\mu(x)$.

The optimal control problem can now be formulated: Given the continuous-time system (1), the set $u \in \Psi(\Omega)$ of admissible control policies, and the infinite horizon cost functional (2), find an admissible control policy such that the cost index (2) associated with the system (1) is minimized.

Defining the Hamiltonian of the problem

$$H(x, u, V_x^*) = r(x(t), u(t)) + V_x^{*T}(f(x(t)) + g(x(t))u(t)), \tag{5}$$

the optimal cost function $V^*(x)$ satisfies the HJB equation

$$0 = \min_{u \in \Psi(\Omega)} [H(x, u, V_x^*)]. \tag{6}$$

Assuming that the minimum on the right-hand side of equation (6) exists and is unique then the optimal control function for the given problem is

$$u^*(x) = -R^{-1} g^T(x) V_x^*(x). \tag{7}$$

Inserting this optimal control in the Hamiltonian we obtain the HJB equation in terms of V_x^*

$$0 = Q(x) + V_x^{*T}(x) f(x) - \frac{1}{4} V_{x*}^T(x) g(x) R^{-1} g^T(x) V_x^*(x); \quad V^*(0) = 0. \tag{8}$$

This is a necessary and sufficient condition for the optimal value function [15]. For the linear system case, considering a quadratic cost functional, the equivalent of this HJB equation is the well-known Riccati equation.

In order to find the optimal control solution for the problem, one only needs to solve the HJB equation (8) for the value function and then substitute the solution in (7) to obtain the optimal control. However, solving the HJB equation is generally difficult as it is a nonlinear differential equation, quadratic in the cost function. Even if a solution of this equation would be readily available, in order to obtain it one needs to have complete knowledge of the system dynamics, i.e., the system dynamics described by the functions $f(x), g(x)$ need to be known.

3 The Policy Iteration Algorithm

In order to solve the optimal control problem, instead of directly solving the HJB equation (8) for the optimal cost and then finding the optimal control policy given by (7), the policy iteration method starts by evaluating the cost of a given initial admissible policy and then makes use of this information to improve the control policy. The two steps are repeated until the policy improvement step no longer changes the actual policy. The following online reinforcement learning algorithm will solve the infinite horizon optimal control problem without using knowledge regarding the system internal dynamics (i.e., the system function $f(x)$).

First note that given an admissible policy for (1), $\mu(x)$, such that the closed-loop system is asymptotically stable on Ω, then the infinite horizon cost for any $x(t) \in \Omega$ is given by (3) and $V^\mu(x(t))$ serves as a Lyapunov function for (1). The cost function (3) can thus be written as

$$V^\mu(x(t)) = \int_t^{t+T} r(x(\tau), \mu(x(\tau)))d\tau + V^\mu(x(t+T)). \tag{9}$$

Based on (9) and (6), considering an initial admissible control policy $\mu^{(0)}(x)$, the following policy iteration scheme can be derived:

1. Solve for $V^{\mu^{(i)}}(x)$ using

$$V^{\mu^{(i)}}(x(t)) = \int_t^{t+T} r(x(\tau), \mu^{(i)}(x(\tau)))d\tau + V^{\mu^{(i)}}(x(t+T)), \quad V^{\mu^{(i)}}(0) = 0. \tag{10}$$

2. Update the control policy using

$$\mu^{(i+1)}(x) = \arg\min_{\mu}\{H(x, \mu, V_x^{\mu^{(i)}})\}, \tag{11}$$

which in this case is

$$\mu^{(i+1)}(x) = -R^{-1}g^T(x)V_x^{\mu^{(i)}}(x). \tag{12}$$

Equations (10) and (12) formulate a new policy iteration algorithm to solve for the optimal control without making use of any knowledge of the system internal dynamics $f(x)$. The online implementation of the algorithm will be discussed in Section 4. This algorithm is an online version of the offline algorithms proposed in [1–3] inspired by the online adaptive critic techniques proposed by computational intelligence researchers in [4, 19, 27]. The convergence of the algorithm is now discussed.

Lemma 1. *Solving for $V^{\mu^{(i)}}$ in (10) is equivalent to finding the solution of the Lyapunov equation*

$$0 = r(x, \mu^{(i)}(x)) + V_x^{\mu^{(i)}T}(f(x) + g(x)\mu^{(i)}(x)), \; V^{\mu^{(i)}}(0) = 0. \tag{13}$$

Proof. See [23]. Note that although the same solution is obtained whether solving (10) or (13), solving (10) does not require any knowledge on the system dynamics $f(x)$. It thus follows that the algorithm (10) and (12) is equivalent to iterating between (13) and (12), without using knowledge of the system internal dynamics.

Theorem 1 (convergence). *The policy iteration algorithm (10) and (12) converges to the optimal control solution on the trajectories having initial state $x_0 \in \Omega$.*

Proof. In [1–3] it was shown that using policy iteration conditioned by an initial admissible policy $\mu^{(0)}(x)$, all the subsequent control policies will be admissible and the iteration between (13) and (12) will converge to the solution of the HJB equation. Based on the proven equivalence between (10) and (13) we can conclude that the proposed adaptive optimal control algorithm will converge to the solution of the optimal control problem with infinite horizon cost (2) without using knowledge on the internal dynamics of the controlled system (1).

4 Online Adaptive Optimal Control Solution Using Neural Network Elements in an Actor–Critic Structure

In order to solve for the cost function $V^{\mu^{(i)}}(x)$ in (10) we will use a neural network, which has the universal approximation property [11], to obtain an expression for the value function for any initial state $x \in \Omega$. The cost function $V^{\mu^{(i)}}(x(t))$ will be approximated by

$$V^{\mu^{(i)}}(x) = \sum_{j=1}^{L} w_j^{\mu^{(i)}} \phi_j(x) = \left(w_L^{\mu^{(i)}}\right)^T \phi_L(x), \tag{14}$$

a neural network with L neurons on the hidden layer, where $w_j^{\mu^{(i)}}$ denote the weights of the neural network and $\phi_j(x) \in C^1(\Omega),\ \phi_j(0) = 0$ denote the activation functions. In a compact representation, ϕ_L is the vector of activation functions and $w_L^{\mu^{(i)}}$ is the weight vector. In the following we assume that the neural network structure can result in an exact description of the cost function. Using the neural network description for the value function, (14), (10) can be written as

$$w_L^{\mu^{(i)}T} \phi_L(x(t)) = \int_t^{t+T} r(x, \mu^{(i)}(x)) d\tau + w_L^{\mu^{(i)}T} \phi_L(x(t+T)). \tag{15}$$

As the cost function was replaced with the neural network approximation, (15) will have the residual error

$$\delta_L^i(x(t)) = \int_t^{t+T} r(x, \mu^{(i)}(x)) d\tau + w_L^{\mu^{(i)}T} [\phi_L(x(t+T)) - \phi_L(x(t))]. \tag{16}$$

From the perspective of temporal difference learning methods, e.g., [6], this error can be viewed as temporal difference residual error. To determine the parameters of the neural network approximating the cost function, in the least squares sense, we use the method of weighted residuals. Thus we seek to minimize the objective

$$S = \int_{\Omega_{\{x_0\}_n}^{\mu^{(i)}}} \delta_L^i(x,T) \delta_L^i(x,T) dx \tag{17}$$

where $\Omega_{\{x_0\}_n}^{\mu^{(i)}}$ denotes a set of trajectories generated by the policy $\mu^{(i)}$ starting from the initial conditions $\{x_0\}_n \subset \Omega$. Using the inner product notation for the Lebesgue integral the minimization of the objective function (17) amounts to

$$\left\langle \frac{d\delta_L^i(x,T)}{dw_L^{\mu^{(i)}}}, \delta_L^i(x,T) \right\rangle_{\Omega_{\{x_0\}_n}^{\mu^{(i)}}} = 0. \tag{18}$$

Conditioned by $\Phi = \left\langle [\phi_L(x(t+T)) - \phi_L(x(t))], [\phi_L(x(t+T)) - \phi_L(x(t))]^T \right\rangle$ being invertible, then we obtain the solution

$$w_L^{\mu^{(i)}} = -\Phi^{-1} \left\langle [\phi_L(x(t+T)) - \phi_L(x(t))], \int_t^{t+T} r(x(s), \mu^{(i)}(x(s)))ds \right\rangle. \tag{19}$$

Results showing that matrix Φ is invertible, conditioned by an excitation requirement related to the selection of the sample time T, are available in [23]. The parameters $w_L^{\mu^{(i)}}$ of the cost function can be calculated using only online measurements of the state vector and the integrated reward over a finite time interval $[t, t+T]$. The solution given by (19) can be obtained in real time after a sufficient number of data points are collected along a finite number of state trajectories in Ω. In practice, the inversion of matrix Φ is not performed; the solution given by (19) is obtained using algorithms that involve techniques such as Gaussian elimination, back-substitution, and Householder reflections.

The flowchart of the online algorithm is presented in Fig. 1. The iterations will be stopped (i.e., the critic will stop updating the control policy) when the error between the system performance evaluated at two consecutive steps will cross below a designer-specified threshold ε. Also, when this error becomes bigger than the above-mentioned threshold the critic will take again the decision to start tuning the actor parameters.

It has to be emphasized that, in order to successfully apply the algorithm, enough excitation must be present in the system. Thus, if the system state reached the equilibrium point (this is often the case since the algorithm iterates only on stabilizing controllers) the data measured from the system can no longer be used in the adaptive algorithm; in this case the system must

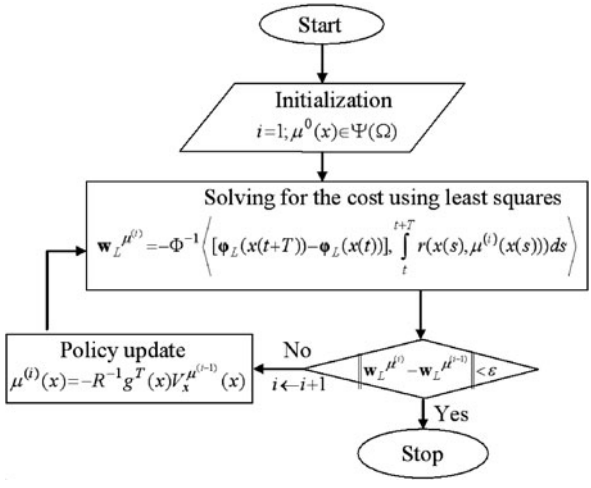


Fig. 1. Flowchart of the online algorithm

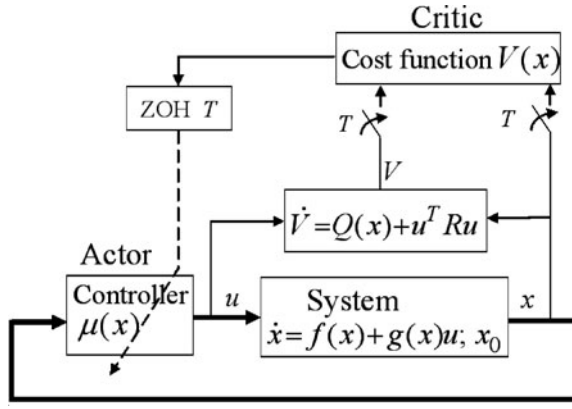


Fig. 2. Structure of the system with adaptive controller

be again excited to a nonzero initial state and a new experiment needs to be conducted having as starting point the last policy obtained in the previous experiment. Figure 2 presents the structure of the system with optimal adaptive controller.

The proposed optimal adaptive procedure requires only measurements of the states at discrete moments in time (measured using a sample time T), as well as knowledge of the observed cost over several time intervals of size T. The control policy remains unchanged until a sufficient number of measurements (say N) have been taken, such that the solution given by (19) becomes feasible. After the parameters of a new policy have been determined, the control policy is updated and it will be used for controlling the system during several time intervals starting with the time $t + NT$; thus the algorithm is suitable for online implementation from the control theory point of view. One also observes that the dynamics of the system has been augmented with an additional state $V(t)$, with dynamics given by $\dot{V} = Q(x) + u^T$ Ru. Measuring this state at discrete moments in time is equivalent with extracting the reward information regarding the cost associated with the given policy. Thus, having little information about the system states, x, and the augmented system state, V, measured from the system only at discrete moments in time, the critic is able to evaluate the performance of the system associated with a given control policy; this is then followed by policy improvement.

It is observed that the update of both the actor and the critic is performed at discrete moments in time. However, the control action is a full-fledged continuous-time control, with its constant gain updated at discrete moments in time. Moreover, the critic update is based on the observations of the continuous-time cost over a finite sample interval. As a result, the algorithm converges to the solution of the continuous-time optimal control problem.

It is important to note that no knowledge of the system internal dynamics is required in either of the two steps of the policy iteration algorithm. The

information regarding the system $f(x)$ matrix is in fact embedded in the states $x(t)$ and $x(t+T)$ which are sampled online; nevertheless no identification procedure needs to be performed. The $g(x)$ matrix is though required in the second step of the policy iteration algorithm as it appears explicitly in (13) for the update of the control policy, and this makes the online tuning algorithm partially model free. However, in practice the knowledge requirement on the input to state dynamics is not a real issue as the function $g(x)$ is related to the, regularly known, actuator dynamics.

Looking at the actor–critic structure of the adaptive system one notes that even if no system model is identified, nor required, in fact the algorithm implies an identification procedure at the level of the critic, i.e., the cost function associated with a given control policy must be identified. It would thus seem that this direct adaptive optimal controller would not have any advantage compared to the indirect adaptive optimal control methods. For this reason we will now bring a couple of reasons as to why this identification procedure is relatively affordable in comparison with a system model identification.

In order to identify a system model, data must be collected such that it contains information related to all the system natural modes and in the case in which the plant to be controlled has unstable dynamics such a procedure could lead to destabilizing the system. When the identification is performed in closed loop, excitation signals must be either injected through the reference of the closed-loop system or added over the controller output signal. In the first case, the presence of a supervisor external to the adaptive control system is required to prescribe the excitatory reference signal. In the second situation, it appears again the exploitation–exploration dilemma characterizing the dual optimal controllers. When a model of the performance of a closed-loop system is identified, one only needs to use data sampled during the normal, stable operation of the control system when that specific policy is used. The exploration requirement is thus removed from the controller level while sufficient excitation required for cost function learning can be obtained based on measurements taken on different state trajectories in Ω.

5 Simulation Results

We now illustrate the results of the adaptive optimal control algorithm considering the nonlinear system given by the equations

$$\begin{aligned} \dot{x}_1 &= -x_2^3 - x_2 \\ \dot{x}_2 &= x_1 + x_2 + u. \end{aligned} \tag{20}$$

We first consider a linear version of the system (20), not including the cubic term in the dynamics of the first state. The simulation was conducted using data measured from the system at a sample rate of $T = 0.09\,\mathrm{s}$. The required initial stabilizing controller was taken as

$$\mu^{(0)}(x) = [0.4142 \ -2.35]x. \tag{21}$$

The cost function parameters, namely the Q and R matrices, were chosen to be identity matrices of appropriate dimensions. The following smooth function with 15 unknown parameters was used to approximate the cost function of the system

$$\begin{aligned} V(x_1, x_2) = {} & w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 + w_4 x_1^4 + w_5 x_1^3 x_2 + \\ & + w_6 x_1^2 x_2^2 + w_7 x_1 x_2^3 + w_8 x_2^4 + w_9 x_1^6 + w_{10} x_1^5 x_2 + w_{11} x_1^4 x_2^2 + \\ & + w_{12} x_1^3 x_2^3 + w_{13} x_1^2 x_2^4 + w_{14} x_1 x_2^5 + w_{15} x_2^6. \end{aligned} \tag{22}$$

In order to solve online for the neural network weights $w_i, i = \overline{1,15}$ which parameterize the cost function, before each iteration step one needs to set up a least squares problem with the solution given by (19). As the considered neural network has 15 weights we can set up a solvable least squares problem by measuring the cost function associated with a given control policy over 15 time intervals, together with the initial state and the final state at each time interval. The result of applying the algorithm is presented in Fig. 3. The cost function converged to

$$V(x_1, x_2) = 3.3784 x_1^2 - 0.8284 x_1 x_2 + 2.6818 x_2^2, \tag{23}$$

the last 12 parameters being close to zero. The resulting control policy is

$$\mu_5(x) = 0.4142 x_1 - 2.6818 x_2. \tag{24}$$

This result is consistent with the solution of the Riccati equation underlying the optimal control problem in the linear case.

From Fig. 3 it is clear that the parameters of the cost function, and implicitly the parameters of the control policy, converged after two iteration steps were performed. In other words, after two iteration steps the system will be controlled in an optimal fashion; the parameters of the controller have been obtained online without using knowledge of the system's internal dynamics.

The adaptive optimal control algorithm was then used to determine the optimal controller for the nonlinear system (20). The required initial stabilizing controller for this system was (21). The cost function was approximated as in (22). The evolution of the cost function parameters is presented in Fig. 4. The obtained optimal controller is

$$\begin{aligned} \mu_5(x) = {} & 0.4142 x_1 - 2.6849 x_2 - 0.1924 x_1^3 + 1.3401 x_1^2 x_2 + \\ & + 0.8078 x_1 x_2^2 - 0.3481 x_2^3 + 0.2038 x_1^5 - 1.018 x_1^4 x_2 + \\ & + 0.8618 x_1^3 x_2^2 - 1.6026 x_1^2 x_2^3 + 0.122 x_1 x_{.2}^4 - 0.5628 x_2^5. \end{aligned} \tag{25}$$

Another experiment considering a cost function having terms up to the power 8 was also performed. The result, i.e., the weights corresponding to the high-order terms were close to zero, indicates that the sixth-order polynomial (22) provides a good approximation for the cost function.

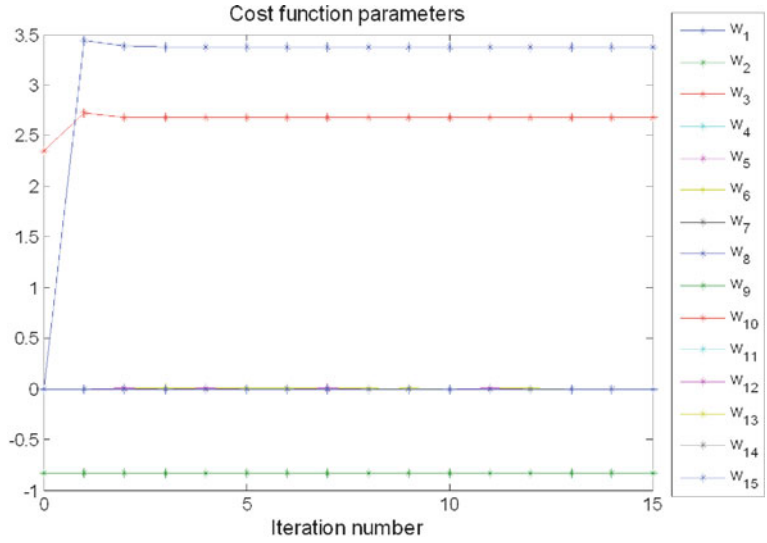


Fig. 3. Parameters of the cost function converging to the optimal values

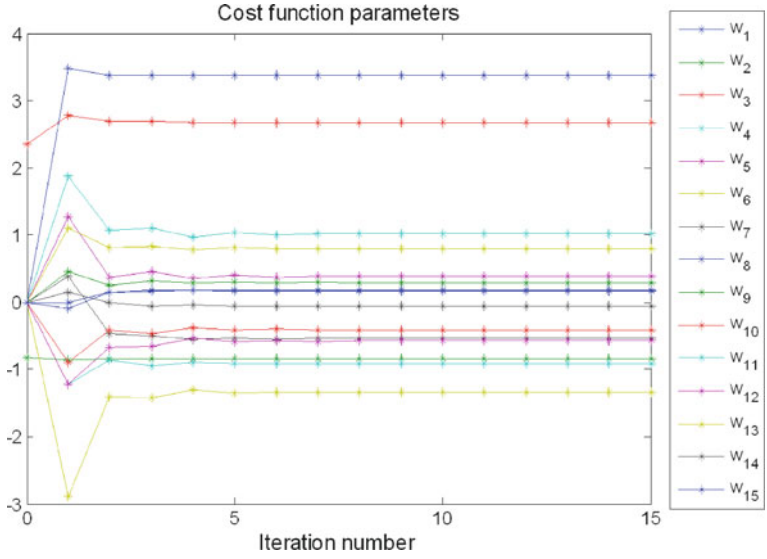


Fig. 4. Parameters of the cost function converging to the optimal values

6 Conclusions

We have presented a new adaptive controller based on a reinforcement learning algorithm, namely policy iteration, to solve online the continuous-time optimal control problem without using knowledge about the system's internal dynamics. Convergence of the proposed algorithm, under the condition

of initial stabilizing controller, to the solution of the optimal control problem has been established. Simulation results support the effectiveness of the online adaptive optimal controller.

Acknowledgment

The research was supported by the National Science Foundation ECS-0501451, ECCS-0801330 and the Army Research Office W911NF-05-1-0314.

References

1. Abu-Khalaf, M., Lewis, F.L.: Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. Automatica 41(5), 779–791 (2005).
2. Beard, R., Saridis, G., Wen, J.: Galerkin approximations of the generalized hamilton-jacobi-bellman equation. Automatica 33(12), 2159–2177 (1997).
3. Beard, R., Saridis, G., Wen, J.: Approximate solutions to the time-invariant hamilton-jacobi-bellman equation. J. Optim. Theor. Appl. 96(3), 589–626 (1998).
4. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming, Athena Scientific, Nashua, NH (1996).
5. Bradtke, S.J., Ydestie, B.E., Barto, A.G.: Adaptive linear quadratic control using policy iteration., *ACC Proceedings* doi: 10.1109/ACC.1994.735224 (1994).
6. Doya, K.: Reinforcement learning in continuous time and space. Neural Comput. 12(1), 219–245 (2000).
7. Feldbaum, A.A.: Dual control theory i-ii. Autom. Remote Control 21 (874–880), 1033–1039 (1960).
8. Filatov, N.M., Unbehauen, H.: Survey of adaptive dual control methods, IEE *Proceedings of the Control Theory and Applications* doi: 10.1049/ip-cta:20000107 (2000).
9. Hanselmann, T., Noakes, L., Zaknich, A.: Continuous-time adaptive critics. IEEE Trans. Neural Netw. 18(3), 631–647 (2007).
10. Hewer, G.: An iterative technique for the computation of the steady state gains for the discrete optimal regulator. IEEE Trans. Automat. Cont. 16, 382–384 (1971).
11. Hornik, K., Stinchcombe, M., White, H.: Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Netw. 3, 551–560 (1990).
12. Howard, R.A.: Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA (1960).
13. Kleinman, D.: On an iterative technique for riccati equation computations. IEEE Trans. Automat. Cont. 13, 114–115 (1968).
14. Leake, R.J., Liu, R.-W.: Construction of suboptimal control sequences, J.SIAM Cont. 5(1), 54–63 (1967).
15. Lewis, F., Syrmos, V.: Optimal Control, Wiley, New York, (1995).

16. Li, Z.H., Krstic, M.: Optimal design of adaptive tracking controllers for nonlinear systems. *ACC Proceedings* doi: 10.1109/ACC.1997.609721 (1997).
17. Miller, W.T., Sutton, R., Werbos, P.: Neural Networks for Control, MIT, Cambridge, MA (1990).
18. Murray, J.J., Cox, C.J., Lendaris, G.G., Saeks, R.: Adaptive dynamic programming. IEEE Trans. Syst. Man Cybern. 32(2), 140–153 (2002).
19. Prokhorov, D., Wunsch, D.: Adaptive critic designs. IEEE Trans. Neur. Netw. 8(5), 997–1007 (1997).
20. Saridis, G., Lee, C.S.: An approximation theory of optimal control for trainable manipulators. IEEE Trans. Syst. Man Cybern. 9(3), 152–159 (1979).
21. Slotine, J.J., Li, W.: Applied Nonlinear Control Prentice Hall, Englewood Cliffs, NJ (1991).
22. Sutton, R.S., Barto, A.G.: Reinforcement Learning – An Introduction, MIT, Cambridge, MA (1998).
23. Vrabie, D., Lewis, F.: Adaptive optimal control algorithm for continuous-time nonlinear systems based on policy iteration, *CDC Proceedings* 73–79, (2008).
24. Vrabie, D., Pastravanu, O., Lewis, F.L.: Policy iteration for continuous-time systems with unknown internal dynamics, *MED Proceedings* doi: 10.1109 / MED. 2007. 4433882 (2007).
25. Watkins, C.J.C.H.: Learning from Delayed Rewards, Phd thesis, University of Cambridge, England (1989).
26. Werbos, P.: Neural networks for control and system identification, *CDC Proceedings* doi: 10.1109/CDC.1989.70114 (1989).
27. Werbos, P.: Approximate dynamic programming for real-time control and neural modeling. In: D.A. White, D.A. Sofge (Eds.), Handbook of Intelligent Control 493–525 (1992).
28. Wittenmark, B.: Adaptive dual control methods: An overview, 5th IFAC Symp. on Adaptive Systems in Control and Signal Processing 67–73, (1995).

Perturbation Methods in Optimal Control Problems

Alexander S. Buldaev

Buryat State University, Ulan-Ude, Russia
buldaev@bsu.ru

Summary. We propose a new numerical approach for polynomial and other nonlinear optimal control problems including problems with time delays. The approach is based on the procedure of perturbation of the conditions of nonlocal improvement and the conditions of control's optimality. The suggested iterative perturbation methods possess characteristic nonlocal improvements of control, do not require parametric search of the improving approximations on each iteration, and have possibility for strong improvement of non-optimal controls satisfied to Pontryagin's maximum principle.

Key words: control system, improvement of control, condition for improvement, perturbation method

1 Introduction

The current variety of optimal control methods is caused by continuously arising demands of applications in many fields of science, techniques, and economics. Applied problems differ from one another by such distinct peculiarities as dimension of state spaces, types of nonlinearities, structure of restrictions, multi-extremality, singularity. It is hard to expect a universal calculating procedure sufficiently effective for solving various control problems appearing. That is why it is actual and justified to elaborate specialized optimal control methods, directed to consideration of peculiarities of applied problem classes.

Historically, the development of calculating methods in optimal control problems is closely connected with theory of necessary and sufficient optimality conditions. This development is also connected with obtaining different constructions and approximations of target functionals. The following basic directions can be extracted from existing approaches in this field:

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_17,

1. improvement methods in control space, characterized by operation of weak or needle-shaped control variation (gradient procedures, maximum principle methods, extension principle methods);
2. variation methods of a controlled process in space of variables in state and control, to which methods for solving boundary-value problem of the maximum principle, quasigradient procedures, procedures and methods of phase approximation of functional can be referred;
3. finite-difference approximation methods on the basis of partial or full digitization of control, and state problem with reduction to technology of mathematical programming.

These directions were developed by A.V. Arguchintsev, V.A. Baturin, O.V. Vasil'yev, F.P. Vasil'yev, R. Gabasov, V.I. Gurman, Yu.M. Danilin, V.F. Dem'yanov, V.V. Dikusar, Yu.G. Evtushenko, Yu.M. Ermol'yev, F.M. Kirillova, N.E. Kirin, V.F. Krotov, I.A. Krylov, A.A. Lyubushin, A.A. Milyutin, N.N. Moiseev, A.I.Moskalenko, D.A. Ovsyannikov, B.N. Pshenichniy, A.M. Rubinov, V.A. Srochko, R.P. Fedorenko, F.L. Chernous'ko, D. Mayne, E. Polak, K.L. Teo, L.T. Yeo, and many other researchers.

As for alternative directions, let us note the following:

- group of nonclassical methods of search for programmed and positional optimal controls for linear and other system classes. These methods are offered in works by R. Gabasov and F.M. Kirillova;
- methods for solving problems with impulse controls and discontinuous trajectories (V.I. Gurman, V.A. Dykhta, S.T. Zavalishchin, B.M. Miller, A.N. Sesekin, and others);
- global optimization methods in nonconvex problems with special structure, constructed in works by A.S. Strekalovskiy;
- variation methods for solving certain problem classes of mathematical physics, represented as optimal control problems for initial boundary conditions (V.I. Agoshkov, ZH.-L. Lions, G.I. Marchuk, V.P. Shutyaev, and others).

Algorithmic and program software of optimal control methods together with numerical solving of test and model problems were considered in works by Yu.G. Evtushenko, A.I. Tyatyuchkin, R.P. Fedorenko, and others.

In recent years the methods for nonlocal control improvement in systems, linear with respect to state, were developed in works by V.A. Srochko and his disciples. These methods are based on nonstandard formulas for the increment of the functional without remainder term (exact formulas). The complexity of nonlocal improvement is determined at the cost of solving two Cauchy problems. Absence of operation of parametric control variation on each iteration along with possibility of extremum controls improvement stipulates increased efficiency of constructed methods. It is actual and perspective to develop this direction on the way of constructing nonlocal improvement methods for class of optimal control problems that are quadratic and total polynomial with respect to state.

2 The Perturbation Methods in Optimal Control Problems That Are Polynomial with Respect to State

The perturbation methods are developed and modified for optimal control problems that are polynomial with respect to state. These methods are widely used for solving nonlinear problems of mathematical physics. The perturbation methods are considered in the context of a problem that is quadratic with respect to state. The following objects of perturbations are proposed to use:

- boundary-value problems of nonlocal improvement;
- conditions which are equivalent to boundary-value improvement problems in control space.

The proposed approach easily generalizes to problems polynomial in state, including problems with time delay.

2.1 The Optimal Control Problem That Is Polynomial with Respect to State

We consider that optimal control problem

$$\Phi(u) = \varphi(x(t_1)) + \int_T F(x(t), u(t), t)dt \to \min_{u \in V}, \tag{1}$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad x(t_0) = x^0, u(t) \in U, \quad t \in T = [t_0, t_1], \tag{2}$$

where $x(t) = (x_1(t), ..., x_n(t))$ is the state vector, $u(t) = (u_1(t), ..., u_m(t))$ is the control vector. The vector-valued function $f(x, u, t)$ and the function $F(x, u, t)$ are polynomial of degree of integer $k \geq 1$ with respect to variable x with coefficients continuously depending on u, t, on the set $R^n \times U \times T$, the function $\varphi(x)$ is polynomial of degree k in R^n. For admissible controls $u(t)$, $t \in T$, the set V of piecewise continuous functions with values in the compact set $U \subset R^m$ is considered. The initial state x^0 and the control interval T are given.

We use the following notations: q_x, q_u, q_{xx}, q_{uu}, q_{xu} are the first and the second partial derivatives of function q with respect to corresponding arguments; $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ is a scalar product of vectors x, y in Euclidean space E^n; $\|x\|$ is a norm of the vector x in Euclidean space; A^T is the transpose of the matrix A.

Let us introduce the Pontryagin's function

$$H(\psi, x, u, t) = \langle \psi, f(x, u, t) \rangle - F(x, u, t),$$

with $\psi \in R^n$ being the adjoint variable and standard adjoint vector system

$$\dot{\psi}(t) = -H_x(\psi(t), x(t), u(t), t), \quad t \in T. \tag{3}$$

For admissible control $v \in V$, by $x(t,v)$, $t \in T$, we denote a solution of the system (2) as $u(t) = v(t)$, $x(t_0,v) = x^0$; by $\psi(t,v)$, $t \in T$, we denote a solution of the system (3) as $u(t) = v(t)$, $x(t) = x(t,v)$, $\psi(t_1,v) = \varphi_x(x(t_1,v))$. Introduce a mapping u^* using the following relation:

$$u^*(\psi,x,t) = \arg\max_{w\in U} H(\psi,x,w,t), \quad \psi \in R^n,\ x \in R^n,\ t \in T. \tag{4}$$

Suppose by analogy with [1], the formula (4) defines the vector-valued function $u^*(\psi,x,t)$. This vector-valued function is piecewise continuous on $R^n \times R^n \times T$, i.e. this function has a finite number of discontinuity surfaces. Each discontinuity surface is prescribed by the following equation $s(\psi,x,t) = 0$, where $s(\psi,x,t)$ is differentiable in arguments ψ, x and continuous with respect to t on $R^n \times R^n \times T$. Assume that in the considered problem class operation for the maximum (4) admits an analytical solving, i.e., the control $u^*(\psi,x,t)$ has the explicit form of the corresponding formula.

The known necessary optimality control condition for $u \in V$ in the form of maximum principle [2], using mapping (4), may be represented as

$$u(t) = u^*(\psi(t,u),x(t,u),t), \quad t \in T. \tag{5}$$

Here and in what follows, equalities on the set T for admissible controls are interpreted accurate to sets of zero measure.

Let us extract the subclass of problems which are linear with respect to control. This subclass is important for applications described in

$$\Phi(u) = \varphi(x(t_1)) + \int_T (\langle a(x(t),t),u(t)\rangle + d(x(t),t))dt \to \min_{u\in V}, \tag{6}$$

$$\dot{x}(t) = A(x(t),t)u(t) + b(x(t),t),\ x(t_0) = x^0,\ u(t) \in U,\ t \in T. \tag{7}$$

The matrix function $A(x,t)$, vector-valued functions $b(x,t)$ and $a(x,t)$, and functions $\varphi(x)$, $d(x,t)$ are polynomial with respect to x and continuous with respect to t on the set $R^n \times T$. $U \subset R^m$ is a convex compact set.

In the problem (6) and (7) the Pontryagin's function has the following structure:

$$H(\psi,x,u,t) = H_0(\psi,x,t) + \langle H_1(\psi,x,t),u\rangle,$$

$$H_0(\psi,x,t) = \langle \psi, b(x,t)\rangle - d(x,t), \quad H_1(\psi,x,t) = A^T(x,t)\psi - a(x,t).$$

The mapping u^* is represented in the following form:

$$u^*(\psi,x,t) = \arg\max_{w\in U} \langle H_1(\psi,x,t),w\rangle.$$

In particular, for scalar control ($m = 1$) with the domain $U = [u^-,u^+]$ (bilateral constraints), we have

$$u^*(\psi,x,t) = \begin{cases} u^-, & H_1(\psi,x,t) < 0, \\ u^+, & H_1(\psi,x,t) > 0, \\ w \in U, & H_1(\psi,x,t) = 0. \end{cases}$$

Here, if $U = [-l, l]$, then the mapping u^* can be represented in the form $u^*(\psi, x, t) = l \cdot sign(H_1(\psi, x, t))$.

The maximum principle (5) for the control $u \in V$ in the problem (6) and (7) can be written in the form

$$\langle H_1(\psi(t,u), x(t,u), t), w - u(t)\rangle \le 0, \quad w \in U, \quad t \in T. \tag{8}$$

Let P_U be a projection operator on the set U in Euclidean norm:

$$P_U(z) = \arg \min_{w \in U}(\|w - z\|), \quad z \in R^m.$$

By the analogy with [1], for admissible control $u \in V$, let us form the vector-valued function u^α with the parameter $\alpha > 0$ using the relation

$$u^\alpha(\psi, x, t) = P_U\left(u(t) + \alpha H_1(\psi, x, t)\right), \ x \in R^n, \ \psi \in R^n, \ t \in T. \tag{9}$$

In view of fulfillment of the Lipschitz condition for the operator P_U, the function u^α is continuous in $(\psi, x) \in R^n \times R^n$ and piecewise continuous with respect to $t \in T$. According to the known projection property, the following inequality exists:

$$\langle H_1(\psi, x, t), u^\alpha(\psi, x, t) - u(t)\rangle \ge \frac{1}{\alpha}\left\|u^\alpha(\psi, x, t) - u(t)\right\|^2. \tag{10}$$

The function u^α has equivalent representation

$$u^\alpha(\psi, x, t) = \arg\max_{w \in U}\left(H(\psi, x, w, t) - \frac{1}{2\alpha}\left\|w - u(t)\right\|^2\right). \tag{11}$$

Using the function u^α the maximum principle (8) for the control $u \in V$ in the problem (6) and (7) can be written in the following form:

$$u(t) = u^\alpha(\psi(t,u), x(t,u), t), \quad t \in T. \tag{12}$$

Note that, to fulfill the maximum principle (8), it is sufficient to examine the condition (12), at least for one $\alpha > 0$. Conversely, from the condition (8), it follows that the condition (12) is fulfilled for any $\alpha > 0$.

Let us introduce a modified adjoint vector system

$$\dot{p}(t) = -H_x - \frac{1}{2!}\langle H_x, z\rangle_x - \cdots - \frac{1}{k!}\langle \cdots \langle\langle H_x, z\rangle_x, z\rangle_x \cdots, z\rangle_x. \tag{13}$$

For admissible controls u, v designate by $p(t, u, v)$, $t \in T$ – a solution of system (13), for $\psi = p(t)$, $x = x(t, u)$, $u = u(t)$, $z = x(t, v) - x(t, u)$ satisfying the boundary condition

$$p(t_1, u, v) = -\varphi_x - \frac{1}{2!}\langle \varphi_x, z\rangle_x - \cdots - \frac{1}{k!}\langle \cdots \langle\langle \varphi_x, z\rangle_x, z\rangle_x \cdots, z\rangle_x,$$

where partial derivatives with respect to x are calculated as $x = x(t_1, u)$ and $z = x(t_1, v) - x(t_1, u)$. It is evident that $p(t, u, u) = \psi(t, u)$, $t \in T$.

Using modification of adjoint system in the problem (1) and (2), symmetric formulas for the increment of the functional without remainder term of the Taylor series expansion can be obtained [3, 4]:

$$\Delta_v \Phi(u) = -\int_T \Delta_{v(t)} H(p(t,u,v), x(t,v), u(t), t)dt,$$

$$\Delta_v \Phi(u) = -\int_T \Delta_{v(t)} H(p(t,v,u), x(t,u), u(t), t)dt.$$

These formulas are the basis for construction of nonlocal improvement methods.

2.2 Perturbation Method for Boundary-Value Improvement Problem

Let us set the improvement control problem for $u^0 \in V$ with respect to functional (1): to find a control $v \in V$ satisfying the condition $\Phi(v) \leq \Phi(u^0)$.

As it is shown in works [3, 4], nonlocal control improvement can be provided at the cost of solving of special double-point boundary-value problem for system of ordinary differential equations. This problem is much easier than boundary-value problem of maximum principle. As a rule application of standard methods for numerical solving of boundary-value improvement problems (shooting method, linearization method, finite-difference method) leads to computational instability. This instability is caused by presence of positive real eigenvalues of Jacobi matrix and possible discontinuity of right-hand sides of the problem.

The perturbation methods are illustrated for the quadratic state optimal control problem (1) and (2). In this case the boundary-value improvement problem based on the map $u^\star$ has the following form [3]:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u^*(p(t), x(t), t), t), \quad x(t_0) = x^0, \\ \dot{p}(t) &= -H_x(p(t), x(t, u^0), u^0(t), t) - \\ &\quad -\frac{1}{2} H_{xx}(p(t), x(t, u^0), u^0(t), t)(x(t) - x(t, u^0)), \end{aligned} \tag{14}$$

$$p(t_1) = -\varphi_x(x(t_1, u^0)) - \frac{1}{2}\varphi_{xx}(x(t_1, u^0))(x(t_1) - x(t_1, u^0)). \tag{15}$$

Assume that the solution $(x(t), p(t)), \quad t \in T$, of the boundary-value problem (14) and (15) (probably, not unique) exists on the interval T and formed control $v(t) = u^*(p(t), x(t), t), \quad t \in T$ is piecewise continuous. Then $\Delta_v \Phi(u^0) \leq 0$. Note nonlocal character of improvement: parameter, characterizing a proximity of improving and being improved control is absent.

For the problem, linear with respect to state (1) and (2) (functions $f(x,u,t)$, $F(x,u,t)$, $\varphi(x)$ are linear with respect to x), the boundary-value

problem (14) and (15) split in to two independent Cauchy problems for adjoint and phase systems of variables. Here the solution of the adjoint system coincides with the solution of standard adjoint system $\psi(t, u^0), \quad t \in T$. Let $p^s(t) = \psi(t, u^0), \quad t \in T$.

In this case for the phase system we obtain Cauchy problem that is discontinuous with respect to state.

$$\dot{x}(t) = f(x(t), u^*(\psi(t, u^0), x(t), t), t), \ x(t_0) = x^0, \ t \in T. \tag{16}$$

Assume that the problem (16) has (probably, not unique) the solution $x^s(t), \quad t \in T$.

We will now consider a nonlinear, quadratic with respect to state, problem (1) and (2). Designate by $C(T)$ a space of vector-valued functions discontinuous on T with the norm $\|x\|_C = \max_{t \in T} \|x(t)\|$.

Assume that for continuous function $p(t)$, $t \in T$, belonging to a certain ball $B(p^s, l) = \{p \in C(T) : \|p - p^s\|_C \leq l\}$ of radius $l > 0$ in the space $C(T)$ centered at a point p^s, there is a solution $x^*(t, p), \quad t \in T$, of the phase system

$$\dot{x}(t) = f(x(t), u^*(p(t), x(t), t), t), \quad x(t_0) = x^0, \quad t \in T.$$

Assume that the corresponding operator X^*, defined by the relation

$$X^*(p) = x^*, \quad p \in C(T), \quad x^*(t) = x^*(t, p), \quad t \in T,$$

satisfies the Lipschitz condition in the ball $B(p^s, l)$ with a constant $M = M(p^s, l) > 0$

$$\|X^*(p) - X^*(q)\|_C \leq M \|p - q\|_C, \quad p \in B(p^s, l), \quad q \in B(p^s, l).$$

The Lipschitz condition guarantees the uniqueness of the solution $x^*(t, p)$, $t \in T$ of the phase system for $p \in B(p^s, l)$.

The operator X^* induces the corresponding operator $p \to x^*(t_1, p)$. It is evident that this operator also satisfies the Lipschitz condition in the ball $B(p^s, l)$ with a constant $M = M(p^s, l) > 0$

$$\|x(t_1, p) - x(t_1, q)\| \leq M \|p - q\|_C, \quad p \in B(p^s, l), \quad q \in B(p^s, l).$$

Using the operator X^* the boundary-value problem (14) and (16) in pair neighborhood (x^s, p^s) amount to Cauchy problem for the adjoint system with the right-hand side continuously dependent on adjoint variables

$$\begin{aligned} \dot{p}(t) = \ & -H_x(p(t), x(t, u^0), u^0(t), t) - \\ & -\frac{1}{2} H_{xx}(p(t), x(t, u^0), u^0(t), t)(x^*(t, p) - x(t, u^0)), \end{aligned} \tag{17}$$

$$p(t_1) = -\varphi_x(x(t_1, u^0)) - \frac{1}{2} \varphi_{xx}(x(t_1, u^0))(x^*(t_1, p) - x(t_1, u^0)). \tag{18}$$

Represent the problem (17) and (18) in the form of evolutionary problem with analogy to [5–7],

$$\dot{p}(t) + A(t)p(t) + G(p)|_t = h(t), \quad t \in T, \tag{19}$$

$$p(t_1) + D(p) = d, \tag{20}$$

$$A(t) = f_x^T(x(t,u^0),u^0(t),t),\ h(t) = F_{xx}(x(t,u^0),u^0(t),t),$$

$$G(p)|_t = \frac{1}{2}H_{xx}(p(t),x(t,u^0),u^0(t),t)(x^*(t,p) - x(t,u^0)),$$

$$D(p) = \frac{1}{2}\varphi_{xx}(x(t_1,u^0))(x^*(t_1,p) - x(t_1,u^0)), \quad d = -\varphi_x(x(t_1,u^0)).$$

On the strength of made assumptions, nonlinear operators G and D satisfy the Lipschitz condition in the ball $B(p^s,l)$ with certain constant $M_0 = M_0(p^s,l) > 0$

$$\|G(p) - G(q)\|_C \le M_0 \|p - q\|_C,$$

$$\|D(p) - D(q)\| \le M_0 \|p - q\|_C,\ p,q \in B(p^s,l). \tag{21}$$

Let us introduce a perturbed evolutionary problem with parameter $\varepsilon \in [0,1]$.

$$\dot{p}(t) + A(t)p(t) + \varepsilon G(p)|_t = h(t), \quad t \in T, \tag{22}$$

$$p(t_1) + \varepsilon D(p) = d. \tag{23}$$

The perturbed boundary-value improvement problem

$$\dot{x}(t) = f(x(t),u^*(p(t),x(t),t),t), \quad x(t_0) = x^0,$$

$$\dot{p}(t) = -H_x(p(t),x(t,u^0),u^0(t),t) -$$

$$-\varepsilon\frac{1}{2}H_{xx}(p(t),x(t,u^0),u^0(t),t)(x(t) - x(t,u^0)),$$

$$p(t_1) = -\varphi_x(x(t_1,u^0)) - \varepsilon\frac{1}{2}\varphi_{xx}(x(t_1,u^0))(x(t_1) - x(t_1,u^0)).$$

corresponds to the problem (22) and (23).

The unperturbed evolutionary problem is resulting from (22) and (23) as $\varepsilon = 0$ and has the following form:

$$\dot{p}(t) + A(t)p(t) = h(t), \quad t \in T, \tag{24}$$

$$p(t_1) = d. \tag{25}$$

The corresponding unperturbed boundary-value improvement problem is obtained from the perturbed boundary-value problem as $\varepsilon = 0$.

It is clear that a solution of the unperturbed problem (24) and (25) coincides with the solution $p^s(t) = \psi(t, u^0)$, $t \in T$. For the problem (24) and (25) the following decidability statement [5–7], formulated as lemma, is correct.

Lemma 1. *For any continuous functions $A(t)$, $h(t)$, $t \in T$ and any vector d of the problem* (24) *and* (25) *a unique continuous solution $p(t)$, $t \in T$, exists. For this solution the estimate exists*

$$\|p\|_C \le C_0(\|h\|_C + \|d\|), \quad C_0 = \text{const} > 0. \tag{26}$$

Using this statement, the theorem of perturbed problem decidability (22) and (23) and convergence method of successive approximations for problem solution can be proved [4].

Theorem 1. *Let operators G and D in the perturbed problem* (22) *and* (23) *satisfy the Lipschitz condition* (21) *in the ball $B(p^s, l)$ of radius l in the space $C(T)$ centered at p^s, where p^s is the unperturbed solution and $(\|G(p^s)\|_C + \|D(p^s)\|) \ne 0$.*

Then as $\varepsilon \le \bar{\varepsilon} = \left[C_0(2M_0 + \frac{1}{l}(\|G(p^s)\|_C + \|D(p^s)\|))\right]^{-1}$, where a constant $C_0 > 0$ is defined by the condition (26),

(1) the perturbed problem (22) and (23) has a unique solution $\bar{p} \in B(p^s, l)$;

(2) approximations p^k of the iterative process

$$\dot{p}^{k+1}(t) + A(t)p^{k+1} = -\varepsilon G(p^k)\big|_t, \quad t \in T, \tag{27}$$

$$p^{k+1}(t_1) = -\varepsilon D(p^k) \tag{28}$$

with initial $p^0 \in B(p^s, l)$ do not fall outside the limits of the ball $B(p^s, l)$ and converge to $\bar{p}$ in the norm $\|\cdot\|_C$;

(3) the estimate of iterative process convergence occurs

$$\|\bar{p} - p^k\|_C \le \|p^k - p^{k-1}\|_C \frac{\varepsilon C_0(2M_0)}{1 - \varepsilon C_0(2M_0)}. \tag{29}$$

As an initial approximation p^0 of the iterative process (27) and (28) under the condition $\|G(p^s)\|_C + \|D(p^s)\| \ne 0$, it is possible to choose the unperturbed solution p^s.

Remark 1. Let simultaneously $G(p^s) = 0$ and $D(p^s) = 0$, i.e., the unperturbed solution p^s is a solution of the perturbed problem (22) and (23) for any admissible $\varepsilon \in [0, 1]$. Then the statement of Theorem 1 is fulfilled for $\varepsilon < \bar{\varepsilon} = [C_0(2M_0)]^{-1}$. In this case there are no solutions of perturbed problem besides p^s in the ball $B(p^s, l)$ and for any initial approximation $p^0 \in B(p^s, l)$ the succession p^k converges to p^s.

If constant M_0 in (21) does not depend on radius l, i.e., the Lipschitz condition (21) is fulfilled on the whole set of admissible functions, then repeating our arguments similar to Theorem 1, we obtain the following statement [4].

Theorem 2. *Let operators G and D in the perturbed problem* (22) *and* (23) *satisfy the Lipschitz condition (21) on the whole set of admissible functions $C(T)$ and $\|G(p^s)\|_C + \|D(p^s)\| \neq 0$. Then as $\varepsilon \leq \bar{\varepsilon} = [C_0(2M_0)]^{-1}$, where constant $C_0 > 0$ is defined by the condition* (26),

(1) the perturbed problem (22) *and* (23) *has a unique solution $\bar{p} \in C(T)$ and in this case the estimate is correct $\|\bar{p} - p^s\|_C \leq \frac{1}{2M_0}(\|G(p^s)\|_C + \|D(p^s)\|)$;*

(2) the iterative process (27) *and* (28) *for any admissible initial approximation p^0 converges in the norm $\|\cdot\|_C$ to the solution $\bar{p}$;*

(3) the estimate (29) *of iterative process convergence occurs.*

Using proof of Theorem 1, we can obtain the statement of perturbed problem decidability with (22) and (23) and as $\varepsilon = 1$ [4].

Theorem 3. *Let operators G and D in the perturbed problem* (22) *and* (23) *satisfy the Lipschitz condition (21) in the ball $B(p^s, l)$ for any radius l with a constant dependent on l ($M_0 = M_0(p^s, l)$) and $\|G(p^s)\|_C + \|D(p^s)\| \neq 0$. Besides, additional conditions are fulfilled:* (1) *$G(0) = 0$, $F(0) = 0$;* (2) *as $l = \|p^s\|_C$ the inequality is fulfilled $C_0(2M_0) < \frac{1}{2}$, where constant $C_0 > 0$ is defined by the condition* (26). *Then*

(1) the perturbed problem (22) *and* (23) *as $\varepsilon = 1$ has a unique solution $\bar{p} \in B(p^s, l)$, and in this case the estimate $\|\bar{p}\|_C \leq 2C_0(\|h\|_C + \|d\|)$ is correct;*

(2) the iterative process (27) *and* (28) *as $\varepsilon = 1$ with any admissible initial approximation converges in the norm $\|\cdot\|_C$ to the solution $\bar{p}$;*

(3) the estimate (29) of iterative process convergence occurs.

Actually, if $G(0) = 0$, $F(0) = 0$, then from the Lipschitz condition (21) it follows that $\|G(p^s)\|_C \leq M_0 \|p^s\|_C$, $\|D(p^s)\| \leq M_0 \|p^s\|_C$. Therefore, for proof of Theorem 1 as $\bar{\varepsilon}$ it is possible to accept $\bar{\varepsilon} = \left[C_0(2M_0)(1 + \frac{1}{l}\|p^s\|_C)\right]^{-1}$. From here, as $l = \|p^s\|_C$, we obtain $\bar{\varepsilon} = [C_0(4M_0)]^{-1} > 1$. So, as $\varepsilon = 1$ the statement of theorem 1 is correct. In this case the estimate for $\bar{p}$ is obtained from the condition (26) and from the estimate $\|\bar{p} - p^s\|_C \leq l$.

Let us apply the perturbation method for boundary-value improvement problem based on the map u^α in problem quadratic in state (6) and (7)

$$\dot{x}(t) = f(x(t), u^\alpha(p(t), x(t), t), t), \quad x(t_0) = x^0, \tag{30}$$

$$\dot{p}(t) = -H_x(p(t), x(t, u^0), u^0(t), t) - \\ -\frac{1}{2}H_{xx}(p(t), x(t, u^0), u^0(t), t)(x(t) - x(t, u^0)),$$

$$p(t_1) = -\varphi_x(x(t_1, u^0)) - \frac{1}{2}\varphi_{xx}(x(t_1, u^0))(x(t_1) - x(t_1, u^0)). \tag{31}$$

Let $x^\alpha(t) = x(t, v^\alpha)$, $p^\alpha(t) = p(t, u^0, v^\alpha)$, $t \in T$, be a solution of this problem. Then the output control $v^\alpha(t) = u^\alpha(p^\alpha(t), x^\alpha(t), t)$, $t \in T$, provides lack of increase of target functional with the estimate

$$\Phi(v^\alpha) - \Phi(u^0) \le -\frac{1}{\alpha} \int_T \left\| v^\alpha(t) - u^0(t) \right\|^2 dt. \tag{32}$$

The perturbed boundary-value improvement problem with perturbation parameter $\varepsilon \in (0, 1]$ looks as follows:

$$\dot{x}(t) = f(x(t), u^\alpha(p(t), x(t), t), t), \quad x(t_0) = x^0, \tag{33}$$

$$\dot{p}(t) = -H_x(p(t), x(t, u^0), u^0(t), t) -$$

$$-\varepsilon \frac{1}{2} H_{xx}(p(t), x(t, u^0), u^0(t), t)(x(t) - x(t, u^0)),$$

$$p(t) = -\varphi_{xx}(x(t_1, u^0)) - \varepsilon \frac{1}{2} \varphi_{xx}(x(t_1, u^0))(x(t_1) - x(t_1, u^0)). \tag{34}$$

Here the unperturbed solution as $\varepsilon = 0$ is the pair $(x_0^\alpha(t), p_0^\alpha(t))$, $t \in T$, where $p_0^\alpha(t) = \psi(t, u^0)$, $t \in T$ and $x_0^\alpha(t)$, $\quad t \in T$ solutions are of the phase system

$$\dot{x}(t) = f(x(t), u^\alpha(\psi(t, u^0), x(t), t), t), \quad x(t_0) = x^0.$$

Note that on the strength of projection operator properties the unperturbed solution $x_0^\alpha(t)$, $t \in T$, exists and is unique. Similar to this, for arbitrary continuous function $p(t)$, $t \in T$, there exists a unique solution $x^\alpha(t, p)$, $\quad t \in T$, of the system

$$\dot{x}(t) = f(x(t), u^\alpha(p(t), x(t), t), t), \quad x(t_0) = x^0.$$

Introduce corresponding operator X^α applying the relation

$$X^\alpha(p) = x^\alpha, \quad p \in C(T), \quad x^\alpha(t) = x^\alpha(t, p), \quad t \in T.$$

Iterative method for solving the perturbed problem (33) and (34) has the form

$$\dot{x}^{k+1}(t) = f(x^{k+1}(t), u^\alpha(p^{k+1}(t), x^{k+1}(t), t), t), \; x^{k+1}(t_0) = x^0, \tag{35}$$

$$\dot{p}^{k+1}(t) = -H_x(p^{k+1}(t), x(t, u^0), u^0(t), t) -$$

$$-\varepsilon \frac{1}{2} H_{xx}(p^k(t), x(t, u^0), u^0(t), t)(x^k(t) - x(t, u^0)),$$

$$p^{k+1}(t_1) = -\varphi_x(x(t_1, u^0)) -$$

$$-\varepsilon \frac{1}{2} \varphi_{xx}(x(t_1, u^0))(x^k(t_1) - x(t_1, u^0)). \tag{36}$$

The initial approximation $x^0 \in C(T)$, $p^0 \in C(T)$ is defined. It is clear that $x^k(t) = x^\alpha(t, p^k)$, $t \in T$, $k > 0$.

As the initial approximation of iterative process (35) and (36) for improvement of control $u^0 \in V$, not satisfying the maximum principle, it is possible to choose the unperturbed solution $(x_0^\alpha(t), p_0^\alpha(t))$, $t \in T$.

Let us analyze the process convergence (35) and (36) in assumption about boundedness of the family of phase system trajectories (7)

$$x(t, u) \in X, \quad t \in T, \quad u \in V,$$

where $X \subset R^n$ is a convex compact set.

In this case, since function $f(x, u, t)$ is quadratic with respect to x, the Lipschitz condition is fulfilled

$$\|f(x, u, t) - f(y, u, t)\| \le M_1 \|x - y\|, \quad x, y \in X, \quad u \in U, \quad t \in T,$$

where $M_1 = \text{const} > 0$. Using the Gronwall–Bellman lemma [8, 9], it is easy to obtain the estimate

$$\|x(t, u) - x(t, v)\| \le M_2 \int_T \|u(t) - v(t)\| \, dt, \quad t \in T, \quad u \in V, \quad v \in V,$$

where $M_2 = \text{const} > 0$.

For the control $u^0 \in V$ and given $\alpha > 0$ introduce operator V^α using the relation

$$V^\alpha(p, x) = v^\alpha, \quad p \in C(T), \quad x \in C(T),$$
$$v^\alpha(t) = P_U(u^0(t) + \alpha H_1(p(t), x(t), t)), \quad t \in T.$$

Then the equality occurs

$$X^\alpha(p) = V^\alpha(p, X^\alpha(p)), \quad p \in C(T).$$

Hence on the basis of the Lipschitz condition for the projection operator P_U we obtain

$$\|x^\alpha(t, p) - x^\alpha(t, q)\| = \|x(t, V^\alpha(p, X^\alpha(p))) - x(t, V^\alpha(q, X^\alpha(q)))\| \le$$

$$\le M_2 \int_T \|V^\alpha(p, X^\alpha(p))|_t - V^\alpha(q, X^\alpha(q))|_t\| \, dt \le$$

$$\le \alpha M_3 \int_T \|H_1(p(t), x^\alpha(t, p), t) - H_1(q(t), x^\alpha(t, q), t)\| \, dt \le$$

$$\le \alpha M_0(\|p - q\|_C + \|x^\alpha(t, p) - x^\alpha(t, q)\|_C), \quad t \in T, \ p \in C(T), \ q \in C(T),$$

where $M_3 = \text{const} > 0$, $M_0 = \text{const} > 0$. Therefore, at sufficiently low $\alpha > 0$ we obtain

$$\|x^\alpha(t,p) - x^\alpha(t,q)\|_C \le \frac{\alpha M_0}{1-\alpha M_0}\|p-q\|_C\,,$$

$$\|x^\alpha(t_1,p) - x^\alpha(t_1,q)\| \le \frac{\alpha M_0}{1-\alpha M_0}\|p-q\|_C\,,$$

where $0 < \alpha M_0 < 1$. So at a sufficiently small $\alpha > 0$ the operator X^α and the corresponding operator $p \to x^\alpha(t_1,p)$ satisfy the Lipschitz condition in the space $C(T)$. Hence, similarly to Theorem 2, we obtain the statement [4].

Theorem 4. *Assume that in the problem* (6) *and* (7) *the boundedness condition for phase trajectories $x(t,u) \in X$, $t \in T$, $u \in V$, is fulfilled, where $X \subset R^n$ is a convex compact set. Then for sufficiently small $\alpha > 0$ as $0 < \varepsilon < \bar{\varepsilon} = C_0 \frac{1-\alpha M_0}{2\alpha M_0}$, $C_0 = \text{const} > 0$, $M_0 = \text{const} > 0$*

(1) the perturbed boundary value problem (33) *and* (34) *has a unique solution $\bar{x}^\alpha \in C(T)$, $\bar{p}^\alpha \in C(T)$;*

(2) the iterative process (35) *and* (36) *converges in the norm $\|\cdot\|_C$ to the solution $(\bar{x}^\alpha, \bar{p}^\alpha)$ of the perturbed problem* (33) *and* (34) *for any initial approximation $x^0 \in C(T)$, $p^0 \in C(T)$.*

From the theorem it follows that for the control u^0, that is satisfying the maximum principle, by virtue of uniqueness, the perturbed solution at small $\alpha > 0$ coincides with the unperturbed one.

Corollary 1. *Under conditions of theorem 4 at sufficiently small $\alpha > 0$*

(1) the perturbed problem (33) *and* (34) *as $\varepsilon = 1$ has a unique solution $\bar{x}^\alpha \in C(T)$, $\bar{p}^\alpha \in C(T)$;*

(2) the iterative process (35) *and* (36) *as $\varepsilon = 1$ with any initial admissible approximation $x^0 \in C(T)$, $p^0 \in C(T)$ converges in the norm $\|\cdot\|_C$ to the solution $(\bar{x}^\alpha, \bar{p}^\alpha)$.*

Similarly, it is possible to put and solve a perturbed boundary-value problem in modified methods of nonlocal improvement [3, 4] on the basis of operation for the maximum u^*.

On the whole, we note that iterative process convergence to solution of the perturbed problem is guaranteed only at sufficiently small perturbation parameters.

Perturbation methods admit a sequential conversion procedure of perturbed problems, those solutions under certain conditions can converge to solution of desired problem.

2.3 Transformation Method for Perturbed Boundary-Value Improvement Problems

In practice, it is difficult to estimate a priori a domain of convergence with respect to perturbation parameter $\varepsilon \in [0,1]$ for iterative solution process of

the perturbed problem (22) and (23). That is why, if unperturbed state of the problem essentially differs from the correct one, then in case of divergence for perturbation parameter $\varepsilon = 1$, it is possible to construct convergent process only for sufficiently small $\varepsilon > 0$. This means the application of the small perturbation theory. Obtained solution $p_1 \neq p^s$ as $0 < \varepsilon_1 < 1$ can be considered only as the first approximation to correct solution for $\varepsilon = 1$.

The way of approximation refining by means of the perturbation method is based on the assumption that obtained perturbed solution is possible more close to true solution than the unperturbed one. This is the transformation of the input problem (19) and (20), thus that obtained perturbed solution becomes unperturbed in the transformed problem. Let us transform the input problem (19) and (20) so, that obtained approximation p_1 as $\varepsilon_1 < 1$ becomes unperturbed in the problem. From (19), (20) and (22), (23) we obtain the problem that is equivalent to (19) and (20)

$$\dot{p}(t) + A(t)p(t) + \left(G(p)|_t - \varepsilon_1 G(p_1)|_t\right) = \dot{p}_1(t) + A(t)p_1(t), \tag{37}$$

$$p(t_1) + (D(p) - \varepsilon_1 D(p_1)) = p_1(t_1). \tag{38}$$

Let us transform the corresponding unperturbed and perturbed problem with parameter $\varepsilon > 0$

$$\dot{p}(t) + A(t)p(t) = \dot{p}_1(t) + A(t)p_1(t), \quad t \in T, \tag{39}$$

$$p(t_1) = p_1(t_1). \tag{40}$$

$$\dot{p}(t) + A(t)p(t) + \varepsilon(\, G(p)|_t - \varepsilon_1\, G(p_1)|_t) = \dot{p}_1(t) + A(t)p_1(t), \tag{41}$$

$$p(t_1) + \varepsilon(D(p) - \varepsilon_1 D(p_1)) = p_1(t_1). \tag{42}$$

It is clear that p_1 is the unique solution of the unperturbed problem (39) and (40). If there is no convergence as $\varepsilon = 1$ in (41) and (42), then we continue transformation process of the input problem (19) and (20).

Let p_2 be a solution of the perturbed problem (41) and (42) at certain $\varepsilon_2 < 1$. Then from (37), (38) and (41), (42) we obtain the problem that is equivalent to the input problem (19) and (20)

$$\begin{aligned} &\dot{p}(t) + A(t)p(t) + \left(G(p)|_t - \varepsilon_2 G(p_2)|_t - \right. \\ &\left. -\varepsilon_1(1-\varepsilon_2)\, G(p_1)|_t\right) = \dot{p}_2(t) + A(t)p_2(t), \end{aligned} \tag{43}$$

$$p(t_1) + (D(p) - \varepsilon_2 D(p_2) - \varepsilon_1(1-\varepsilon_2)D(p_1)) = p_2(t_1). \tag{44}$$

The unperturbed and perturbed problems that are corresponding to (43) and (44), with parameter $\varepsilon > 0$, have the following form:

$$\dot{p}(t) + A(t)p(t) = \dot{p}_2 t) + A(t)p_2(t), \quad t \in T, \tag{45}$$

$$p(t_1) = p_2(t_1), \tag{46}$$

$$\begin{aligned} &\dot{p}(t) + A(t)p(t) + \varepsilon\, \left(G(p)|_t - \varepsilon_2 G(p_2)|_t - \right. \\ &\left. -\varepsilon_1(1-\varepsilon_2)\, G(p_1)|_t\right) = \dot{p}_2(t) + A(t)p_2(t), \end{aligned} \tag{47}$$

$$p(t_1) + \varepsilon(D(p) - \varepsilon_2 D(p_2) - \varepsilon_1(1-\varepsilon_2)D(p_1)) = p_2(t_1), \tag{48}$$

where p_2 is a solution of the unperturbed problem (45) and (46).

Similarly, on the basis of the solution p_3 for the perturbed problem (47), (48), as $\varepsilon_3 < 1$ let us construct the perturbed problem

$$\dot{p}(t) + A(t)p(t) + \varepsilon\,(G(p)|_t - \varepsilon_3 G(p_3)|_t - \varepsilon_2(1-\varepsilon_3)\,G(p_2)|_t - \\ - \varepsilon_1(1-\varepsilon_2)(1-\varepsilon_3)G(p_1)|_t) = \dot{p}_3(t) + A(t)p_3(t),$$

$$p(t_1) + \varepsilon(D(p) - \varepsilon_3 D(p_3) - \varepsilon_2(1-\varepsilon_3)D(p_2) - \\ -\varepsilon_1(1-\varepsilon_2)(1-\varepsilon_3)D(p_1)) = p_3(t_1).$$

This problem has the unperturbed solution p_3, etc. If there exists a perturbation parameter $\varepsilon_k = 1$ at certain $k \geq 1$ that provides convergence, then the corresponding solution p_k is a solution of the input problem (19) and (20).

We will now show that under certain conditions a sequence of perturbation parameters $\varepsilon_k > 0$, providing convergence of the corresponding iterative process in kth perturbed problem, $k \geq 1$, can be chosen, so that the solution p_k of the perturbed problem tends to the solution of the input problem (19) and (20). Note that in this case it is possible to interpret the convergence as a convergence of the unperturbed solutions for transformed problems to true solution.

Consider the case of fulfillment of the Lipschitz condition (21) with one constant $M_0 > 0$ on the whole set of admissible functions p. According to Theorem 2, existence of solution and convergence to solution in sequence of perturbed transformed problems are guaranteed at perturbation parameter value that is satisfying the condition

$$0 < \varepsilon < \bar{\varepsilon} = [C_0(2M_0)]^{-1},$$

where constant $C_0 > 0$ is defined by the condition (26).

If $\bar{\varepsilon} > 1$ then the perturbed solution p_1 that is corresponding to parameter $\varepsilon_1 = 1$, is a solution of the input problem.

Let $\bar{\varepsilon} \leq 1$. In kth perturbed problem if the perturbation parameter $\varepsilon = 1$ does not provide convergence of the iterative process we set

$$0 < \varepsilon_0 \leq \varepsilon_k < \bar{\varepsilon},\ \varepsilon_0 \geq \bar{\varepsilon}(1-\delta), \tag{49}$$

where $1 > \delta > 0$ is fixed.

According to Theorem 2 we have the following estimates for solutions of perturbed problems p_k:

$$\|p_1 - p^s\|_C \leq \frac{1}{2M_0}(\|G(p^s)\|_C + \|D(p^s)\|),$$

$$\|p_2 - p_1\|_C \leq \frac{(1-\varepsilon_1)}{2M_0}(\|G(p_1)\|_C + \|D(p_1)\|),$$

$$\|p_3 - p_2\|_C \le \frac{(1-\varepsilon_2)}{2M_0}(\|G(p_2) - \varepsilon_1 G(p_1)\|_C + \|D(p_2) - \varepsilon_1 D(p_1)\|),$$

$$\|p_4 - p_3\|_C \le \tfrac{(1-\varepsilon_3)}{2M_0}(\|G(p_3) - \varepsilon_2 G(p_2) - \varepsilon_1(1-\varepsilon_2)G(p_1)\|_C + \\ + \|D(p_3) - \varepsilon_2 D(p_2) - \varepsilon_1(1-\varepsilon_2)D(p_1)\|),\ \dots .$$

It is convenient to introduce the following notation:

$$p_0 = p^s, \quad \alpha_k = \|p_k - p_{k-1}\|_C\,,\ b_0 = (\|G(p^s)\|_C + \|D(p^s)\|),$$

$$b_1 = (\|G(p_1)\|_C + \|D(p_1)\|),$$

$$b_2 = (\|G(p_2) - \varepsilon_1 G(p_1)\|_C + \|D(p_2) - \varepsilon_1 D(p_1)\|),$$

$$b_3 = (\|G(p_3) - \varepsilon_2 G(p_2) - \varepsilon_1(1-\varepsilon_2)G(p_1)\|_C + \\ + \|D(p_3) - \varepsilon_2 D(p_2) - \varepsilon_1(1-\varepsilon_2)D(p_1)\|),$$

etc. Using the Lipschitz condition we obtain

$$a_1 \le \frac{b_0}{2M_0}, \quad a_2 \le (1-\varepsilon_1)\frac{b_1}{2M_0},$$

$$a_3 \le (1-\varepsilon_2)\frac{b_2}{2M_0}, \quad a_4 \le (1-\varepsilon_3)\frac{b_3}{2M_0}, ...,$$

$$b_1 \le 2M_0 a_1 + b_0 \le 2b_0 \quad \Rightarrow \quad a_2 \le 2(1-\varepsilon_1)\frac{b_0}{2M_0},$$

$$b_2 \le 2M_0 a_2 + (1-\varepsilon_1)b_1 \le 2(1-\varepsilon_1)b_1 \quad \Rightarrow \quad a_3 \le 4(1-\varepsilon_2)(1-\varepsilon_1)\frac{b_0}{2M_0},$$

etc. Total estimate as $k \ge 0$ has the following form:

$$a_{k+1} \le 2^k(1-\varepsilon_k)(1-\varepsilon_{k-1})\cdots(1-\varepsilon_1)\frac{b_0}{2M_0} \le 2^k(1-\varepsilon_0)^k\frac{b_0}{2M_0}. \qquad (50)$$

From the obtained estimate (50) it follows that at $\varepsilon_0 > 2^{-1}$ the sequence p_k is fundamental and converges to $\bar{p} \in C(T)$. So, for $1 \ge \bar{\varepsilon} > 2^{-1}$ at choice ε_k according to the rule (49), where $\delta > 0$ is sufficiently small integer, sequence convergence of solutions p_k for perturbed problems to certain $\bar{p} \in C(T)$.

Performed analysis, using the upper estimate (50), illustrates the possibility of convergence of solutions p_k for perturbed problems when using the rule (49). Assume that the convergence condition of solutions p_k for perturbed problems is fulfilled at a choice of the rule (49).

The perturbed solution p_k is defined by the conditions

$$\dot{p}_k(t) + A(t)p_k(t) + \varepsilon_k(G(p_k)|_t - \varepsilon_{k-1}G(p_{k-1})|_t - \\ -\varepsilon_{k-2}\,(1-\varepsilon_{k-1})G(p_{k-2})|_t - \\ -\varepsilon_{k-3}(1-\varepsilon_{k-2})(1-\varepsilon_{k-1})\,G(p_{k-3})|_t - \cdots - \\ -\varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_{k-1})\,G(p_1)|_t) = \\ = \dot{p}_{k-1}(t) + A(t)p_{k-1}(t),$$

$$p_k(t_1) + \varepsilon_k(D(p_k) - \varepsilon_{k-1}D(p_{k-1}) - \varepsilon_{k-2}(1-\varepsilon_{k-1})D(p_{k-2}) - \\ -\varepsilon_{k-3}(1-\varepsilon_{k-2})(1-\varepsilon_{k-1})D(p_{k-3}) - \cdots - \\ -\varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_{k-1})D(p_1)) = p_{k-1}(t_1).$$

Hence, it follows that using assumption about sequence convergence p_k and the condition (49) we obtain as $k \to \infty$

$$g_k = (G(p_k) - \varepsilon_{k-1}G(p_{k-1}) - \varepsilon_{k-2}(1-\varepsilon_{k-1})G(p_{k-2}) - \cdots - \\ -\varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_{k-1})G(p_1)) \to 0,$$

$$d_k = (D(p_k) - \varepsilon_{k-1}D(p_{k-1}) - \varepsilon_{k-2}(1-\varepsilon_{k-1})D(p_{k-2}) - \cdots - \\ -\varepsilon_1(1-\varepsilon_2)...(1-\varepsilon_{k-1})D(p_1)) \to 0.$$

On the basis of obtained p_k let us consider the next$(k+1)$ transformed problem that is equivalent to the input problem (19) and (20)

$$\begin{aligned} &\dot{p}(t) + A(t)p(t) + G(p)|_t - \varepsilon_k G(p_k)|_t - \\ &- \varepsilon_{k-1}(1-\varepsilon_k)G(p_{k-1})|_t - \\ &- \varepsilon_{k-2}(1-\varepsilon_{k-1})(1-\varepsilon_k)G(p_{k-2})|_t - \cdots - \\ &- \varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_k)G(p_1)|_t = \\ &= \dot{p}_k(t) + A(t)p_k(t), \end{aligned} \tag{51}$$

$$\begin{aligned} &p(t_1) + D(p) - \varepsilon_k D(p_k) - \varepsilon_{k-1}(1-\varepsilon_k)D(p_{k-1}) - \\ &- \varepsilon_{k-2}(1-\varepsilon_{k-1})(1-\varepsilon_k)D(p_{k-2}) - \cdots - \\ &- \varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_k)D(p_1) = p_k(t_1). \end{aligned} \tag{52}$$

Let us set the following notation:

$$y_k = \varepsilon_k G(p_k) + \varepsilon_{k-1}(1-\varepsilon_k)G(p_{k-1}) + \\ +\varepsilon_{k-2}(1-\varepsilon_{k-1})(1-\varepsilon_k)G(p_{k-2}) + \cdots + \varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_k)G(p_1),$$
$$z_k = \varepsilon_k D(p_k) + \varepsilon_{k-1}(1-\varepsilon_k)D(p_{k-1}) + \\ +\varepsilon_{k-2}(1-\varepsilon_{k-1})(1-\varepsilon_k)D(p_{k-2}) + \cdots + \varepsilon_1(1-\varepsilon_2)\cdots(1-\varepsilon_k)D(p_1).$$

It is obvious, that the following relations are fulfilled:

$$y_k = \varepsilon_k G(p_k) + (1-\varepsilon_k)(G(p_k) - g_k) = G(p_k) - g_k(1-\varepsilon_k),$$

$$z_k = \varepsilon_k D(p_k) + (1-\varepsilon_k)(D(p_k) - d_k) = D(p_k) - d_k(1-\varepsilon_k).$$

Hence, we obtain a convergence in the corresponding norms

$$y_k \to G(p_k), \quad z_k \to D(p_k), \quad k \to \infty. \tag{53}$$

Since the input problem (19) and (20) is equivalent to transformed problem (51) and (52) we have

$$\dot{p}_k + A(t)p_k + y_k = h,$$

$$p_k(t_1) + z_k = d.$$

Hence from (53), as $k \to \infty$, it follows that

$$\dot{p}_k + A(t)p_k + G(p_k) \to h,$$

$$p_k(t_1) + D(p_k) \to d.$$

Therefore, $\bar{p}$ is a solution of the input problem.

So, by assumption of sequence convergence of solutions p_k for transformed perturbed problems under the condition of choice of perturbation parameters according the rule (49), it is possible to find solution of the input problem with arbitrarily high precision.

In real calculations we can use various rules of successive decrease of perturbation parameter from $\varepsilon = 1$ to a value, whereby, convergence of the iterative process occurs. Here it is possible to calculate the perturbed boundary-value problem till first input control improvement occurs.

2.4 Perturbation Method for Improvement Condition in Control Space

The optimal control problem, that is quadratic with respect to state (1) and (2) is considered. Define improvement condition in control space that is equivalent to the nonlocal boundary-value improvement problem (14) and (15) in state space.

Let $(x(t), p(t))$, $t \in T$, be a solution of the boundary-value problem (14) and (15) in state space. Then the admissible control $v(t) = u^*(p(t), x(t), t)$, $t \in T$, satisfies the condition

$$v(t) = u^*(p(t, u^0, v), x(t, v), t), \quad t \in T, \tag{54}$$

in control space. On the contrary, if $v(t)$, $t \in T$, is an admissible control that is satisfying the relation (54), then pair $(x(t, v), p(t, u^0, v))$, as $t \in T$, satisfies the boundary-value problem (14) and (15). So, the boundary-value improvement problem (14) and (15) in state space reduces to the condition (54) on the set of admissible controls V.

In the problem, linear with respect to state (1) and (2) for solving the improvement control problem $u^0 \in V$ it is sufficient to solve two Cauchy problems in state space. These problems are received from decomposition of the boundary-value improvement problem. Note that here the condition (54) has the following form

$$v(t) = u^*(\psi(t, u^0), x(t, v), t), \quad t \in T. \tag{55}$$

In the problem, nonlinear with respect to state (1) and (2) for improvement u^0 it is possible to use method for solving the relation (54) on the set of admissible controls.

The difficulties in realization of the condition (54) are analogous to difficulties in solving the corresponding boundary-value improvement problem. In common case these difficulties are connected with possible multi-valuedness and discontinuity of the mapping u^*.

In this section perturbation method for the condition (54) is justified. To solve the perturbed condition the method of successive approximations is used. On each iteration of this method the problem as complicated as unperturbed one is solved. As the unperturbed condition the relation of type (55) is defined. To solve this relation it is sufficient to solve two Cauchy problems. Let us isolate the linear part with respect to state from the nonlinear problem (1) and (2) and represent the problem in the form

$$\Phi(u) = \langle c_0, x(t_1)\rangle + \varphi_1(x(t_1)) + \int_T (\langle a_0(u(t),t), x(t)\rangle + \tag{56}$$
$$+ d_0(u(t),t) + F_1(x(t),u(t),t))dt \to \min_{u\in V},$$
$$\dot{x}(t) = A_0(u(t),t)x(t) + b_0(u(t),t) + f_1(x(t),u(t),t), \quad x(t_0) = x^0,$$
$$u(t) \in U, \quad t \in T = [t_0, t_1], \tag{57}$$

where the vector-valued function $f_1(x,u,t)$ and the function $F_1(x,u,t)$ are quadratic with respect to variable x and discontinuous in the variables u, t on the set $R^n \times U \times T$; the function $\varphi_1(x)$ is quadratic on R^n, the matrix function $A_0(u,t)$, the vector-valued functions $a_0(u,t)$, $b_0(u,t)$, the function $d_0(u,t)$ are continuous in the variables u, t on the set $U \times T$, c_0 is a constant vector.

The Pontryagin function and modified adjoint system in the problem (56) and (57) take the form

$$H(\psi,x,u,t) = \langle \psi, A_0(u,t)x + b_0(u,t) + f_1(x,u,t)\rangle -$$
$$- \langle a_0(u,t), x\rangle - d_0(u,t) - F_1(x,u,t),$$
$$\dot{p}(t) = - A_0^T(u(t),t)p(t) + a_0(u,t) - f_{1x}^T(x(t),u(t),t)p(t) +$$
$$+ F_{1x}(x(t),u(t),t) -$$
$$- \frac{1}{2}[f_{1x}^T(x(t),u(t),t)p(t) - F_{1x}(x(t),u(t),t)]_x y(t), t \in T,$$
$$p(t_1) = -c_0 - \varphi_{1x}(x(t_1)) - \frac{1}{2}\varphi_{1xx}(x(t_1))y(t_1).$$

Let us define the perturbed improvement condition with a perturbation parameter $\varepsilon \in [0,1]$

$$v(t) = u_\varepsilon^*(p_\varepsilon(t,u^0,v), x_\varepsilon(t,v), t), \quad t \in T, \tag{58}$$

where $x_\varepsilon(t,v)$, $t \in T$, is a solution of the perturbed phase system

$$\dot{x}(t) = A_0(u(t),t)x(t) + b_0(u(t),t) + \varepsilon f_1(x(t),u(t),t), \quad x(t_0) = x^0,$$

as $u(t) = v(t)$; and $p_\varepsilon(t,u^0,v)$, $t \in T$, is a solution of the perturbed adjoint system

$$\dot{p}(t) = -A_0^T(u(t),t)p(t) + a_0(u(t),t) - \varepsilon(f_{1x}^T(x(t),u(t),t)p(t) -$$
$$-F_{1x}(x(t),u(t),t) + \frac{1}{2}[f_{1x}^T(x(t),u(t),t)p(t) -$$
$$-F_{1x}(x(t),u(t),t)]_x y(t)),\ t \in T,$$
$$p(t_1) = -c_0 - \varepsilon(\varphi_{1x}(x(t_1)) + \frac{1}{2}\varphi_{1xx}(x(t_1))y(t_1))$$

as $u(t) = u^0(t)$, $x(t) = x(t,u^0)$, $y(t) = x(t,v) - x(t,u^0)$. The perturbed mapping u_ε^* is formed by the perturbed Pontryagin function.

$$H_\varepsilon(\psi,x,u,t) = \langle\psi, A_0(u,t)x + b_0(u,t)\rangle - \langle a_0(u,t),x\rangle - d_0(u,t) +$$
$$+\varepsilon(\langle\psi, f_1(x,u,t)\rangle - F_1(x,u,t)),$$

using the formula

$$u_\varepsilon^*(\psi,x,t) = \arg\max_{w\in U} H_\varepsilon(\psi,x,w,t)', \quad \psi \in R^n, \quad x \in R^n, \quad t \in T.$$

So, the perturbed condition (58) is an improvement condition for the perturbed optimal control problem

$$\Phi_\varepsilon(u) = \langle c_0, x(t_1)\rangle + \varepsilon\varphi_1(x(t_1)) +$$
$$+\int_T (\langle a_0(u(t),t),x(t)\rangle +$$
$$+d_0(u(t),t) + \varepsilon F_1(x(t),u(t),t))dt \to \min_{u\in V}, \tag{59}$$
$$\dot{x}(t) = A_0(u(t),t)x(t) + b_0(u,t) + \varepsilon f_1(x(t),u(t),t), \quad x(t_0) = x^0,$$
$$u(t) \in U, \quad t \in T = [t_0,t_1]. \tag{60}$$

The input optimal control problem (56) and (57) corresponds to the perturbed problem (59) and (60) for $\varepsilon = 1$.

Unperturbed improvement condition is obtained from the perturbed one (58) as $\varepsilon = 0$, and has the form

$$v(t) = u_0^*(p_0(t,u^0),x_0(t,v),t), \quad t \in T, \tag{61}$$

where $x_0(t,v)$, $t \in T$, is a solution of the unperturbed phase system

$$\dot{x}(t) = A_0(u(t),t)x(t) + b_0(u(t),t), \quad x(t_0) = x^0, \quad t \in T,$$

as $u(t) = v(t)$; and $p_0(t,u^0)$, $t \in T$, is a solution of the unperturbed adjoint system

$$\dot{p}(t) = -A_0^T(u(t),t)p(t) + a_0(u(t),t), \quad t \in T, \quad p(t_1) = -c_0$$

as $u(t) = u^0(t)$. The unperturbed mapping u_0^* is formed by the unperturbed Pontryagin function.

$$H_0(\psi, x, u, t) = \langle \psi, A_0(u,t)x + b_0(u,t)\rangle - \langle a_0(u,t), x\rangle - d_0(u,t)$$

and is defined by the formula

$$u_0^*(\psi, x, t) = \arg\max_{w\in U} H_0(\psi, x, w, t)', \quad \psi \in R^n, \quad x \in R^n, \quad t \in T.$$

The unperturbed Pontryagin function, phase and adjoint systems are obtained from the corresponding perturbed ones as $\varepsilon = 0$.

So, the unperturbed improvement condition (61) corresponds to improvement condition in the unperturbed optimal control problem that has the following form

$$\begin{aligned} &\Phi_0(u) = \langle c_0, x(t_1)\rangle + \\ &+ \int_T (\langle a_0(u(t),t), x(t)\rangle + d_0(u(t),t))dt \to \min_{u\in V}, \end{aligned} \tag{62}$$

$$\dot{x}(t) = A_0(u(t),t)x(t) + b_0(u(t),t),\ x(t_0) = x^0,\ t \in T = [t_0, t_1]. \tag{63}$$

Obviously, the unperturbed problem (62) and (63) is obtained from the perturbed optimal control problem (59) and (60) as $\varepsilon = 0$.

Complexity of solving the unperturbed relation (61) is defined by Cauchy problem for the unperturbed adjoint system and by Cauchy problem for the phase system

$$\begin{aligned} &\dot{x}(t) = A_0(u_0^*(\bar{p}_0(t), x(t), t), t)x(t) + b_0(u_0^*(\bar{p}_0(t), x(t), t), t), \\ &x(t_0) = x^0, \end{aligned} \tag{64}$$

where $\bar{p}_0(t) = p(t, u^0)$, $t \in T$.

Let $\bar{x}_0(t)$, $t \in T$, be a solution of the Cauchy problem (64) (probably, not unique), moreover, the output control $\bar{v}_0(t) = u_0^*(\bar{p}_0(t), \bar{x}_0(t), t)$, $t \in T$, is admissible. Then $\bar{x}_0(t) = x_0(t, \bar{v}_0)$, $t \in T$, and, therefore, $\bar{v}_0(t)$, $t \in T$, is a solution of the unperturbed condition (61).

Iterative solution process for the perturbed relation (58) with a perturbation parameter $\varepsilon \in [0, 1]$ consists in solving the problem as difficult as the unperturbed condition (61) on each iteration and has the form

$$v^{k+1}(t) = u_\varepsilon^*(p_\varepsilon(t, u^0, v^k), x_\varepsilon(t, v^{k+1}), t),\ t \in T,\ k \geq 0. \tag{65}$$

The initial approximation $v^0 \in V$ is given.

In the context of the phase and adjoint system of variables the iterative process (65) corresponds to the process

$$\begin{aligned} \dot{x}_\varepsilon^{k+1}(t) &= A_0(u_\varepsilon^*(p_\varepsilon^k(t), x_\varepsilon^{k+1}(t), t), t)x_\varepsilon^{k+1}(t) + b_0(u_\varepsilon^*(p_\varepsilon^k(t), x_\varepsilon^{k+1}(t), t), t) + \\ &+ \varepsilon f_1(x_\varepsilon^{k+1}(t), u_\varepsilon^*(p_\varepsilon^k(t), x_\varepsilon^{k+1}(t), t), t), \quad x_\varepsilon^{k+1}(t_0) = x^0, \\ \dot{p}_\varepsilon^k(t) &= -A_0^T(u^0(t), t)p_\varepsilon^k(t) + a_0(u^0(t), t) - \varepsilon(f_{1x}^T(x(t, u^0), u^0(t), t)p_\varepsilon^k(t) - \\ &- F_{1x}(x(t, u^0), u^0(t), t) + \frac{1}{2}[f_{1x}^T(x(t, u^0), u^0(t), t)p_\varepsilon^k(t) - \end{aligned}$$

$$-F_{1x}(x(t,u^0),u^0(t),t)]_x(x_\varepsilon^k(t)-x(t,u^0)),\quad t\in T,$$

$$p_\varepsilon^k(t_1)=-c_0-\varepsilon(\varphi_{1x}(x(t_1,u^0))+\frac{1}{2}\varphi_{1xx}(x(t_1,u^0))(x_\varepsilon^k(t_1)-x(t_1,u^0))),$$

where $x_\varepsilon^k(t)=x_\varepsilon(t,v^k)$, $p_\varepsilon^k(t)=p_\varepsilon(t,u^0,v^k)$, $t\in T$. Transition $k\to k+1$ is realized by successive solving of two Cauchy problems and formation of the control $v^{k+1}(t)=u_\varepsilon^*(p_\varepsilon^k(t),x_\varepsilon^{k+1}(t),t)$, $t\in T$.

In case of nonuniqueness for recurrent v^{k+1}, it is possible to choose a solution that provides the greatest decrease in target functional of the input problem. In practice, the iteration process is performed till first improvement of the input control u^0 occurs. For the initial approximation $v^0\in V$ it is possible to choose solution $\bar{v}_0$ of the unperturbed problem (61).

Under certain assumptions it is possible to justify the iterative process convergence (65).

Let us represent the iterative process in the form

$$v^{k+1}=G_\varepsilon(v^k),\quad k\geq 0, \tag{66}$$

where operator G_ε is a superposition of three operators.

The first operator P_ε is defined on the basis of the perturbed adjoint system using the relation

$$P_\varepsilon(v)=p_\varepsilon,\quad v\in V,\quad p_\varepsilon(t)=p_\varepsilon(t,u^0,v),\quad t\in T.$$

The second operator X_ε^* is defined by solution $x_\varepsilon^*(t,p)$, $t\in T$, of the perturbed Cauchy problem

$$\begin{aligned}\dot{x}(t)=A_0(u_\varepsilon^*(p(t),x(t),t),t)x(t)+b_0(u_\varepsilon^*(p(t),x(t),t),t)+\\ +\varepsilon f_1(x(t),u_\varepsilon^*(p(t),x(t),t),t),\quad x(t_0)=x^0,\quad t\in T=[t_0,t_1]\end{aligned} \tag{67}$$

on the basis of the relation

$$X_\varepsilon^*(p)=x_\varepsilon^*,\quad p\in C(T),\quad x_\varepsilon^*(t)=x_\varepsilon^*(t,p),\quad t\in T.$$

The third operator V_ε^* has the following form:

$$V_\varepsilon^*(p,x)=v_\varepsilon^*,\quad p\in C(T),\quad x\in C(T),\quad v_\varepsilon^*(t)=u_\varepsilon^*(p(t),x(t),t),\quad t\in T.$$

Finally, G_ε is represented in the form of composition

$$G_\varepsilon(v)=V_\varepsilon^*(P_\varepsilon(v),X_\varepsilon^*(P_\varepsilon(v))).$$

Let us introduce an operator on the basis of the solution $x_\varepsilon(t,v)$, $t\in T$ for the perturbed phase system (60)

$$X_\varepsilon(v)=x_\varepsilon,\quad v\in V,\quad x_\varepsilon(t)=x_\varepsilon(t,v),\quad t\in T.$$

Then the mapping X_ε^* satisfies the relation

$$X_\varepsilon^*(p) = X_\varepsilon(V_\varepsilon^*(p, X_\varepsilon^*(p))).$$

Hence, we obtain the condition $X_\varepsilon^*(P_\varepsilon(v^k)) = X_\varepsilon(v^{k+1})$, i.e. the iterative process (65) can be represented in the implicit form

$$v^{k+1} = V_\varepsilon^*(P_\varepsilon(v^k), X_\varepsilon(v^{k+1})), \quad k \geq 0.$$

Consider the operator equation

$$v = G_\varepsilon(v), \quad v \in V. \tag{68}$$

Let $v \in V$ be a solution of the perturbed problem (58). Then $x_\varepsilon(t, v)$, $t \in T$ satisfies the problem (67) as $p(t) = p_\varepsilon(t, u^0, v)$, $t \in T$. Therefore, $x_\varepsilon(t, v) = x_\varepsilon^*(t, p_\varepsilon)$, $t \in T$, where $p_\varepsilon(t) = p_\varepsilon(t, u^0, v)$, $t \in T$. So v is a solution of the problem (68). In contrary, if $v \in V$ is a solution (68) then $x_\varepsilon^*(t, p_\varepsilon) = x_\varepsilon(t, v)$, $t \in T$, i.e. v satisfies the condition (58).

So, the perturbed problem (58) with parameter $\varepsilon \in [0, 1]$ is equivalent to the perturbed operator equation (68).

The input problem (55) is written by using the operator G_ε at parameter's value $\varepsilon = 1$ and has the form

$$v = G_1(v), \quad v \in V.$$

The following unperturbed operator equation

$$v = G_0(v), \quad v \in V,$$

corresponds to the unperturbed problem (61). This unperturbed operator equation is obtained from the perturbed equation (68) as $\varepsilon = 0$. Operator G_0 is defined by using the corresponding operators P_0, X_0^*, V_0^*, X_0. In this case $P_0(v) = \bar{p}_0 \in C(T)$, $v \in V$, $X_0^*(p_0) = \bar{x}_0 \in C(T)$, $V_0^*(\bar{p}_0, \bar{x}_0) = \bar{v}_0 \in V$ is a solution of (61).

The iterative process (66) has a form of standard simple iteration method for solving the operator Equation (68). Conditions of convergence of simple iteration method can be defined on the basis of the known principle of contraction mappings. Let us formulate an analog of the known theorem [10].

Consider the operator $G : V \to V$, acting on the set V in completed normalized space of functions, that are defined on the set T with values in the compact set $U \subset R^m$, with the norm $\|\cdot\|_V$.

For solving the operator equation

$$v = G(v), \quad v \in V \tag{69}$$

the simple iteration method is considered

$$v^{k+1} = G(v^k), \quad k \geq 0. \tag{70}$$

Theorem 5. *Let the operator G satisfies the Lipschitz condition in the ball $B(v_0, l) = \{v \in V : \|v - v_0\|_V \le l, v_0 \in V, l > 0\}$ with a constant $0 < M = M(v_0, l) < 1$*

$$\|G(v) - G(u)\|_V \le M \|v - u\|_V, \ v \in B(v_0, l), \ u \in B(v_0, l), \tag{71}$$

moreover, the following condition is fulfilled

$$\|G(v_0) - v_0\|_V \le (1 - M)l. \tag{72}$$

Then the Equation (69) *has a unique solution $\bar{v} \in B(v_0, l)$ and the simple iteration method* (70) *converges to $\bar{v}$ in the norm $\|\cdot\|_V$ at any initial approximation $v^0 \in B(v_0, l)$. The following estimate is correct for method error:*

$$\left\|v^k - \bar{v}\right\|_V \le M^k \left\|v^0 - \bar{v}\right\|_V, \quad k \ge 0.$$

The theorem proof is similar to the proof illustrated in the work [10].

Note that the condition (72) is introduced in order that the iterative process approximations (70) fall outside the limits of the set $B(v_0, l)$, where the Lipschitz condition (71) is fulfilled.

Let us use Theorem 5 to justify convergence of the iterative process (66) on the set

$$V = \{u \in L_\infty(T) : u(t) \in U, t \in T\}$$

of essentially restricted functions measurable on T with values in the convex compact set $U \subset R^m$ with the norm $\|\cdot\|_\infty$.

The necessity of imbedding of piecewise continuous admissible controls on T into space $L_\infty(T)$ is connected with using properties of convergence of fundamental element sequence in complete normalized space.

Assume that the family of phase trajectories of perturbed system (60) at sufficiently small $\varepsilon > 0$ is bounded

$$x_\varepsilon(t, u) \in X, \quad t \in T, \ u \in V, \tag{73}$$

where $X \subset R^n$ is a convex compact set, and the function $f_\varepsilon(x, u, t) = A_0(u, t)x + b_0(u, t) + \varepsilon f_1(x, u, t)$ satisfies the Lipschitz condition with a constant $M > 0$

$$\|f_\varepsilon(x, u, t) - f_\varepsilon(x, v, t)\| \le M \|u - v\|, \quad u, \ v \in U, \quad x \in X, \quad t \in T.$$

Note that the sufficient boundedness condition (73) is fulfillment of the known estimate [1, 8, 11]

$$\|f_\varepsilon(x, u, t)\| \le C(\|x\| + 1), \ x \in R^n, \ u \in U, \ t \in T. \tag{74}$$

Since the function $f_\varepsilon(x, u, t)$ is quadratic with respect to x, the Lipschitz condition is fulfilled

$$\|f_\varepsilon(x,u,t) - f_\varepsilon(y,u,t)\| \le M_1 \|x-y\|, \ x \in X, \ y \in X, \ u \in U, \quad t \in T.$$

where $M_1 = \text{const} > 0$.

Hence, using the Gronwall–Bellman lemma [8, 9], it is possible to show that the operator $X_\varepsilon : u \to x_\varepsilon(t,u)$, $t \in T$, satisfies the Lipschitz condition

$$\|X_\varepsilon(u) - X_\varepsilon(v)\|_C \le M_2 \|u-v\|_\infty, \ u \in V, \ v \in V, \tag{75}$$

where $M_2 = \text{const} > 0$.

Note that by virtue of linearity of the adjoint system the boundedness condition for the family of adjoint trajectories at small $\varepsilon > 0$ is fulfilled on the basis of the sufficient condition (74)

$$p_\varepsilon(t,u^0,u) \in P, \quad t \in T, \quad u \in V,$$

where $P \subset R^n$ is a convex compact set.

The difference $q_\varepsilon(t,u^0,v,u) = p_\varepsilon(t,u^0,v) - p_\varepsilon(t,u^0,u)$, $t \in T$, satisfies the linear problem

$$\begin{aligned}
\dot q(t) = &-A_0^T(u^0(t),t)q(t) - \varepsilon(f_{1x}^T(x(t,u^0),u^0(t),t)q(t) + \\
&+\frac{1}{2}([f_{1x}^T(x(t,u^0),u^0(t),t)p_\varepsilon(t,u^0,v)]_x x_\varepsilon(t,v) - \\
&-[f_{1x}^T(x(t,u^0),u^0(t),t)p_\varepsilon(t,u^0,u)]_x x_\varepsilon(t,u) - \\
&-[f_{1x}^T(x(t,u^0),u^0(t),t)q(t)]_x x(t,u^0) - \\
&-F_{1xx}(x(t,u^0),u^0(t),t)(x_\varepsilon(t,v) - x_\varepsilon(t,u)))), \\
q(t_1) = &-\varepsilon\frac{1}{2}\varphi_{1xx}(x(t_1,u^0))(x_\varepsilon(t_1,v) - x_\varepsilon(t_1,u)), \quad t \in T.
\end{aligned}$$

Hence, taking into account lemma 1 and the Lipschitz condition (75), it is easy to obtain the following estimate for the function $q_\varepsilon(t,u^0,v,u)$, $t \in T$:

$$\|q_\varepsilon\|_C \le \varepsilon C_1 \|v-u\|_\infty, \quad v \in V, \quad u \in V,$$

where $C_1 = \text{const} > 0$. So, the operator P_ε satisfies the Lipschitz condition with a constant of order $\varepsilon > 0$

$$\|P_\varepsilon(v) - P_\varepsilon(u)\|_C \le \varepsilon C_1 \|v-u\|_\infty, \ v \in V, \ u \in V. \tag{76}$$

Note that as $\varepsilon = 0$ we have $P_0(v) = \bar p_0$, $v \in V$, and therefore, the Lipschitz condition (76) is correct as $\varepsilon = 0$.

Assume that operators X_ε^* and V_ε^* at sufficiently small $\varepsilon \ge 0$ satisfy the Lipschitz condition with respect to variables $p \in C(T)$, $x \in C(T)$ in the corresponding balls $B_1(\bar p_0, l_1)$ and $B_2(\bar p_0, \bar x_0, l_2)$ of radii $l_1 > 0$ and $l_2 > 0$ centered at points $\bar p_0$ and $(\bar p_0, \bar x_0)$, where $\bar p_0$ and $\bar x_0 = X_0^*(\bar p_0)$ are corresponding solutions of unperturbed phase and adjoint systems,

$$\|X_\varepsilon^*(p) - X_\varepsilon^*(q)\|_C \le C_2 \|p-q\|_C, \quad p \in B_1(\bar{p}_0, l_1), \quad q \in B_1(\bar{p}_0, l_1),$$

$$\|V_\varepsilon^*(p,x) - V_\varepsilon^*(q,y)\|_2 \le C_3(\|p-q\|_C + \|x-y\|_C),$$
$$(p,x) \in B_2(\bar{p}_0, \bar{x}_0, l_2), \quad (p,x) \in B_2(\bar{p}_0, \bar{x}_0, l_2),$$

where $C_2 = C_2(\bar{p}_0, l_1) > 0$, $C_3 = C_3(\bar{p}_0, \bar{x}_0, l_2) > 0$ do not depend on ε.

Then the operator G_ε at small $\varepsilon \ge 0$ satisfies the Lipschitz condition in certain ball $B(\bar{v}_0, l)$ of radius $l > 0$ with the Lipschitz constant of order ε

$$\begin{aligned} &\|G_\varepsilon(v) - G_\varepsilon(u)\|_\infty \le \varepsilon C \|v-u\|_\infty, \\ &v \in B(\bar{v}_0, l), \ u \in B(\bar{v}_0, l), \end{aligned} \tag{77}$$

where $\bar{v}_0 = V_0^*(\bar{p}_0, \bar{x}_0)$ is a solution of the unperturbed problem (61), $C = C(\bar{v}_0, l) > 0$.

Note that there is single-valuedness of mappings X_ε^*, V_ε^*, and G_ε in accepted assumptions at sufficiently small $\varepsilon \ge 0$ in view of fulfillment of Lipschitz conditions. From single-valuedness of mappings X_0^* and V_0^* it follows that solutions $\bar{p}_0$ and $\bar{v}_0$ are unique.

Assume that there is a continuity of the operator G_ε with respect to parameter ε at small $\varepsilon \ge 0$ in the ball $B(\bar{v}_0, l)$. Then the operator G_ε is close to G_0 at small $\varepsilon > 0$ and, therefore, the condition (72) of Theorem 5 is realized for the operator G_ε at sufficiently small $\varepsilon > 0$ in the ball $B(\bar{v}_0, l)$.

As a result, by Theorem 5 and taking into account the estimate (77), the iterative process (66) at small $\varepsilon > 0$ converges in the norm $\|\cdot\|_\infty$ to unique solution $\bar{v} \in B(\bar{v}_0, l)$ of the perturbed problem (68) for any initial approximation $v^0 \in B(\bar{v}_0, l)$.

Formulated conditions can be useful for analysis of proof scheme for iterative process convergence, but usually it is difficult to verify them in practice.

Let us apply the perturbation method for realizing the projection improvement condition in the problem linear in control and quadratic in state (6) and (7) with the convex compact set U.

The boundary-value improvement problem for control $u^0 \in V$ on the basis of projection operation with a given parameter $\alpha > 0$ (30) and (31) is equivalent to projective condition on the set of admissible controls

$$v(t) = u^\alpha(p(t, u^0, v), x(t, v), t), \quad t \in T. \tag{78}$$

Represent the problem (6) and (7) in the form (56) and (57), where the corresponding functions $a_0(u,t)$, $d_0(u,t)$, $F_1(x,u,t)$, $A_0(u,t)$, $b_0(u,t)$, $f_1(x,u,t)$ are linear with respect to control. Let us form the perturbed optimal control problem (59) and (60) with perturbation parameter $\varepsilon \in [0,1]$ and the unperturbed optimal control problem (62) and (63) ($\varepsilon = 0$).

The perturbed improvement condition with parameter $\varepsilon \in [0,1]$ is defined as corresponding projective improvement condition in the perturbed problem (59) and (60) and has the following form:

$$v(t) = u_\varepsilon^\alpha(p_\varepsilon(t, u^0, v), x_\varepsilon(t, v), t), \quad t \in T. \tag{79}$$

In this case the mapping u_ε^α is introduced by using the perturbed Pontryagin function in the form

$$H_\varepsilon(\psi, x, u, t) = \langle H_{\varepsilon 1}(\psi, x, t), u\rangle + H_{\varepsilon 0}(\psi, x, t)$$

by the relation

$$u_\varepsilon^\alpha(\psi, x, t) = P_U(u^0(t) + \alpha H_{\varepsilon 1}(\psi, x, t)), \quad \psi \in R^n, \quad x \in R^n, \quad t \in T.$$

The condition (78) is obtained from (79) as $\varepsilon = 1$.

Represent the perturbed condition (79) in the operator form

$$v = G_\varepsilon^\alpha(v), \quad v \in V,$$

where the operator G_ε^α is a superposition of three operators.

The first operator $P_\varepsilon : v \to p_\varepsilon(t, u^0, v)$, $t \in T$, is introduced on the basis of the perturbed adjoint system. The second operator X_ε^α is defined by the solution $x_\varepsilon^\alpha(t, p)$, $t \in T$, for the continuous perturbed Cauchy problem

$$\begin{aligned}\dot{x}(t) = {} & A_0(u_\varepsilon^\alpha(p(t), x(t), t), t)x(t) + b_0(u_\varepsilon^\alpha(p(t), x(t), t), t) + \\ & + \varepsilon f_1(x(t), u_\varepsilon^\alpha(p(t), x(t), t), t), \quad x(t_0) = x^0, \quad t \in T = [t_0, t_1],\end{aligned}$$

using the relation

$$X_\varepsilon^\alpha(p) = x_\varepsilon^\alpha, \quad p \in C(T), \quad x_\varepsilon^\alpha(t) = x_\varepsilon^\alpha(t, p), \quad t \in T.$$

The third operator is defined by the formula

$$V_\varepsilon^\alpha(p, x) = v_\varepsilon^\alpha, \ p \in C(T), \ x \in C(T), \ v_\varepsilon^\alpha(t) = u_\varepsilon^\alpha(p(t), x(t), t), \ t \in T.$$

On the whole, G_ε^α is formed in the composition

$$G_\varepsilon^\alpha(v) = V_\varepsilon^\alpha(P_\varepsilon(v), X_\varepsilon^\alpha(P_\varepsilon(v))).$$

Using the defined operator $X_\varepsilon : v \to x_\varepsilon(t, v)$, $t \in T$, on the basis of solution for the perturbed phase system (60), the mapping X_ε^α is represented in the form

$$X_\varepsilon^\alpha(p) = X_\varepsilon(V_\varepsilon^\alpha(p, X_\varepsilon^\alpha(p))).$$

To solve the perturbed problem (79) the iterative process is considered

$$v^{k+1}(t) = u_\varepsilon^\alpha(p_\varepsilon(t, u^0, v^k), x_\varepsilon(t, v^{k+1}), t), \quad t \in T, \quad k \geq 0; \tag{80}$$

this has the operator form

$$v^{k+1} = G_\varepsilon^\alpha(v^k), \quad k \geq 0.$$

Note that, since the condition $X_\varepsilon^\alpha(P_\varepsilon(v^k)) = X_\varepsilon(v^{k+1})$ is fulfilled, the iterative process (80) can be represented in the implicit operator form

$$v^{k+1} = V_\varepsilon^\alpha(P_\varepsilon(v^k), X_\varepsilon(v^{k+1})), \quad k \geq 0.$$

Let us analyze convergence of the method (80) on the set of admissible controls

$$V = \{v \in C(T) : v(t) \in U, t \in T\}$$

using the operator representation and Theorem 5.

Assume that the boundedness condition (73) at all $\varepsilon \in [0,1]$ is fulfilled.

Hence on the basis of fulfillment of the Lipschitz condition for projection operator P_U we obtain

$$\begin{aligned}\|x_\varepsilon^\alpha(t,p)x_\varepsilon^\alpha(t,q)\| &= \|x_\varepsilon(t, V_\varepsilon^\alpha(p, X_\varepsilon^\alpha(p))) - x_\varepsilon(t, V_\varepsilon^\alpha(q, X_\varepsilon^\alpha(q)))\| \leq \\ &\leq M_3 \int_T \|V_\varepsilon^\alpha(p, X_\varepsilon^\alpha(p)) - V_\varepsilon^\alpha(q, X_\varepsilon^\alpha(q))\|\, dt \leq \\ &\leq \alpha M_4 \int_T \|H_{\varepsilon 1}(p(t), x_\varepsilon^\alpha(t,p), t) - H_{\varepsilon 1}(q(t), x_\varepsilon^\alpha(t,q), t)\|\, dt \leq \\ \leq \alpha M_0(\|p-q\|_C &+ \|x_\varepsilon^\alpha(t,p) - x_\varepsilon^\alpha(t,q)\|_C), \ t \in T, \ p \in C(T), \ q \in C(T),\end{aligned}$$

where $M_3 = \text{const} > 0$, $M_4 = \text{const} > 0$, $M_0 = \text{const} > 0$. Therefore, at sufficiently small $\alpha > 0$ the operator X_ε^α satisfies the Lipschitz condition in the space $C(T)$

$$\|X_\varepsilon^\alpha(p) - X_\varepsilon^\alpha(q)\|_C \leq \frac{\alpha M_0}{1 - \alpha M_0} \|p-q\|_C,$$

where $0 < \alpha M_0 < 1$.

For operator V_ε^α we have

$$\begin{aligned}&\|u_\varepsilon^\alpha(p(t), x(t), t) - u_\varepsilon^\alpha(q(t), y(t), t)\| = \\ &\quad = \alpha \|H_{\varepsilon 1}(p(t), x(t), t) - H_{\varepsilon 1}(q(t), y(t), t)\| \leq \\ &\quad \leq \alpha C_4(\|p-q\|_C + \|x-y\|_C), \ t \in T, \ p, x, q, y \in C(T),\end{aligned}$$

where $C_4 = \text{const} > 0$. Therefore

$$\|V_\varepsilon^\alpha(p,q) - V_\varepsilon^\alpha(q,y)\|_C \leq \alpha C_4(\|p-q\|_C + \|x-y\|_C), \quad p, x, q, y \in C(T).$$

So, the operator V_ε^α satisfies the Lipschitz condition in the variables with a constant proportional to parameter $\alpha > 0$.

In view of fulfillment of the Lipschitz condition (76) for the operator P_ε with a constant proportional to $\varepsilon > 0$, finally, we obtain the Lipschitz condition for the operator G_ε^α at all $\varepsilon \in [0,1]$ in the form

$$\|G_\varepsilon^\alpha(v) - G_\varepsilon^\alpha(u)\|_C \leq \varepsilon \frac{\alpha C_0}{1 - \alpha M_0} \|v-u\|_C, \quad v \in V, \quad u \in V,$$

where $C_0 = \text{const} > 0$.

On the whole, by Theorem 5 we obtain the following statement concerning decidability of the perturbed problem (79) and the process convergence (80).

Theorem 6. *Let the family of perturbed phase trajectories of system* (60) *in the problem* (6) *and* (7) *with the convex compact set* U *be bounded:* $x_\varepsilon(t,u) \in X$, $t \in T$, $u \in V$, $\varepsilon \in [0,1]$, *where* $X \subset R^n$ *is a convex compact set. Then for given sufficiently small projection parameter* $\alpha > 0$ *as* $0 < \varepsilon < \bar{\varepsilon} = \frac{1-\alpha M_0}{\alpha C_0}$, $C_0 = const > 0$, $M_0 = const > 0$

(1) The perturbed condition (79) *has a unique solution* $\bar{v}^\alpha \in V$;
(2) The iterative process (80) *converges in the norm* $\|\cdot\|_C$ *to the solution* $\bar{v}^\alpha$ *for any initial approximation* $v^0 \in V$.

Corollary 2. *Under conditions of Theorem* 6 *at sufficiently small* $\alpha > 0$

(1) The perturbed relation (79) *as* $\varepsilon = 1$ *has a unique solution* $\bar{v}^\alpha \in V$;
(2) The iterative process (80) *as* $\varepsilon = 1$ *with any admissible initial approximation* $v^0 \in V$ *converges in the norm* $\|\cdot\|_C$ *to the solution* $\bar{v}^\alpha$.

2.5 Projective Perturbation Method for Improvement Condition

The problem linear in control and quadratic in state (6) and (7) is considered.

For given control $u^0 \in V$ and fixed $\alpha > 0$ let us represent the improvement condition (78) in control space in the form

$$v(t) = P_U(u^0(t) + \alpha H_1(p(t,u^0,v), x(t,v), t)),\ t \in T. \tag{81}$$

Let us consider a projection parameter $\alpha > 0$ as a perturbation parameter and call the condition (81) perturbed. The unperturbed condition is obtained from the perturbed one (1.5.1) as $\alpha = 0$, and has the obvious solution $v(t) = u^0(t)$, $t \in T$.

The iterative process for solving the perturbed relation (81) has the form

$$v^{k+1}(t) = P_U(u^0(t) + \alpha H_1(p(t,u^0,v^k), x(t,v^k), t)), \quad t \in T. \tag{82}$$

The other iterative process for the problem (81) has the implicit form

$$v^{k+1}(t) = P_U(u^0(t) + \alpha H_1(p(t,u^0,v^k), x(t,v^{k+1}), t)), \quad t \in T. \tag{83}$$

An initial approximation $v^0 \in V$ is prescribed on the initial (zero) iteration.

Note that the process (83) coincides with the solving process (80) for the perturbed problem (79) as $\varepsilon = 1$, that corresponds to the input problem (81). Therefore, Corollary 2 defines the condition for process convergence (83).

Let us formulate conditions for process convergence (82) using Theorem 5 on the set $V = \{v \in C(T) : v(t) \in U, t \in T\}$. For this purpose we will describe the perturbed problem (81) with respect to parameter $\alpha > 0$ and the process (82) for solving this problem in the operator form

$$v = G^\alpha(v), \quad v \in V, \tag{84}$$

$$v^{k+1} = G^\alpha(v^k), \quad k \geq 0. \tag{85}$$

The operator G^α can be represented in the form of superposition of operators P_ε, V_ε^α, X_ε introduced in Section 2.4 as $\varepsilon = 1$

$$G^\alpha(v) = V_1^\alpha(P_1(v), X_1(v)), \quad v \in V.$$

In view of properties of the projection operator P_U the mapping G^α, $\alpha > 0$ is a single valued.

The unperturbed problem

$$v = G^0(v), \quad v \in V,$$

is defined by the operator $G^0 : v \to u^0$, $v \in V$. Therefore, u^0 is a unique solution of the unperturbed problem. In this case G^0 is obtained from G^α, if assume $\alpha = 0$.

Assume that the family of phase trajectories is bounded on the set V: $x(t, v) \in X$, $t \in T$, $v \in V$, where $X \subset R^n$ is a convex compact set.

Then similarly to (75) and (76) we obtain that operators P_1 and X_1 satisfy the Lipschitz condition with a constant $C_1 > 0$

$$\|X_1(v) - X_1(u)\|_C \le C_1 \|v - u\|_C, \quad v \in V, \quad u \in V, \tag{86}$$

$$\|P_1(v) - P_1(u)\|_C \le C_1 \|v - u\|_C, \quad v \in V, \quad u \in V. \tag{87}$$

On the basis of the Lipschitz condition for the projection operator P_U we obtain

$$\begin{aligned}
&\|u^\alpha(p(t), x(t), t) - u^\alpha(q(t), y(t), t)\| \\
&\quad = \alpha \|H_1(p(t), x(t), t) - H_1(q(t), y(t), t)\| \\
&\quad \le \alpha C_2(\|p - q\|_C + \|x - y\|_C), \quad t \in T, \quad p, x, q, y \in C(T),
\end{aligned}$$

where $C_2 = \text{const} > 0$. Therefore,

$$\begin{gathered}
\|V_1^\alpha(p, q) - V_1^\alpha(q, y)\|_C \le \alpha C_2(\|p - q\|_C + \|x - y\|_C), \\
p, x, q, y \in C(T).
\end{gathered} \tag{88}$$

So, the operator V_1^α satisfies the Lipschitz condition with a constant, proportional to parameter $\alpha > 0$. From conditions (86), (87), and (88) it follows that the operator G^α satisfies the Lipschitz condition with a constant, proportional to $\alpha > 0$

$$\|G^\alpha(v) - G^\alpha(u)\|_C \le \alpha 2 C_1 C_2 \|v - u\|_C, \quad v \in V, \quad u \in V.$$

On the whole, by Theorem 5, the iterative process (85) at small $\alpha > 0$ converges to a unique solution of the perturbed problem (84) for any initial approximation $v^0 \in V$.

So, the following convergence theorem is proved.

Theorem 7. *Let the family of phase trajectories in the problem linear with respect to control and quadratic with respect to state* (6) *and* (7) *with the convex compact set* $U \subset R^m$ *be bounded:* $x(t,u) \in X$, $t \in T$, $u \in V$, *where* $X \subset R^n$ *is a convex compact set. Then for a sufficient small projection parameter* $\alpha > 0$

(1) the problem (81) *has a unique solution* $\bar{v}^\alpha \in V$;

(2) the iterative process (82) *converges in the norm* $\|\cdot\|_C$ *to a solution* $\bar{v}^\alpha$ *for any initial approximation* $v^0 \in V$.

Note that under conditions of Theorem 7 the solution of the perturbed problem (81) for control $u^0 \in V$, satisfying the maximum principle, coincides with u^0 since its uniqueness.

For initial approximation of iterative processes (82) and (83) in solving the perturbed problem (81) for control $u^0 \in V$ that is not satisfying the maximum principle, it is possible to choose the initial approximation $v^0 = u^0$. In this case for sufficiently small $\alpha > 0$, according to Theorem 7, Corollary 2 and the improvement estimate (32), the strict improvement of control u^0 of iterative processes is guaranteed.

The perturbation method of projective improvement condition in control space without crucial variations generalizes to optimal control problems, which are polynomial with respect to state, including problems with time delay.

Note that projective perturbation method is advantageously different from perturbation methods with artificial perturbation parameter $\varepsilon \in [0,1]$. In projective perturbation method the control $u^0 \in V$ is being improved by solving the perturbed problem (81) for any projective parameter $\alpha > 0$. In common case solving perturbed problems with a parameter $0 < \varepsilon < 1$ does not guarantee improvement of the control u^0.

In conclusion let us extract the primary properties of developed perturbation methods.

1. The absence of operation of parametric search for the improving control.
2. Nonlocal character of improvement that is caused by fixity of perturbation parameter.
3. Possibility for improvement of controls, satisfying the maximum principle.

3 Perturbation Methods in the Main Optimal Control Problem

The known approach to solving optimal control problem is reduction to a double-point boundary-value problem for ordinary differential equations on the basis of necessary optimality condition with consequent solving of the obtained boundary-value problem by numerical method. In common case the difficulties in solving a boundary-value problem in state space are connected with presence of positive real parts of eigenvalues for Jacobi matrix. It is also

connected with possible discontinuity of the right-hand side of boundary-value problem with respect to phase variables.

Methods of maximum principle provide a convergence of residual of the boundary-value problem for maximum principle to zero on improving control approximations. So, these methods permit one to solve a boundary-value problem in control space. The advantages of these methods are computational stability of phase and adjoint subsystems for a boundary-value problem, and relaxation on target functional on each iteration of methods. The relaxation is provided with respect to small parameter that regulates a domain of weak or needle-shaped control variation. This parametric search is the most labor-consuming part of the iterative process. Moreover, the operation of control variation can form a calculated control which is hard to realize in practice. The small deviation of this control leads to inadmissible change of target functional in comparison with the calculated value.

In this chapter the iterative methods for calculation of extremum controls (which are satisfying the maximum principle) are considered. Considered methods do not contain an operation of parametric search for the improving control. Proposed methods are applied for solving necessary optimality conditions in the main optimal control problem. In design the methods are similar to perturbation methods, that were developed in the previous chapter in order to solve improvement conditions in polynomial optimal control problems.

3.1 The Main Optimal Control Problem

The main optimal control problem is considered

$$\Phi(u) = \varphi(x(t_1)) + \int_T F(x(t), u(t), t)dt \to \min_{u \in V}, \tag{89}$$

$$\dot{x}(t) = f(x(t), u(t), t), \quad x(t_0) = x^0,\ u(t) \in U,\ t \in T = [t_0, t_1], \tag{90}$$

where $x(t) = (x_1(t), ..., x_n(t))$ is a state vector, $u(t) = (u_1(t), ..., u_m(t))$ is a control vector. As admissible controls the set V of functions piecewise continuous on T with values in the convex compact set $U \subset R^m$ is considered. The initial state x^0 and the control interval T are given.

Introduce the following set of assumptions for the problem (89) and (90) (DMP conditions):

1. function $\varphi(x)$ is continuously differentiable on R^n, vector-valued function $F(x, u, t)$, vector function $f(x, u, t)$, and its derivatives $F_x(x, u, t)$, $F_u(x, u, t)$, $f_x(x, u, t)$, $f_u(x, u, t)$ are continuous in the arguments (x, u, t) on the set $R^n \times U \times T$;
2. function $f(x, u, t)$ satisfies the Lipschitz condition with respect to x in $R^n \times U \times T$ with a constant $L > 0$

$$\|f(x,u,t) - f(y,u,t)\| \le L\,\|x-y\|\,.$$

DMP conditions guarantee [11] existence and uniqueness of solution $x(t,v)$, $t \in T$, for the system (90) for any admissible control $v(t)$, $t \in T$.

Let us form the Pontryagin function with adjoint variable $\psi \in R^n$

$$H(\psi,x,u,t) = \langle f(x,u,t),\psi\rangle - F(x,u,t).$$

For admissible control $v \in V$ designate by $\psi(t,v),\ t \in T$, a solution of the standard adjoint system

$$\dot{\psi}(t) = -H_x(\psi(t),x(t),u(t),t), \quad t \in T, \quad \psi(t_1) = -\varphi_x(x(t_1)), \tag{91}$$

as $u(t) = v(t)$, $x(t) = x(t,v)$.

Using mapping u^*, introduced in Chapter 2, the maximum principle for control $u \in V$ is represented in the form

$$u(t) = u^*(\psi(t,u),x(t,u),t), \quad t \in T. \tag{92}$$

The boundary-value problem of maximum principle has the following form:

$$\dot{x}(t) = f(x(t),u^*(\psi(t),x(t),t),t), \quad x(t_0) = x^0, \tag{93}$$

$$\dot{\psi}(t) = -H_x(\psi(t),x(t),u^*(\psi(t),x(t),t),t),\ \psi(t_1) = -\varphi_x(x(t_1)). \tag{94}$$

The boundary-value problem (93) and (94) in state space reduces to the pointwise relation (92) on the set of admissible controls. In common case, right-hand sides of the boundary-value problem are discontinuous with respect to phase variables x, ψ.

In DMP conditions the differential maximum principle follows from the maximum principle (92)

$$\langle H_u(\psi(t,u),x(t,u),u(t),t), w-u(t)\rangle \le 0, \quad w \in U,\ t \in T. \tag{95}$$

Define mapping $w^\alpha,\ \alpha > 0$ using the relation

$$\begin{gathered} w^\alpha(\psi,x,u,t) = P_U(u + \alpha H_u(\psi,x,u,t)), \\ \psi \in R^n, \quad x \in R^n, \quad u \in U, \quad t \in T, \end{gathered} \tag{96}$$

where P_U is a projection operator to set U in Euclidean form.

On the basis of the Lipschitz condition for operator P_U function w^α is continuous in the variables $(\psi,x,u) \in R^n \times R^n \times U$ and piecewise continuous with respect to $t \in T$. In this case the inequality takes place

$$\langle H_u(\psi,x,u,t), w^\alpha(\psi,x,u,t) - u\rangle \ge \frac{1}{\alpha}\,\|w^\alpha(\psi,x,u,t) - u\|^2\,. \tag{97}$$

The estimate (97) is caused by properties of the operation of projection.

The differential maximum principle (95) for control $u \in V$ using the mapping (96) is represented in the form

$$u(t) = w^{\alpha}(\psi(t,u), x(t,u), u(t), t), \quad t \in T,\ \alpha > 0. \tag{98}$$

Note that to fulfill (95) it is sufficient to examine the condition (98) at least for one $\alpha > 0$. Conversely, from the condition (95) it follows that (98) is fulfilled for all $\alpha > 0$.

In the problem, linear with respect to control (89) and (90) (functions $f(x,u,t)$, $F(x,u,t)$ are linear with respect to u) the differential maximum principle (98) is equivalent to the maximum principle (92).

Perturbation methods are proposed to use for solving the relation of the maximum principle (92) and sufficient optimality condition (98).

3.2 Perturbation Method for Maximum Principle

Introduce a perturbation parameter $\varepsilon \in [0,1]$ into the condition of maximum principle (92), considered in the form

$$v(t) = u^{*}(\psi(t,v), x(t,v), t), \quad t \in T. \tag{99}$$

To do this let us isolate from the problem (89) and (90) a special part, linear in state, with variables separated in state and control. Represent this isolated part in the form

$$\Phi(u) = \langle c_0, x(t_1)\rangle + \varphi_1(x(t_1)) + \\ + \int_T (\langle a_0(t), x(t)\rangle + d_0(u(t),t) + F_1(x(t),u(t),t))dt \to \min_{u\in V}, \tag{100}$$

$$\dot{x}(t) = A_0(t)x(t) + b_0(u(t),t) + f_1(x(t),u(t),t), \quad x(t_0) = x^0, \\ u(t) \in U, \quad t \in T = [t_0, t_1], \tag{101}$$

where matrix function $A_0(t)$ and vector function $a_0(t)$ are continuous on T, vector function $b_0(u,t)$ and function $d_0(u,t)$ are continuous in the variables u, t on the set $U \times T$, c_0 is a constant vector.

Note that matrix function $A_0(t)$ and vector function $a_0(t)$ do not depend on control, contrary to presentation of the polynomial problem in the form (56) and (57)

Introduce a perturbed optimal control problem with a perturbation parameter $\varepsilon \in [0,1]$

$$\Phi_{\varepsilon}(u) = \langle c_0, x(t_1)\rangle + \varepsilon\varphi_1(x(t_1)) + \\ + \int_T (\langle a_0(t), x(t)\rangle + d_0(u(t),t) + \varepsilon F_1(x(t),u(t),t))dt \to \min_{u\in V}, \tag{102}$$

$$\dot{x}(t) = A_0(t)x(t) + b_0(u(t),t) + \varepsilon f_1(x(t),u(t),t), \quad x(t_0) = x^0, \\ u(t) \in U, \quad t \in T = [t_0, t_1]. \tag{103}$$

The problem (102) and (103) is matched by the perturbed Pontryagin function

$$H_\varepsilon(\psi, x, u, t) = \langle \psi, A_0(t)x + b_0(u,t)\rangle - \langle a_0(t), x\rangle - d_0(u,t) +$$

$$+\varepsilon(\langle \psi, f_1(x,u,t)\rangle - F_1(x,u,t)),$$

the perturbed mapping

$$u_\varepsilon^*(\psi, x, t) = \arg\max_{w\in U} H_\varepsilon(\psi, x, w, t)', \quad \psi \in R^n, \quad x \in R^n, \quad t \in T,$$

and the perturbed adjoint system

$$\dot\psi(t) = -A_0^T(t)\psi(t) + a_0(t) - \varepsilon(f_{1x}^T(x(t), u(t), t)\psi(t) - F_{1x}(x(t), u(t), t)),$$

$$\psi(t_1) = -c_0 - \varepsilon\varphi_{1x}(x(t_1)), \quad t \in T. \tag{104}$$

Designate by $x_\varepsilon(t, v)$, $t \in T$, a solution of the perturbed phase system (103) as $u(t) = v(t)$; by $\psi_\varepsilon(t, v)$, $t \in T$, a solution of the perturbed adjoint system (104) as $u(t) = v(t)$, $x(t) = x_\varepsilon(t, v)$.

We will determine the maximum principle condition for the perturbed problem (102) and (103)

$$v(t) = u_\varepsilon^*(\psi_\varepsilon(t, v), x_\varepsilon(t, v), t), \quad t \in T, \tag{105}$$

as a perturbed condition of maximum principle with a parameter $\varepsilon \in [0, 1]$.

The input problem in the form (100) and (101), Pontryagin function H, mapping u^*, the adjoint system (91), and the maximum principle condition (99) are obtained, respectively, from the perturbed problem (102) and (103), perturbed Pontryagin function H_ε, perturbed mapping u_ε^*, the perturbed adjoint system (104), and the perturbed condition (105) as $\varepsilon = 1$.

The unperturbed condition of maximum principle corresponds to the unperturbed optimal control problem

$$\Phi_0(u) = \langle c_0, x(t_1)\rangle + \int_T (\langle a_0(t), x(t)\rangle + d_0(u(t), t))dt \to \min_{u\in V}, \tag{106}$$

$$\dot x(t) = A_0(t)x(t) + b_0(u(t), t), \quad x(t_0) = x^0, \quad t \in T = [t_0, t_1], \tag{107}$$

with the unperturbed Pontryagin function

$$H_0(\psi, x, u, t) = \langle \psi, A_0(t)x + b_0(u,t)\rangle - \langle a_0(t), x\rangle - d_0(u,t),$$

with the unperturbed mapping

$$u_0^*(\psi, x, t) = \arg\max_{w\in U} H_0(\psi, x, w, t), \quad \psi \in R^n, \quad x \in R^n, \quad t \in T,$$

with the unperturbed adjoint system

$$\dot{\psi}(t) = -A_0^T(t)\psi(t) + a_0(t), \quad t \in T, \quad p(t_1) = -c_0. \tag{108}$$

For $v \in V$ designate by $x_0(t, v)$, $t \in T$, a solution of the unperturbed phase system (107); by $\bar{\psi}_0(t)$, $t \in T$, a solution of the unperturbed adjoint system (108). The unperturbed maximum principle condition is obtained from (105) as $\varepsilon = 0$ and has the form

$$v(t) = u_0^*(\bar{\psi}_0(t), x_0(t, v), t), \quad t \in T. \tag{109}$$

Unperturbed phase and adjoint systems, Pontryagin function H_0, mapping u_0^* are obtained from corresponding perturbed ones as $\varepsilon = 0$.

Note that the unperturbed problem (106) and (107) is linearly convex, for this the maximum principle (109) is necessary and sufficient condition of control optimality [8, 9].

Complexity of solving the unperturbed relation (109) is defined by solving the Cauchy problem for the adjoint system (108) and solving the Cauchy problem for the phase system

$$\dot{x}(t) = A_0(t)x(t) + b_0(u_0^*(\bar{\psi}_0(t), x(t), t), t), x(t_0) = x^0, \quad t \in T. \tag{110}$$

Let $\bar{x}_0(t)$, $t \in T$, be a solution of the problem (110), and also the output control $\bar{v}_0(t) = u_0^*(\bar{\psi}_0(t), \bar{x}_0(t), t)$, $t \in T$, is admissible. Then $\bar{x}_0(t) = x_0(t, \bar{v}_0)$, $t \in T$, and, therefore, $\bar{v}_0(t)$, $t \in T$, is a solution of the unperturbed system (109).

The iterative process of solving the perturbed condition (105) with fixed perturbation parameter $\varepsilon \in (0, 1]$ consists in solving problem, similar to the unperturbed condition (109) on each iteration, and has the form

$$v^{k+1}(t) = u_\varepsilon^*(\psi_\varepsilon(t, v^k), x_\varepsilon(t, v^{k+1}), t), \ t \in T, \ k \geq 0. \tag{111}$$

The initial approximation $v^0 \in V$ is given.

Function $\psi_\varepsilon(t, v^k)$, $t \in T$, is a solution of the adjoint Cauchy problem

$$\dot{\psi}(t) = -A_0^T(t)\psi(t) + a_0(t) - \varepsilon(f_{1x}^T(x_\varepsilon(t, v^k), v^k(t), t)\psi(t) - \\ -F_{1x}(x_\varepsilon(t, v^k), v^k(t), t)), \quad t \in T, \quad \psi(t_1) = -c_0 - \varepsilon\varphi_{1x}(x_\varepsilon(t_1, v^k)).$$

Let $x_\varepsilon(t)$, $t \in T$, be a solution (probably, not unique) of the phase Cauchy problem

$$\dot{x}(t) = A_0(t)x(t) + b_0(u_\varepsilon^*(\psi_\varepsilon(t, v^k), x(t), t), t) + \\ +\varepsilon f_1(x(t), u_\varepsilon^*(\psi_\varepsilon(t, v^k), x(t), t), t), \quad x(t_0) = x^0.$$

Let us form the output control $v^{k+1}(t) = u_\varepsilon^*(\psi_\varepsilon(t, v^k), x_\varepsilon(t), t)$, $t \in T$. It is clear that $x_\varepsilon(t) = x_\varepsilon(t, v^{k+1})$, $t \in T$, and, therefore $v^{k+1}(t)$, $t \in T$, satisfies the process (111).

In case of nonuniqueness, a solution of the system (111), providing the least value of residual for the maximum principle, can be chosen as a recurrent

approximation v^{k+1}. Achievement of given small value of residual for the maximum principle can be a practical criterion for stopping of the iterative process (111).

On initial (zero) iteration the unperturbed solution $\bar{v}_0$ can be chosen as initial approximation $v^0 \in V$ of the process (111).

Under certain assumptions it is possible to justify convergence of the iterative process (111).

Represent the perturbed condition (105) with a parameter $\varepsilon \in [0,1]$ in the operator form

$$v = G_\varepsilon(v), \quad v \in V, \tag{112}$$

where operator G_ε is a superposition of three operators.

The first operator Ψ_ε is defined on the basis of solution $\psi_\varepsilon(t,v)$, $t \in T$, of the perturbed adjoint system (104) using the relation

$$\Psi_\varepsilon(v) = \psi_\varepsilon, \quad v \in V, \quad \psi_\varepsilon(t) = \psi_\varepsilon(t,v), \quad t \in T.$$

The second operator X^*_ε is defined by solution $x^*_\varepsilon(t,p)$, $t \in T$, of the discontinuous perturbed Cauchy problem

$$\dot{x}(t) = A_0(t)x(t) + b_0(u^*_\varepsilon(p(t),x(t),t),t) + \\ +\varepsilon f_1(x(t),u^*_\varepsilon(p(t),x(t),t),t), \quad x(t_0) = x^0, \quad t \in T = [t_0,t_1],$$

on the basis of the relation

$$X^*_\varepsilon(p) = x^*_\varepsilon, \quad p \in C(T), \quad x^*_\varepsilon(t) = x^*_\varepsilon(t,p), \quad t \in T,$$

where $C(T)$ is a space of functions continuous on T.

The third operator V^*_ε is defined by the formula

$$V^*_\varepsilon(p,x) = v^*_\varepsilon, \quad p,x \in C(T), \quad v^*_\varepsilon(t) = u^*_\varepsilon(p(t),x(t),t), \quad t \in T.$$

As a result, G_ε is represented in the form of composition

$$G_\varepsilon(v) = V^*_\varepsilon(\Psi_\varepsilon(v), X^*_\varepsilon(\Psi_\varepsilon(v))).$$

Introduce operator X_ε on the basis of solution $x_\varepsilon(t,v)$, $t \in T$, for the perturbed phase system (103)

$$X_\varepsilon(v) = x_\varepsilon, \quad v \in V, \quad x_\varepsilon(t) = x_\varepsilon(t,v), \quad t \in T.$$

Then mapping X^*_ε can be represented as the relation

$$X^*_\varepsilon(p) = X_\varepsilon(V^*_\varepsilon(p, X^*_\varepsilon(p))).$$

The input condition (99) is written using operator G_ε as $\varepsilon = 1$ and has the form

$$v = G_1(v), \quad v \in V.$$

The unperturbed condition (109) takes the form

$$v = G_0(v), \quad v \in V$$

and is obtained from the perturbed one (112) as $\varepsilon = 0$. Here operator G_0 is defined on the basis of corresponding operators Ψ_0, X_0^*, V_0^*, X_0. We have $\Psi_0(v) = \bar{\psi}_0 \in C(T)$, $v \in V$, $X_0^*(\bar{\psi}_0) = \bar{x}_0 \in C(T)$, $V_0^*(\bar{\psi}_0, \bar{x}_0) = \bar{v}_0 \in V$.

The iterative process (111) using operator G_ε is written in the explicit form

$$v^{k+1} = G_\varepsilon(v^k), \quad k \geq 0. \tag{113}$$

Note that the condition $X_\varepsilon^*(\Psi_\varepsilon(v^k)) = X_\varepsilon(v^{k+1})$ is fulfilled, i.e., the iterative process (111) is represented in the implicit operator form

$$v^{k+1} = V_\varepsilon^*(\Psi_\varepsilon(v^k), X_\varepsilon(v^{k+1})), \quad k \geq 0.$$

Convergence of the process (113) can be justified using Theorem 5 on the set of admissible controls $V = \{u \in L_\infty(T) : u(t) \in U, t \in T\}$.

Let us formulate the conditions such that operator G_ε, at a sufficiently small $\varepsilon > 0$, satisfies the conditions of the stated theorem.

Assume that a family of phase trajectories for the perturbed system (103) is bounded at a sufficiently small $\varepsilon > 0$

$$x_\varepsilon(t, u) \in X, \quad t \in T, \quad u \in V, \tag{114}$$

where $X \subset R^n$ is a convex compact set. In this case, taking into consideration DMP conditions and linearity of the adjoint system on the basis of sufficient condition (74), we obtain the boundedness condition for the family of adjoint trajectories for the system (104) at small $\varepsilon > 0$

$$\psi_\varepsilon(t, u) \in P, \quad t \in T, \quad u \in V, \tag{115}$$

where $P \subset R^n$ is a convex compact set.

Taking into account fulfillment of the Lipschitz condition with respect to $x \in X$ for the function

$$f_\varepsilon(x, u, t) = A_0(t)x + b_0(u, t) + \varepsilon f_1(x, u, t), \quad \varepsilon \in [0, 1]$$

and using the Gronwall–Bellman lemma [8, 9], it is possible to show that operator $X_\varepsilon : u \to x_\varepsilon(t, u)$, $t \in T$, satisfies the Lipschitz condition at small $\varepsilon > 0$

$$\|X_\varepsilon(u) - X_\varepsilon(v)\|_C \leq M_2 \|u - v\|_\infty, \quad u \in V, \quad v \in V, \tag{116}$$

where $M_2 = \text{const} > 0$.

Difference $q_\varepsilon(t, v, u) = \psi_\varepsilon(t, v) - \psi_\varepsilon(t, u)$, $t \in T$, satisfies the linear problem

$$\begin{aligned} \dot{q}(t) = &-A_0^T(t)q(t) - \varepsilon(f_{1x}^T(x_\varepsilon(t, v), v(t), t)\psi_\varepsilon(t, v) - \\ &-f_{1x}^T(x_\varepsilon(t, u), u(t), t)\psi_\varepsilon(t, u) - \\ &-F_{1x}(x_\varepsilon(t, v), v(t), t) + F_{1x}(x_\varepsilon(t, u), u(t), t)), \\ q(t_1) = &-\varepsilon(\varphi_{1x}(x_\varepsilon(t_1, v)) - \varphi_{1x}(x_\varepsilon(t_1, u))), \quad t \in T. \end{aligned}$$

In addition to DMP conditions, assume that functions $f(x,u,t)$, $F(x,u,t)$, $\varphi(x)$ are twice continuously differentiable in the variables x, u, t on the set $R^n \times U \times T$. Then under fulfillment of the boundedness condition (114), functions $f(x,u,t)$, $F(x,u,t)$, $\varphi(x)$, and their derivatives with respect to x, u satisfy the Lipschitz condition in the variables $x \in X$, $u \in U$ with one Lipschitz constant $M_1 > 0$.

Hence, taking into account Lemma 1, the boundedness condition (115) and the Lipschitz condition (116), for function $q_\varepsilon(t,v,u)$, $t \in T$, we obtain the estimate at small $\varepsilon > 0$:

$$\|q_\varepsilon\|_C \le \varepsilon C_1 \|v-u\|_\infty, \quad v \in V, \quad u \in V,$$

where $C_1 = \text{const} > 0$. So, operator Ψ_ε satisfies the Lipschitz condition with a constant of order $\varepsilon > 0$

$$\|\Psi_\varepsilon(v) - \Psi_\varepsilon(u)\|_C \le \varepsilon C_1 \|v-u\|_\infty, \quad v \in V, \quad u \in V. \tag{117}$$

As $\varepsilon = 0$, we have $\Psi_0(v) = \bar\psi_0$, $v \in V$, and therefore, the Lipschitz condition (117) is also fulfilled as $\varepsilon = 0$.

Assume that operators X^*_ε, V^*_ε, at sufficiently small $\varepsilon \ge 0$, satisfy the Lipschitz condition with respect to variables $p \in C_2(T)$, $x \in C_2(T)$ in corresponding balls $B_1(\bar\psi_0, l_1)$ and $B_2(\bar\psi_0, \bar x_0, l_2)$ of radii $l_1 > 0$ and $l_2 > 0$, centered at points $\bar\psi_0$ and $(\bar\psi_0, \bar x_0)$, where $\bar\psi_0$, $\bar x_0 = X^*_0(\bar\psi_0)$ are corresponding solutions of the unperturbed adjoint system (108) and the phase system (107),

$$\|X^*_\varepsilon(p) - X^*_\varepsilon(q)\|_C \le C_2 \|p-q\|_C,$$

$$\|V^*_\varepsilon(p,x) - V^*_\varepsilon(q,y)\|_2 \le C_3(\|p-q\|_C + \|x-y\|_C),$$

where $C_2 = C_2(\bar\psi_0, l_1) > 0$, $C_3 = C_3(\bar\psi_0, \bar x_0, l_2) > 0$ do not depend on ε.

Then at small $\varepsilon \ge 0$, operator G_ε satisfies the Lipschitz condition in a certain ball $B(\bar v_0, l)$ of radius $l > 0$ with the Lipschitz constant of order ε

$$\begin{aligned} &\|G_\varepsilon(v) - G_\varepsilon(u)\|_\infty \le \varepsilon C_0 \|v-u\|_\infty, \\ &v \in B(\bar v_0, l), \ u \in B(\bar v_0, l), \end{aligned} \tag{118}$$

where $\bar v_0 = V^*_0(\bar\psi_0, \bar x_0)$ is a solution of the unperturbed problem (109), $C_0 = C_0(\bar v_0, l) > 0$.

In this case note single valuedness of mappings X^*_ε, V^*_ε, and G_ε for sufficiently small $\varepsilon \ge 0$ in view of fulfillment of Lipschitz conditions. Uniqueness of solutions $\bar\psi_0$ and $\bar v_0$ follows from single valuedness of mappings X^*_0, V^*_0.

Assume that operator G_ε is continuous with respect to parameter ε at small $\varepsilon \ge 0$ in the ball $B(\bar v_0, l)$. Therefore, the condition (72) of Theorem 8 is fulfilled for operator G_ε at small $\varepsilon > 0$ in the ball $B(\bar v_0, l)$.

As a result, according to Theorem 5, in view of the estimate (118) the iterative process (113) at small $\varepsilon > 0$ converges in the norm $\|\cdot\|_\infty$ to a unique solution $\bar v \in B(\bar v_0, l)$ of the perturbed problem (112) for any initial approximation $v^0 \in B(\bar v_0, l)$.

Having obtained convergence at certain $\varepsilon < 1$, we will increase ε, by taking the perturbed solution at previous value ε as an initial approximation for the iterative process. In case of convergence as $\varepsilon = 1$ we obtain a solution of the input relation (99).

With the aim of comparison of proposed perturbation method for solving the maximum principle (99) let us represent the known methods in notation used.

The simplest method of successive approximations [12] for solving (99) can be written in the form

$$v^{k+1}(t) = u_1^*(\psi_1(t, v^k), x_1(t, v^k), t), \quad t \in T.$$

Modification of the simplest method of successive approximations (algorithm M1) [13] in the main problem (89) and (90) with $\varphi(x) = \langle c, x\rangle$ is obtained under formation of the perturbed optimal control problem with a parameter $\varepsilon \in [0, 1]$ like (102) and (103) with $A_0(t) \equiv 0$, $b_0(t) \equiv 0$, $a_0(t) \equiv 0$, $d_0(t) \equiv 0$ and using the iterative process

$$v^{k+1}(t) = u_\varepsilon^*(\psi_\varepsilon(t, v^k), x_\varepsilon(t, v^k), t), \quad t \in T.$$

Standard conditional gradient method [8, 9] for solving (99) is described by the relations

$$\bar{v}^k(t) = u_1^*(\psi_1(t, v^k), x_1(t, v^k), t), \quad t \in T,$$

$$v_\lambda^k(t) = v^k(t) + \lambda(\bar{v}^k(t) - v^k(t)), \quad t \in T,$$

$$\lambda \in [0, 1]: \quad \Phi(v_\lambda^k) \le \Phi(v^k) \quad \Rightarrow \quad v^{k+1}(t) = v_\lambda^k(t), \quad t \in T.$$

The needle-shaped linearization method [1] for solving (99) is characterized by the relations

$$\bar{v}^k(t) = u_1^*(\psi_1(t, v^k), x_1(t, v^k), t), \quad t \in T,$$

$$g^k(t) = \Delta_{\bar{v}^k} H(\psi_1(t, v^k), x_1(t, v^k), v^k(t), t), \quad t \in T,$$

$$\lambda_{\min} = \inf_{t \in T} g^k(t), \quad \lambda_{\max} = \sup_{t \in T} g^k(t),$$

$$v_\lambda^k(t) = \begin{cases} v^k(t), & g^k(t) \le \lambda, \\ \bar{v}^k(t), & g^k(t) > \lambda, \end{cases} \quad \lambda \in [\lambda_{\min}, \lambda_{\max}], \quad t \in T,$$

$$\lambda \in [\lambda_{\min}, \lambda_{\max}]: \quad \Phi(v_\lambda^k) \le \Phi(v^k) \quad \Rightarrow \quad v^{k+1}(t) = v_\lambda^k(t), \quad t \in T.$$

The proposed perturbation method does not possess the property of compulsory relaxation on target functional on each iteration in contrast to gradient methods and methods of maximum principle. Compensation of relaxation property is the absence of operation of parametric search for the improving control and obtaining the output controls, acceptable in practice on each iteration.

Note that perturbation method of maximum principle generalizes to the problems with delay in an obvious way.

3.3 Projective Perturbation Method for Optimality Condition

Let us consider the optimality condition (98) in the main problem (89) and (90), represented in the form

$$v(t) = P_U(v(t) + \alpha H_u(\psi(t,v), x(t,v), v, t)),\ t \in T,\ \alpha > 0. \tag{119}$$

We will consider a projection parameter $\alpha > 0$ as a perturbation parameter, we will call the condition (119) as perturbed condition. The unperturbed condition is obtained from (119) as $\alpha = 0$. Any admissible control $v(t)$, $t \in T$, satisfies this condition.

Explicit iterative process of solving the perturbed condition (119) is represented in the form

$$v^{k+1}(t) = P_U(v^k(t) + \alpha H_u(\psi(t,v^k), x(t,v^k), v^k(t), t)), \quad t \in T. \tag{120}$$

Implicit iterative process for solving the system (119) has the form

$$v^{k+1}(t) = P_U(v^k(t) + \alpha H_u(\psi(t,v^k), x(t,v^{k+1}), v^k(t), t)),\ t \in T. \tag{121}$$

On initial (zero) iteration the initial approximation $v^0 \in V$ is given.

For a fixed $\alpha > 0$ let us formulate convergence conditions of processes (120) and (121) on the basis of theorem 5. For this let us describe processes (120) and (121) in the operator form

$$v^{k+1} = U^\alpha(v^k), \quad v \in V, \tag{122}$$

$$v^{k+1} = G^\alpha(v^k), \quad v \in V. \tag{123}$$

Define auxiliary operator V^α by the relation

$$V^\alpha(\psi, x, v) = v^\alpha, \quad v \in V, \quad \psi \in C(T), \quad x \in C(T),$$

$$v^\alpha(t) = P_U(v(t) + \alpha H_u(\psi(t), x(t), v(t), t)), \quad t \in T.$$

Define operator X^α as

$$X^\alpha(p, v) = x^\alpha, \quad p \in C(T), \quad v \in V, \quad x^\alpha(t) = x^\alpha(t, p, v), \quad t \in T,$$

where $x^\alpha(t, p, v)$, $t \in T$, is a solution of Cauchy problem

$$\dot{x}(t) = f(x(t), w^\alpha(p(t), x(t), v(t), t), t), \quad x(t_0) = x^0, \quad t \in T = [t_0, t_1].$$

By using above considered mappings $\Psi : u \to \psi(t,u)$, $t \in T$, and $X : u \to x(t,u)$, $t \in T$, operators U^α, G^α are represented in the form

$$U^\alpha(v) = V^\alpha(\Psi(v), X(v), v), \quad v \in V,$$

$$G^\alpha(v) = V^\alpha(\Psi(v), X^\alpha(\Psi(v), v), v), v \in V.$$

Note fulfillment of the relation

$$X^\alpha(p,v) = X(V^\alpha(p, X^\alpha(p,v), v)), \quad p \in C(T), \quad v \in V.$$

Hence, we obtain

$$X^\alpha(\Psi(v^k), v^k) = X(v^{k+1}),$$

i.e., process (121) can be represented in the implicit operator form

$$v^{k+1} = V^\alpha(\Psi(v^k), X(v^{k+1}), v^k).$$

In view of properties of projection operator P_U, operators U^α, G^α, $\alpha > 0$ are single valued.

The perturbed problem (119) is equivalent to each of the problems

$$v = U^\alpha(v), \quad v \in V, \tag{124}$$

$$v = G^\alpha(v), \quad v \in V. \tag{125}$$

An unperturbed problem is obtained from perturbed problem (124) and (125) as $\alpha = 0$. In this case we have $U^0 : v \to v$, $v \in V$ and $G^0 : v \to v$, $v \in V$.

Convergence of processes (122) and (123) can be justified by using Theorem 5 on the set of admissible controls $V = \{v \in C(T) : v(t) \in U, t \in T\}$.

Assume that a family of phase trajectories for the system (90) is bounded on the set V:

$$x(t,v) \in X, \quad t \in T, \quad v \in V, \tag{126}$$

where $X \subset R^n$ is a convex compact set. Then taking into consideration DMP conditions and linearity of the adjoint system (91), on the basis of sufficient condition (74), we obtain boundedness condition of a family of adjoint trajectories

$$\psi(t,v) \in P, \quad t \in T, \quad v \in V, \tag{127}$$

where $P \subset R^n$ is a convex compact set.

In addition to DMP conditions, we assume that functions $f(x,u,t)$, $F(x,u,t)$, $\varphi(x)$ are twice continuously differentiable in the variables x, u, t on the set $R^n \times U \times T$.

In made assumptions operators X, Ψ satisfy the Lipschitz condition with a constant $C_1 > 0$

$$\|X(v) - X(u)\|_C \le C_1 \|v - u\|_C, \quad v \in V, \quad u \in V,$$

$$\|\Psi(v) - \Psi(u)\|_C \le C_1 \|v - u\|_C, \quad v \in V, \quad u \in V.$$

On the basis of fulfillment of the Lipschitz condition for projection operator P_U and boundedness conditions (126) and (127) we have

$$\begin{aligned}
&\|x^\alpha(t,p,u) - x^\alpha(t,q,v)\| \\
&\quad = \|x(t, V^\alpha(p, X^\alpha(p,u),u)) - x(t, V^\alpha(q, X^\alpha(q,v),v))\| \\
&\quad \le M_3 \int\limits_T \|V^\alpha(p, X^\alpha(p,u))|_t - V^\alpha(q, X^\alpha(q,v))|_t\| \, dt \\
&\quad \le M_4 \int\limits_T \|u(t) - v(t)\| \, dt + \\
&\quad + \alpha M_4 \int\limits_T \|H_u(p(t), x^\alpha(t,p,u), u(t), t) - H_u(q(t), x^\alpha(t,q,v), v(t), t)\| \, dt,
\end{aligned}$$

where $t \in T$, $p, q, u, v \in C(T)$, $M_3 = \text{const} > 0$, $M_4 = \text{const} > 0$. Hence, at a sufficiently small $\alpha > 0$ it is easy to obtain the estimate

$$\|X^\alpha(\Psi(u), u) - X^\alpha(\Psi(v), v)\|_C \le \frac{(1+\alpha)M_1}{(1-\alpha M_2)} \|u - v\|_C ,$$

where $u \in V$, $v \in V$, $M_1 = \text{const} > 0$, $M_2 = \text{const} > 0$.

On the basis of the Lipschitz condition for projection operator P_U we have

$$\begin{aligned}
&\|w^\alpha(p,x,u,t) - w^\alpha(q,y,v,t)\|^2 \le \\
&\le \|(u-v) + \alpha(H_u(p,x,u,t) - H_u(q,y,v,t))\|^2 \\
&\le \|u-v\|^2 + 2\alpha \langle u - v, H_u(p,x,u,t) - H_u(q,y,v,t)\rangle \\
&+ \alpha^2 \|H_u(p,x,u,t) - H_u(q,y,v,t)\|^2 , \\
&\qquad u, v \in U, \quad p, q \in P, \quad x, y \in X, \quad t \in T.
\end{aligned}$$

Assume that for vector-valued function $H_u(\psi, x, u, t)$ the following condition is fulfilled:

$$\begin{aligned}
&\langle u - v, H_u(p,x,u,t) - H_u(q,y,v,t)\rangle \le -K \|u-v\|^2 , \qquad (128) \\
&\qquad u, v \in U, \quad p, q \in P, \quad x, y \in X, \quad t \in T,
\end{aligned}$$

where $K = \text{const} > 0$.

As a result, on the basis of (128) at sufficiently small $\alpha > 0$, we obtain the estimates

$$\begin{aligned}
&\|V^\alpha(\Psi(u), X(u), u) - V(\Psi(v), X(v), v)\|_C \le (1 - 2\alpha K + \alpha^2 M)^{\frac{1}{2}} \|u - v\|_C , \\
&\|V^\alpha(\Psi(u), X^\alpha(\Psi(u), u), u) - V(\Psi(v), X^\alpha(\Psi(v), v), v)\|_C \le \\
&\le (1 - 2\alpha K + \alpha^2 M)^{\frac{1}{2}} \|u - v\|_C , \quad u \in V, \quad v \in V,
\end{aligned}$$

where $M = \text{const} > 0$.

So, in made assumptions, operators U^α, G^α satisfy the Lipschitz condition with a constant less than 1 at a sufficiently small $\alpha > 0$. Note that at $\alpha = 0$ operators U^0, G^0 of unperturbed problems satisfy the Lipschitz condition with a constant, equal to 1; this is in agreement with estimates obtained.

As a result, on the basis of Theorem 5 we obtain the following statement for convergence of processes (122) and (123).

Theorem 8. *Let*

(1) a family of phase trajectories in the main problem (129) *and* (130) *be bounded:* $x(t,u) \in X$, $t \in T$, $u \in V$, *where* $X \subset R^n$ *is a convex compact set;*

(2) vector-valued function $f(x,u,t)$, *functions* $F(x,u,t)$, $\varphi(x)$ *be twice continuously differentiable in the variables* x, u, t *on the set* $R^n \times U \times T$;

(3) for vector-valued function $H_u(\psi,x,u,t)$ *the condition be fulfilled*

$$\langle u-v, H_u(p,x,u,t) - H_u(q,y,v,t)\rangle \le -K\|u-v\|^2,$$

$$u,v \in U, \quad p,q \in P, \quad x,y \in X, \quad t \in T,$$

where $K = const > 0$, $P \subset R^n$ *is a convex compact set that bounds the family of adjoint trajectories:* $\psi(t,u) \in P$, $t \in T$, $u \in V$.

Then for a sufficiently small projection parameter $\alpha > 0$

(1) the relation (119) *has a unique solution* $\bar{v}^\alpha \in V$;

(2) iterative processes (120) *and* (121) *converge in the norm* $\|\cdot\|_C$ *to the solution* $\bar{v}^\alpha$ *for any initial approximation* $v^0 \in V$.

Note that the projective perturbation method is characterized by the extremum control, determined by the condition (119), at any perturbation parameter $\alpha > 0$.

Projective perturbation method easily generalizes to systems with time delay.

With the aim of comparison of developed projective perturbation method let us represent the standard gradient projection method in the notation used [9]

$$\bar{v}^k(t) = w^1(\psi(t,v^k), x(t,v^k), v^k(t), t), \quad t \in T,$$

$$v^k_\lambda(t) = v^k(t) + \lambda(\bar{v}^k(t) - v^k(t)), \quad t \in T,$$

$$\lambda \in [0,1]: \quad \Phi(v^k_\lambda) \le \Phi(v^k) \quad \Rightarrow \quad v^{k+1} = v^k_\lambda.$$

Modification of standard gradient projection method with $\alpha > 0$ is described by relations

$$\bar{v}^k(t) = w^\alpha(\psi(t,v^k), x(t,v^k), v^k(t), t), \quad t \in T,$$

$$v^k_\lambda(t) = v^k(t) + \lambda(\bar{v}^k(t) - v^k(t)), \quad t \in T,$$

$$\lambda \in [0,1]: \quad \Phi(v^k_\lambda) \le \Phi(v^k) \quad \Rightarrow \quad v^{k+1} = v^k_\lambda.$$

The main distinction of constructed projective perturbation method for optimality condition from standard projective methods, and its modifications [1] consists in that the projection parameter $\alpha > 0$ is fixed in iterative process of successive approximations. In gradient projection methods this parameter varies on each iteration in order to provide improvement of control.

On a whole, developed perturbation methods do not guarantee relaxation on target functional in contrast to conditional gradient methods, gradient

projection methods, and their modifications. But the relaxation property is compensated by absence of parametric search of operation for the improving approximations and obtaining controls that are realizable in practice. The above properties are important factors for rise of computational efficiency for solving optimal control problems.

3.4 Numerical Solution for Test Case

Numerical calculations of test problems by applying proposed perturbation methods have illustrated the possibility of considerable decrease of complexity and improvement of solving realizability in comparison with standard methods (conditional gradient, gradient projection, needle-shaped linearization).

For instance, let us show comparative results of solving the known optimal control problem for step electric motor [14]

$$\Phi(u) = \int_T (x_1^2 + k_1 u_1 + k_2 u_2 + k_3 u_3) dt \to \min, \tag{129}$$

$$\dot{x}_1 = x_2, \quad x_1(0) = \pi/3, \tag{130}$$

$$\dot{x}_2 = -a x_2 - b(u_1 \sin(2x_1) + u_2 \sin(2x_1 + \frac{2\pi}{3}) + u_3 \sin(2x_1 - \frac{2\pi}{3})), x_2(0) = 0,$$
$$u_i = u_i(t) \in [0, 16], \quad i = 1, 2, 3, \quad t \in T = [0, 0.05].$$

Here x_1 is a motor shaft position, x_2 is the velocity, components of control u_1, u_2, u_3 correspond to squares of winding current. Performance criterion (129) is determined by requirement for shaft position reduction to zero at minimal energy costs. The values of parameters are $k_i = 0.001$, $i = 1, 2, 3$, $a = 50$, $b = 1,000$.

In [14] the problem (129) and (130) is solved by conditional gradient method (CGM), the first and the second conditional quasigradient methods (CQM-1 and CQM-2) [1].

Here this problem was solved by projective perturbation method (120) for optimality condition (PPMOC).

The problem was computed on PC Celeron 700. Phase and adjoint Cauchy problems were solved numerically by Runge–Kutta–Felberg method of variable (5–6) order and step [15], realized on Fortran PowerStation 4.0. The ratio error for computation of phase and adjoint Cauchy problems was given at 10^{-10}. During computation process values of computed controlled, phase and adjoint variables were stored in uniform grid nodes Ω with discretization step 0.00025 on interval T. In the intervals among neighboring grid nodes value of control was taken to be constant and equal to value of control in the left node.

For an initial approximation of the iterative process (120) a control identically zero was chosen. For a condition of computation stop the following inequality was chosen:

$$|\Phi(u^{k+1}) - \Phi(u^k)| \leq |\Phi(u^k)| \cdot 10^{-4}.$$

The known results of solving problem by methods CGM, CQM-1, CQM-2 [14] and method PPMOC at perturbation parameter $\alpha = 10^2$ are listed in Table 1 (Φ^* is the best calculated value of functional, N is a total number of solved Cauchy problems).

Table 1. Comparative results of numerical calculations

Method	Φ^*	N
CGM	0.00817	617
CQM-1	0.00988	410
CQM-2	0.00792	287
PPMOC	0.00779	309

Figures 1, 2, 3, and 4 shows total control and phase trajectories of solving problem at a scale of a figure, mentioned in [14]. The total component of calculated control $u_2 \equiv 0$ is not shown in the figure.

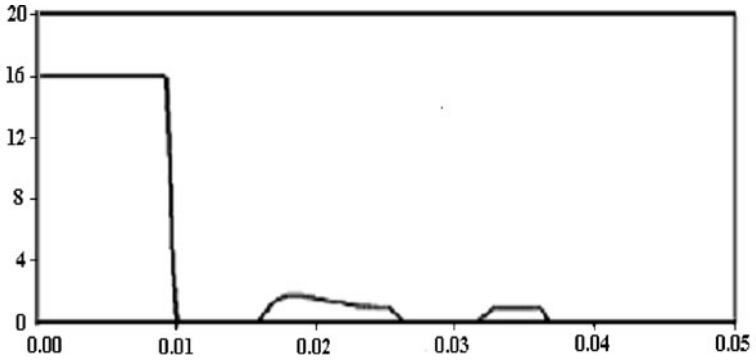


Fig. 1. u_1

According to [5] the controls, obtained by other compared methods, contain oscillatory segments of frequent switchings. This makes the controls unsatisfactory in terms of practical realization. In this case calculated phase trajectories coincide with those shown in Figs. 3 and 4 in a qualitative sense (up to correspondence of figures).

In the context of considered problem, method PPMOC, in contrast to compared methods, permits one to obtain control that is realizable in practice (without oscillatory segments) with the best value of functional. In addition, the iterative process has non-relaxational character.

If perturbation parameter decreases to $\alpha = 50$, the method realizes the same total control in a qualitative sense with $u_2 \equiv 0$ with a precision

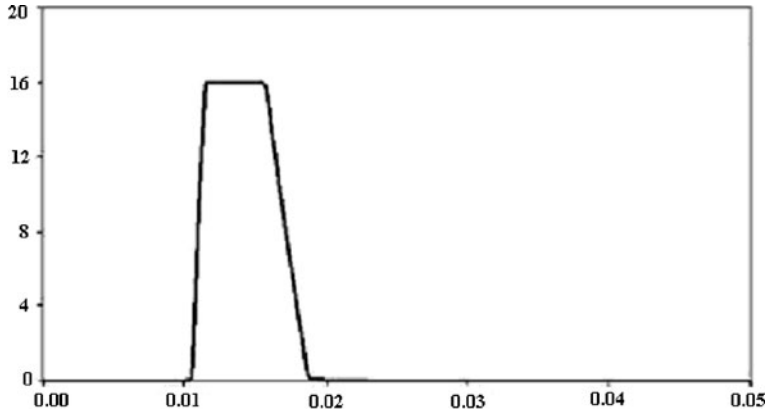


Fig. 2. u_3

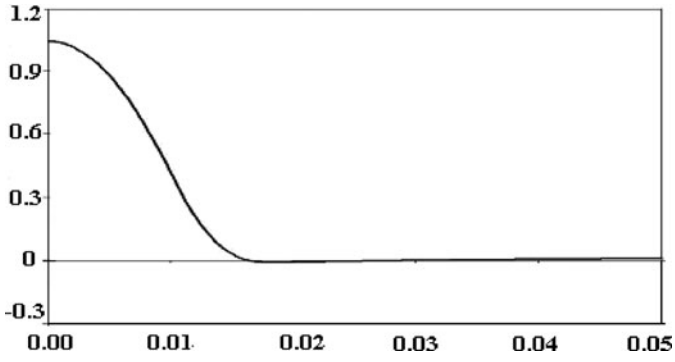


Fig. 3. x_1

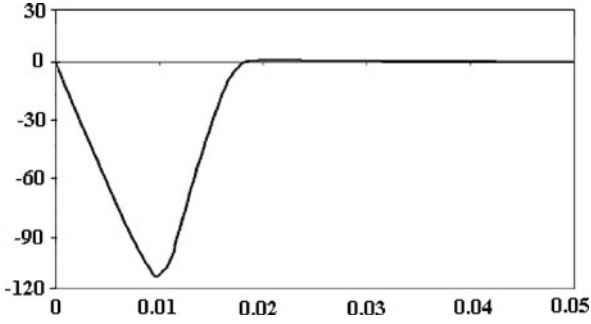


Fig. 4. x_2

acceptable value of functional $\Phi(u) = 0.007830$ at complexity $N = 593$. If perturbation parameter increases to $\alpha = 200$ the iterative process (120) does not converge.

4 Conclusion

The methods, wherein the nonlocal nature of the improvement is provided in quadratic optimal control problem and achieved at the cost of solving boundary-value problem for ordinary differential equations were first developed in the author's works. Such a boundary-value improvement problem is considerably easier than the boundary-value problem of maximum principle, and is reduced to two Cauchy problems in linear case. The proposed approach to nonlocal improvement on the basis of solving boundary-value problem proved possible to generalize to optimal control problem class, that is polynomial with respect to state, including problems with time delay.

The structure of proposed boundary-value problem for nonlocal improvement allows evident isolation of a part linear, with respect to a state that is solved by applying two Cauchy problems and coincides with boundary-value problem in linear case. This property makes it possible to use and prove the perturbation method, known in computational mathematics, in order to solve the problem efficiently. The considered approach does not contain an operation of parametric search for successive approximations and generally forms new perturbation methods for nonlocal improvement in optimal control problems.

The core of the proposed methods consists in entering a parameter into the considered problem, so, that the problem, called as unperturbed, has a simple or evident solution at a certain value of the parameter. As a rule, an unperturbed problem corresponds to zero perturbation parameter. In order to solve perturbed problems at fixed nonzero perturbation parameter iterative algorithms are constructed wherein problem, as difficult as unperturbed problem, is solved on each iteration. In this case a solution of perturbed problem obtained at a smaller value of perturbation parameter is used as an initial approximation of the iterative process.

Constructed perturbation methods do not guarantee relaxation on target functional on each iteration. But this is compensated by absence of operation of parametric search for the improving control, by obtaining solutions admissible on practice, and by simplicity of realization and adjustment to a concrete problem. These properties are essential factors of efficiency upgrading for solving nonlinear optimal control problems.

On the whole, numerical experiments illustrated better quantitative indexes (a number of solved Cauchy problems, value of target functional) and qualitative indexes (realizability of control, approximation of optimal control) for calculating test and model problems by constructed perturbation methods compared to standard methods for local improvement.

The conducted analysis opens new possibilities for efficient use of perturbation method within the framework of optimal control problems, when boundary-value problem of improvement and necessary optimality conditions are proposed to use as parameterization objects.

References

1. Srochko, V. A: Iterative Methods for Solving Optimal Control Problems, Phizmatlit, Moscow. (In Russian) (2000).
2. Pontryagin, L. S., Boltyanskiy, V. G., Gamkrelidze, R. V., Mishchenko, E. F.: Mathematical Theory of Optimal Processes, Nauka, Moscow. (In Russian) (1976).
3. Buldaev, A. S.: Nonlocal control improvement in dynamical systems that are quadratic with respect to state. Izv. Vyssh. Uchebn. Zaved., Mat. 12, 3–9. (In Russian) (2001).
4. Buldaev, A. S.: Nonlocal Improvements and Perturbation Methods in Polinomial and Another Nonlinear Optimal Control Problems. PhD Dissertation, Irkutsk State University, Irkutsk. (In Russian) (2005).
5. Lions Zh, L.: Some Methods of Solving Nonlinear Boundary-value Problems, Mir, Moscow. (In Russian) (1972).
6. Marchuk, G. I., Agoshkov, V. I., Shutyaev, V. P.: Adjoint Equations and Algorithms for Perturbations in Nonlinear Problems of Mathematical Physics, Nauka, Moscow. (In Russian) (1993).
7. Shutyaev, V. P.: Control Operators and Iterative Algorithms in Problems of Variational Data Assimilation, Nauka, Moscow. (In Russian) (2001).
8. Vasil'yev, O. V.: Optimization Methods, World Federation, Atlanta, GA, USA (1996).
9. Vasil'yev, O. V., Arguchintsev, A. V.: Optimization Methods in Problems and Exercises, Moscow: Fizmatlit (In Russian) (1999).
10. Samarskiy, A. A., Gulin, A. V.: Numerical Methods, Nauka, Moscow. (In Russian) (1989).
11. Vasil'yev, F. P.: Numerical Methods for Solving Extremum Problems, Nauka, Moscow. (In Russian) (1980).
12. Krylov, I. A., Chernous'ko, F. L.: On a method of Successive Approximations for Solving Optimal Control. Zh. Vychisl. Mat. Mat. Fiz, 32(6), 1132–1138. (In Russian) (1962).
13. Chernous'ko, F. L.: Phase State Estimation of Dynamical Systems, Nauka, Moscow. (In Russian) (1988).
14. Antonik, V. G., Srochko, V. A.: Questions of Comparative Efficiency of Gradient Type Methods in Optimal Control Problems, Series: Optimization and Control, Issue 9. Irkutsk. State Univ. Publ., Irkutsk. (In Russian), (2003).
15. Novikov, V. A., Novikov, E. A.: Explicit Methods for Solving Stiff Ordinary Differential Equations, Preprint of Computation Centre Siberian Branch Academy USSR No. 629, Novosibirsk. (In Russian) (1985).

Stochastic Optimal Control with Applications in Financial Engineering

Hans P. Geering[1], Florian Herzog[2], and Gabriel Dondi[3]

[1] ETH Zurich, Measurement and Control Laboratory, Zurich, Switzerland
geering@imrt.mavt.ethz.ch
[2] SwissQuant Group AG, Zurich, Switzerland
herzog@swissquant.ch
[3] SwissQuant Group AG, Zurich, Switzerland
dondi@swissquant.ch

Summary. In this chapter, it is shown how stochastic optimal control theory can be used in order to solve problems of optimal asset allocation under consideration of risk aversion. Two types of problems are presented: a problem type with a power utility function with a constant relative risk aversion coefficient and a problem type with an exponential utility function with a constant absolute risk aversion coefficient. The problems can be solved analytically in the unconstrained cases. In order to keep this chapter reasonably self-contained, short introductions to deterministic optimal control theory, stochastic processes, stochastic dynamic systems, and stochastic optimal control theory are given.

Key words: stochastic optimal control, asset management, multi-period portfolio optimization

1 Introduction

The notion "strategic asset allocation" was introduced in Brennan et al. [5] to describe the portfolio optimization problem with time-varying returns and long-term investor objectives. In general, the problem of long-term investments is a well-established research field introduced by Samuelson and Merton [40] and [31–33], respectively. Since then, it is well understood that a short-term portfolio optimization can be very different from long-term portfolio optimization.

In this chapter, continuous-time modeling along the lines of [33] is pursued. Using stochastic optimal control theory, Merton was able to establish important financial economic principles, but due to his very general model formulation, he did not give explicit results for portfolio choice problems. Merton's paper [33] highlights the difficulties in solving complex cases of asset dynamics

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_18,

with stochastic factors, because one has to solve a high-dimensional nonlinear partial differential equation. Until recently, few authors worked on problems similar to [33].

Advances in numerical techniques and the growth of computing power led to the development of numerical solutions to multi-period portfolio optimization problems, which are solved by a discrete state approximation. Examples of this line of research are given in [2, 5, 6, 30]. The use of numerical dynamic programming is very often restricted to few factors, due to the fact that the algorithms use excessive computation time and become numerically unreliable for high dimensions.

Closed-form solutions of the Merton model in continuous time with a single stochastic factor are given in [10, 11, 22, 28, 29]. For closed-form solutions of problems involving two or three stochastic factors, see [7, 34].

In [3], Bielecki et al. present a closed-form solution of the portfolio optimization problem in continuous time for multiple assets and multiple factors with an infinite time horizon. Under the assumption of uncorrelated residuals of the asset prices and the factors, they find the optimal portfolio allocation decision for many assets and many factors. This is an important development for a practical and tractable large-scale asset allocation approach.

Many authors of empirical studies have found evidence that macroeconomic and financial variables, such as long-term interest rates or the dividend-price ratio, are suitable return predictors. Among the identified factors are the short-term interest rate [16, 20], the dividend-price ratio [8, 17], and the yield spread between long-term and short-term bonds [9, 18]. A systematic study to analyze the robustness and the economic significance of return predictors is presented in [37], where 1-month treasury bill rates, 12-month treasury bill rates, the inflation rate, the change in industrial production, and the monetary growth rate were used as factors to explain the US stock returns. Testing a simple allocation strategy, the authors concluded that investors could have exploited the predictability of returns during the volatile markets of the 1970s. In [36], evidence is shown for the predictability of US excess stock returns, based on five monetary policy factors as well as on interest rate spreads and 1-month real interest rates. In [26], it is empirically shown that the excess returns of long-term T-bonds are predictable with factors such as term spread or momentum factors. Furthermore, in [41] the spread between long-term and short-term interest rates and price-earnings ratios were used to predict future up- or downturns of the S&P 500 index. Additional studies on return predictability are cited in the bibliographies of these papers.

These and other studies provide evidence that a dynamic asset allocation strategy provides significant portfolio improvements for investors. None of these studies, however, developed a systematic allocation strategy but rather relied on ad hoc portfolio allocation methods.

In this chapter, a systematic method for dynamic optimal asset allocation is presented which has been proposed by the authors in [23, 25]:

- A utility function is chosen, which is a function of the wealth W of the portfolio at the chosen final time t_1. Furthermore, the utility function involves a parameter γ, which controls the risk aversion of the investment strategy.
- The dynamics of the values of the n risky investment opportunities and of the risk-free money market account are Brownian motions. The drift terms in the dynamics of the risky components and of the return rate of the risk-free account are functions of m economic influence factors. These factors are Brownian motions, too. Moreover, as has been observed in the above-mentioned empirical studies, the increments of the Brownian motions driving the prices and those driving the factors are correlated.
- The rules of asset allocation may allow short selling and hedging for the risky investment opportunities and borrowing money from the risk-free account.
- The solution of an optimal dynamic asset allocation is found using the stochastic optimal control theory. This involves solving the so-called stochastic Hamilton-Jacobi-Bellman partial differential equation and leads to an optimal *feedback* solution at all times t in the investment horizon $t_0 \leq t \leq t_1$.

As discussed above, the correlation between the increments of the Brownian motions driving the prices and those of driving the economic factors can be exploited by using *dynamic* optimal asset allocation.

Conversely, in the uncorrelated case, the stochastic dynamic optimal control problem degenerates to a static stochastic optimization problem. Hence, the resulting optimal investment strategy is "myopic" in this case.

This chapter is structured as follows: In order for this chapter to be fairly self-contained, short introductions into deterministic optimal control theory and to stochastic optimal control theory are given in Sections 2 and 3, respectively. The general formulation of a stochastic optimal control problem for dynamic asset allocation is given in Sections 4.1, 4.2, and 4.3.

In Section 4.4, two problems with the power utility function $\frac{1}{\gamma}W^{\gamma}(t_1)$ are solved. This utility function has a constant coefficient of relative risk aversion.[1] Problem 1 with an unconstrained control set $U = R^n$ admits an analytic solution of the Hamilton-Jacobi-Bellman partial differential equation and of the optimal state feedback control law. In Problem 2, the control constraint set is required to be closed and bounded; more specifically, in this problem, no short selling and no borrowing of money are allowed (see (142) and (143)). Unfortunately, this problem cannot be solved analytically, i.e., numerical methods are needed.

In Section 4.5, two problems with the exponential utility function $-\frac{1}{\gamma}e^{-\gamma W(t_1)}$ are solved. This utility function has a constant coefficient of absolute risk aversion.[2] Problem 3 with an unconstrained control set $U = R^n$

[1] CRRA: constant relative risk aversion.
[2] CARA: constant absolute risk aversion.

also admits an analytic solution. In Problem 4, the control constraint set is required to be closed and bounded. Unfortunately, this problem must be solved with numerical methods.

In Section 5, the potential of the stochastic optimal control approach to asset management is highlighted and open problems for further research are outlined.

The reader is encouraged to consult Appendix A, which contains a compendium of some of the notations used throughout this chapter. In particular, note that for a function $f : R \to R^n$, the row vector of its partial derivatives is denoted by f_x (Jacobian), whereas the column vector of its partial derivatives is denoted by $\nabla_x f$ (gradient).

2 Deterministic Optimal Control

In this section, the following deterministic optimal control problem is considered for a dynamic system with the state vector $x(t) \in R^n$ and the admissible control vector $u(t) \in U \subseteq R^m$ (where U is a time-invariant, convex, and closed subset of R^m).

Problem: For the dynamic system described by the differential equation

$$\dot{x}(t) = f(x(t), u(t)) \tag{1}$$

with the given initial state x_0 at the fixed initial time t_0

$$x(t_0) = x_0 \ , \tag{2}$$

find a piecewise continuous control vector $u(t) \in U$ for all times t in the fixed time interval $[t_0, t_1]$, such that the objective functional

$$J = K(x(t_1)) + \int_{t_0}^{t_1} L(x(t), u(t))\, dt \tag{3}$$

is maximized.

In order to solve this optimal control problem, it is useful to introduce the so-called Hamilton function or Hamiltonian

$$H(x(t), u(t), \lambda(t)) = L(x(t), u(t)) + \lambda^T(t) f(x(t), u(t)) \ , \tag{4}$$

where $\lambda(t)$ is an unspecified n-vector function (at this time).

2.1 Theory

In order to find an open-loop optimal solution for this problem, the following necessary conditions (generally called Pontryagin's maximum principle) can be exploited.

Pontryagin's maximum principle: If the control trajectory $u^o(\cdot)$ generating the state trajectory $x^o(\cdot)$ is optimal, the following equations are satisfied:

$$\dot{x}^o(t) = f(x^o(t), u^o(t)), \tag{5}$$
$$x^o(t_0) = x_0, \tag{6}$$
$$u^o(t) \in U \quad \text{for all} \quad t \in [t_0, t_1], \tag{7}$$
$$\dot{\lambda}^o(t) = -\nabla_x H(x^o(t), u^o(t), \lambda^o(t)), \tag{8}$$
$$\lambda^o(t_1) = \nabla_x K(x^o(t_1)), \tag{9}$$
$$H(x^o(t), u^o(t), \lambda^o(t)) \geq H(x^o(t), u, \lambda^o(t)) \quad \text{for all } u \in U \text{ and all } t \in [t_0, t_1]\ . \tag{10}$$

Proof. A proof based on geometrical ideas can be found in the seminal paper [21] and in [1]. A proof based on the calculus of variations can be found in [19] and many other publications.

Often, it is possible to transform the optimal open-loop solution into the preferred closed-loop solution in the form of a state feedback control law. In some cases, the closed-loop solution can directly be obtained using the sufficient conditions of the Hamilton-Jacobi-Bellman theory.

Hamilton-Jacobi-Bellman Theorem: If the Hamiltonian H has a unique admissible H-maximizing control $\widetilde{u}(x, \lambda)$, i.e., if the inequality

$$H(x, \widetilde{u}(x, \lambda), \lambda) \geq H(x, u, \lambda) \tag{11}$$

is satisfied for all $u \in U$, all $x \in R^n$, and all $\lambda \in R^n$, and if the following partial differential equation for the so-called optimal cost-to-go function $\mathcal{J}(x, t)$

$$-\frac{\partial \mathcal{J}(x,t)}{\partial t} = H\left(x, \widetilde{u}(x, \nabla_x \mathcal{J}(x,t)), \nabla_x \mathcal{J}(x,t)\right) \tag{12}$$

with the boundary condition

$$\mathcal{J}(x, t_1) = K(x) \tag{13}$$

admits a unique solution, the optimal state feedback control is

$$u^o(t) = \widetilde{u}\left(x^o(t), \nabla_x \mathcal{J}(x^o(t), t)\right)\ . \tag{14}$$

Proof. A proof can be found in [1].

Remark 1. So far, optimal control problems have been considered, where the objective functional is to be maximized. For optimal control problems, where the objective functional is to be minimized, the Hamiltonian H must be minimized (Pontryagin's minimum principle) and we are looking for an H-minimizing control $\widetilde{u}(x, \lambda)$ in the Hamilton-Jacobi-Bellman theorem.

2.2 Example: An LQ Optimal Control Problem

Problem: For the linear, completely controllable[3] system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t) \tag{15}$$

with the initial state

$$x(t_0) = x_0, \tag{16}$$

find the unconstrained optimal control u (i.e., $U = R^m$), such that the objective functional

$$\begin{aligned} J = \; & \frac{1}{2}x^T(t_1)Fx(t_1) + g^T x(t_1) \\ & + \frac{1}{2}\int_{t_0}^{t_1} \big(x^T(t)Q(t)x(t) + u^T(t)R(t)u(t) \\ & \qquad + x^T(t)S(t)u(t) + u^T(t)S^T(t)x(t)\big)\, dt \end{aligned} \tag{17}$$

with

$$F \succeq 0 \tag{18}$$

and

$$\begin{bmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{bmatrix} \succeq 0 \text{ for all } t \in [t_0, t_1] \tag{19}$$

with

$$R(t) \succ 0 \text{ for all } t \in [t_0, t_1] \tag{20}$$

is minimized.

For this optimal control problem, the Hamiltonian is

$$H = \frac{1}{2}\left(x^T Q x + u^T R u + x^T S u + u^T S^T x\right) + \lambda^T \left(Ax + Bu + b\right) \; . \tag{21}$$

The H-minimizing control is

$$u = -R^{-1}(B^T\lambda + S^T x) \; . \tag{22}$$

Therefore, exploiting Pontryagin's minimum principle leads to the following two-point boundary value problem:

$$\dot{x} = (A - BR^{-1}S^T)x - BR^{-1}B^T\lambda + b, \tag{23}$$

$$x(t_0) = x_0, \tag{24}$$

$$\dot{\lambda} = -(Q - SR^{-1}S^T)x - (A - BR^{-1}S^T)^T\lambda, \tag{25}$$

$$\lambda(t_1) = Fx(t_1) + g \; . \tag{26}$$

[3] For some background material about controllability, the reader is referred to Appendix B.

Solving this two-point boundary value problem yields the optimal trajectories $x^o(\cdot)$ and $\lambda^o(\cdot)$ and hence the open-loop optimal control law

$$u^o(t) = -R^{-1}(t)\left(B^T(t)\lambda^o(t) + S^T(t)x^o(t)\right) \ . \tag{27}$$

In order to convert this open-loop optimal control law into a theoretically equivalent (but preferable) closed-loop control law, the following ansatz is useful:

$$\lambda(t) = K(t)x(t) + k(t) \ , \tag{28}$$

where $K(\cdot)$ is an n by n matrix function and $k(\cdot)$ an n-vector function which remains to be found.

Combining its differentiated form

$$\dot{\lambda}(t) = \dot{K}(t)x(t) + K(t)\dot{x}(t) + \dot{k}(t) \tag{29}$$

with the differential equations of the two-point boundary value problem yields the equation

$$\begin{aligned} &\big[\dot{K} + K(A - BR^{-1}S^T) + (A - BR^{-1}S^T)^T K \\ &\qquad\qquad - KBR^{-1}B^T K + Q - SR^{-1}S^T\big]x \\ &= \ -Kb - \dot{k} - (A - BR^{-1}B^T K - BR^{-1}S^T)^T k \ . \end{aligned} \tag{30}$$

This equation must hold for an arbitrary vector $x(t) \in R^n$ since the initial state x_0 is an arbitrary vector in R^n. In other words, both the brackets on the left-hand side and the right-hand side of this equation vanish. This yields the following differential equations for the matrix function $K(t)$ and the vector function $k(t)$:

$$\begin{aligned} \dot{K} = \ &-K(A - BR^{-1}S^T) - (A - BR^{-1}S^T)^T K \\ &+ KBR^{-1}B^T K - Q + SR^{-1}S^T, \end{aligned} \tag{31}$$

$$\dot{k} = \ -(A - BR^{-1}B^T K - BR^{-1}S^T)^T k - Kb, \tag{32}$$

with the boundary conditions

$$K(t_1) = F, \tag{33}$$

$$k(t_1) = g \ . \tag{34}$$

Of course, the resulting optimal feedback control law

$$u(t) = \ -R^{-1}(t)\left(B^T(t)K(t) + S^T(t)\right)x(t) - R^{-1}(t)k(t) \tag{35}$$

can directly be determined using the Hamilton-Jacobi-Bellman theory. This is left to the reader as an exercise, because the stochastic version of this problem is treated in Section 3.5 in detail.

3 Stochastic Optimal Control

3.1 Stochastic Processes

Definition 1. Brownian motion [4]
A stochastic process W(t) is called Brownian motion if it satisfies the following conditions:

1. *Independence:* $W(t+\Delta t) - W(t)$ *is independent of* $\{W(\tau)\}$ *for all* $\tau \le t$.
2. *Stationarity: The distribution of* $W(t+\Delta t) - W(t)$ *does not depend on* t.
3. *Continuity:* $\lim\limits_{\Delta t \downarrow 0} \dfrac{P(|W(t+\Delta t) - W(t)| \ge \delta)}{\Delta t} = 0$ *for all* $\delta > 0$.

Please note that the third assumption is expressed with probabilities: discontinuities in sample functions can only occur with probability zero. Hence, there is a version of the Brownian motion with *all* sample functions continuous. (This technicality is not of any practical importance.)

This definition induces the distribution of the process $W(t)$.

Theorem 1. (Normally distributed increments of Brownian motion)
If $W(t)$ *is a Brownian motion, then* $W(t) - W(0)$ *is a normal random variable with mean* μt *and variance* $\sigma^2 t$*, where* μ *and* σ *are constant real numbers.*

As a result of this theorem, we have the following density function of a Brownian motion:

$$f_{W(t)}(x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{(x-\mu t)^2}{2\sigma^2 t}} . \tag{36}$$

An irritating property of Brownian motion is that its sample paths are not differentiable. This is easily verified in the mean-square sense:

$$E\left[\left(\frac{W(t+\Delta t) - W(t)}{\Delta t}\right)^2\right] = \frac{E[(W(t+\Delta t) - W(t))^2]}{\Delta t^2} = \frac{\sigma^2}{\Delta t} . \tag{37}$$

This diverges for $\Delta t \to 0$ and therefore $W(\cdot)$ is not differentiable in L^2.

The Brownian motion $W(\cdot)$ (starting at $W(0) = 0$) has many more bizarre and intriguing properties. Some of them are listed below:

- Autocovariance function: $E\{(W(t) - \mu t)(W(\tau) - \mu\tau)\} = \sigma^2 \min(t, \tau)$
- $\mathrm{Var}\left\{\dfrac{W(t)}{t}\right\} = \dfrac{\sigma^2}{t}$
- $\lim\limits_{t\to\infty} \frac{W(t)-\mu t}{t} = 0$ with probability 1
- The total variation of the Brownian motion over a finite interval $[0, T]$ is infinite!
- The "sum of squares" of a drift-free Brownian motion is deterministic:
  $\lim\limits_{N\to\infty} \sum\limits_{k=1}^{N} \left(W\left(k\frac{T}{N}\right) - W\left((k-1)\frac{T}{N}\right)\right)^2 = \sigma^2 T.$
  Important consequence: Whenever the term dW^2 appears in a stochastic differential equation, it should be replaced by $\sigma^2 dt$.

- Zero-crossings: In a finite interval $[0, T]$, every sample of a drift-free Brownian motion has infinitely many zero-crossings. The set of zero-crossings is dense in $[0, T]$, i.e., no sample path has isolated zero-crossings!

Definition 2. Standard Brownian motion
A Brownian motion is called standard if

$$W(0) = 0, \tag{38}$$
$$E[W(t)] = 0 \quad (\mu = 0), \tag{39}$$
$$E[W^2(t)] = t \quad (\sigma^2 = 1). \tag{40}$$

In the sequel, a Brownian motion is assumed to be a standard Brownian motion unless explicitly stated otherwise. In most cases, we use the differential form

$$dW(t) = \lim_{\tau \downarrow 0} W(t + \tau) \tag{41}$$

with $E[dW(t)] = 0$ and the sum-of-squares property $E[dW^2(t)] = dt$.

The generalization of a Brownian motion from the scalar case to the vector case is straightforward: The scalar drift parameter μ becomes a vector; and the "volatility parameter" σ and the "intensity parameter" σ^2 become symmetric, positive-definite matrices. The notation in the vector case will be Σ instead of σ^2 and $\Sigma^{1/2}$ instead of σ.

In the case of a vector-valued standard Brownian motion, it will be assumed that the component processes of the vector are mutually independent.

3.2 Stochastic Differential Equations

A non-standard Brownian motion $X(\cdot)$ satisfies the stochastic differential equation

$$dX(t) = \mu dt + \sigma dW(t), \tag{42}$$
$$X(0) = 0\ , \tag{43}$$

where $W(\cdot)$ is a standard Brownian motion.

In financial engineering, the following stochastic processes are also of interest:

The geometric Brownian motion $X(\cdot)$ is described by the differential equation

$$dX(t) = \mu X(t)dt + \sigma X(t)dW(t)\ . \tag{44}$$

It is popular for modeling stock prices.

A mean reverting stochastic process $X(\cdot)$ can be modeled by the differential equation

$$dX(t) = \kappa[\mu - X(t)]dt + \sigma dW(t) \tag{45}$$

with $\kappa > 0$. It is popular for modeling interest rates.

In the most general nonlinear case, the stochastic differential equation for a stochastic process can be written as follows:

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t) \ . \tag{46}$$

3.3 Stochastic Calculus

Due to the "sum-of-squares" property of the Brownian motion, the rules of differentiation in the stochastic case differ from those in the deterministic case.

Consider the following problem: Given a stochastic differential equation for the process $X(\cdot)$

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t), \tag{47}$$
$$X(t_0) = X_0 \ , \tag{48}$$

find the differential equation for the process $Y(t)$ which is a function of $X(t)$,

$$Y(t) = \phi(t, X(t)) \ , \tag{49}$$

where the function $\phi(t, X)$ is continuously differentiable in t and twice continuously differentiable in X.

Let us do a Taylor series expansion of (49) up to second-order terms:

$$\begin{aligned} dY(t) &= \phi_t(t, X)dt + \phi_x(t, X)dX(t) + \frac{1}{2}\phi_{tt}(t, X)dt^2 \\ &\quad + \frac{1}{2}\phi_{xx}(t, X)(dX(t))^2 + \phi_{xt}(t, X)dX(t)dt + \text{higher order terms} \\ &= \phi_t(t, X)dt + \phi_x(t, X)[f(t, X(t))dt + g(t, X(t))dW(t)] \\ &\quad + \frac{1}{2}\phi_{tt}(t, X)dt^2 + \frac{1}{2}\phi_{xx}(t, X)[f(t, X(t))dt + g(t, X(t))dW(t)]^2 \\ &\quad + \phi_{xt}(t, X)[f(t, X(t))dt + g(t, X(t))dW(t)]dt + \text{h.o.t.} \end{aligned} \tag{50}$$

Notice that the term $dW^2(t)$ appears when the square factor of ϕ_{xx} is expanded. Replacing it by dt and retaining only the terms of first order yield the following result:

$$\begin{aligned} dY(t) = &\left[\phi_t(t, X) + \phi_x(t, X)f(t, X(t)) + \frac{1}{2}\phi_{xx}(t, X)g^2(t, X(t))\right]dt \\ &+ \phi_x(t, X)g(t, X(t))dW(t), \end{aligned} \tag{51}$$
$$Y(t_0) = \phi(t_0, X_0) \ . \tag{52}$$

The term $\frac{1}{2}\phi_{xx}g^2dt$ is called "Itô correction term."

In the more general case where the stochastic process $X(t) \in R^n$ and the standard Brownian motion $W(t) \in R^m$ are vectors but where the function ϕ is still scalar-valued, the generalized form of (51) is

$$dY(t) = \tilde{f}(t, X(t))dt + \tilde{g}(t, X(t))dW(t) \tag{53}$$

with

$$\begin{aligned}\tilde{f}(t, X(t)) = \; & \phi_t(t, X(t)) + \phi_x(t, X(t))f(t, X(t)) \\ & + \frac{1}{2}\mathrm{tr}\left(\phi_{xx}(t, X(t))g(t, X(t))g^T(t, X(t))\right)\end{aligned} \tag{54}$$

and

$$\tilde{g}(t, X(t)) = \phi_x(t, X(t))g(t, X(t)) \ , \tag{55}$$

where "tr" denotes the trace operator.

Due to the properties of the trace operator for square matrices, the Itô correction term can be written in the following three equivalent forms:

$$\begin{aligned}& \frac{1}{2}\mathrm{tr}\left(\phi_{xx}(t, X(t))g(t, X(t))g^T(t, X(t))\right) \\ = \; & \frac{1}{2}\mathrm{tr}\left(g(t, X(t))g^T(t, X(t))\phi_{xx}(t, X(t))\right) \\ = \; & \frac{1}{2}\mathrm{tr}\left(g^T(t, X(t))\phi_{xx}(t, X(t))g(t, X(t))\right) \ .\end{aligned} \tag{56}$$

Besides its aesthetic symmetric form, the last version has the advantage that the trace operator is not needed in the case of a scalar Brownian motion $W(\cdot)$, i.e., for $m = 1$.

For more information on stochastic calculus, the reader is referred to [15, 35, 42].

3.4 Stochastic Optimal Control Theory

In this section, the following stochastic optimal control problem is considered for a dynamic system with the state vector $x(t) \in R^n$, the admissible control vector $u(t) \in U \subseteq R^m$ (where U is a time-invariant, convex, and closed subset of R^m), and the standard vector Brownian motion $W(t) \in R^k$.

Problem: For the dynamic system described by the stochastic differential equation

$$dx(t) = f(x(t), u(t))dt + g(x(t), u(t))dW(t) \tag{57}$$

with the given deterministic initial state x_0 at the fixed initial time t_0,

$$x(t_0) = x_0 \ , \tag{58}$$

find a piecewise continuous control vector $u(t) \in U$ for all times t in the fixed time interval $[t_0, t_1]$, such that the objective functional

$$J = E\left[K(x(t_1)) + \int_{t_0}^{t_1} L(x(t), u(t))\, dt\right] \tag{59}$$

is maximized.

Pontryagin's maximum principle

It is possible to postulate a Pontryagin's maximum principle for the considered stochastic optimal control problem. Of course, the rules of stochastic differentiation have to be considered in order to arrive at its correct formulation.

However, this is not of interest here because solving the two-point boundary value problem is not practical in the stochastic case.

Hamilton-Jacobi-Bellman Theory:

Theorem 2. Stochastic Hamilton-Jacobi-Bellman Theorem
If the partial differential equation

$$-\mathcal{J}_t(x,t) = \max_{u\in U}\left\{L(x,u) + \mathcal{J}_x(x,t)f(x,u) + \frac{1}{2}\mathrm{tr}\left(\mathcal{J}_{xx}(x,t)g(x,u)g^T(x,u)\right)\right\} \tag{60}$$

with the boundary condition

$$\mathcal{J}(x,t_1) = K(x) \tag{61}$$

admits a unique solution, the globally optimal state feedback control law is

$$u(x) = \arg\max_{u\in U}\ \left\{L(x,u) + \mathcal{J}_x(x,t)f(x,u) + \frac{1}{2}\mathrm{tr}\left(\mathcal{J}_{xx}(x,t)g(x,u)g^T(x,u)\right)\right\}\ . \tag{62}$$

Proof. A rigorous proof of this theorem can be found in [44].

Of course, $\mathcal{J}(x_0,t_0)$ is the optimal value of the objective functional. And again, for stochastic optimal control problems, where the objective functional (59) is to be minimized, the max operator appearing in (60) and (62) must be replaced by the min operator.

In practice (when the problem cannot be solved analytically), the following iterative procedure is applied:

1. For a given function $\mathcal{J}(x,t)$, find $u(x,\mathcal{J}_x,\mathcal{J}_{xx},t)$ satisfying (62), with $x(t)$ replaced by x.
2. Solve the Hamilton-Jacobi-Bellman partial differential equation (60), eliminating the max operator and plugging in the control $u(x,\mathcal{J}_x,\mathcal{J}_{xx},t)$ found in step 1.
3. Return to step 1.

Under suitable convexity assumptions for $K(x)$ and $L(x,u)$ (for the existence of a unique optimal control), this procedure converges, see [24, 38].

3.5 Example: An LQ Optimal Control Problem

In this section, a stochastic version of the deterministic optimal control problem of Section 2.2 is analyzed.

Problem: For the linear, completely controllable, stochastic system

$$dx(t) = [A(t)x(t) + B(t)u(t) + b(t)]dt + [C(t)x(t) + \sigma(t)]dW(t) \tag{63}$$

with the deterministic initial state

$$x(t_0) = x_0, \tag{64}$$

find the unconstrained optimal control u, such that the objective functional

$$\begin{aligned} J = \; E\Bigg[& \frac{1}{2}x^T(t_1)Fx(T) + g^T x(t_1) \\ & + \frac{1}{2}\int_0^T \big(x(t)^T Q(t)x(t) + x(t)^T S(t)u(t) \\ & + u(t)^T S^T(t)x(t) + u(t)^T R(t)u(t)\big)dt\Bigg] \end{aligned} \tag{65}$$

with

$$F \geq 0 \tag{66}$$

and

$$\begin{bmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{bmatrix} \geq 0 \text{ for all } t \in [t_0, t_1] \tag{67}$$

with

$$R(t) > 0 \text{ for all } t \in [t_0, t_1] \tag{68}$$

is minimized.

For a yet unknown cost-to-go function $\mathcal{J}(x,t)$, (62) yields

$$u(x, \mathcal{J}_x, \mathcal{J}_{xx}, t) = -R^{-1}(t)\left(B^T(t)\mathcal{J}_x^T(x,t) + S^T(t)x\right) \; . \tag{69}$$

With this optimal control law, the Hamilton-Jacobi-Bellman partial differential equation (60) has the following form (using simplified notation):

$$\begin{aligned} 0 = \; & \mathcal{J}_t + \frac{1}{2}\Big\{ x^T Q x - x^T S R^{-1} S^T x - \mathcal{J}_x B R^{-1} B^T \mathcal{J}_x^T \\ & + x^T (A - BR^{-1}S^T)^T \mathcal{J}_x^T + \mathcal{J}_x (A - BR^{-1}S^T)x \\ & + x^T C^T \mathcal{J}_{xx} C x + \sigma^T \mathcal{J}_{xx}\sigma + x^T C^T \mathcal{J}_{xx}\sigma \\ & + \sigma^T \mathcal{J}_{xx} C x + b^T \mathcal{J}_x^T + \mathcal{J}_x b \Big\} \; . \end{aligned} \tag{70}$$

Since the objective functional (65) is quadratic and the side constraints (63) and (64) are linear, a quadratic ansatz for the cost-to-go function $\mathcal{J}(x,t)$ of the form

$$J(x,t) = \frac{1}{2}x^T K(t)x + k^T(t)x + c(t) \tag{71}$$

with

$$\mathcal{J}_x(x,t) = K(t)x + k(t) \tag{72}$$

$$\mathcal{J}_{xx}(x,t) = K(t) \tag{73}$$

should be successful, where $K(\cdot)$ is a symmetric matrix function.

Combining (65, 70, 72, and 73) and taking into account that $x \in R^n$ is an arbitrary vector lead to the following equations defining the coefficients $K(t)$, $k(t)$, and $c(t)$ of the ansatz (71):

$$\begin{aligned} \dot{K} = & -K(A - BR^{-1}S^T) - (A - BR^{-1}S^T)^T K \\ & + KBR^{-1}B^T K - Q + SR^{-1}S^T - C^T KC, \end{aligned} \tag{74}$$

$$K(t_1) = F, \tag{75}$$

$$\dot{k} = -(A - BR^{-1}B^T K - BR^{-1}S^T)^T k - Kb - C^T K\sigma, \tag{76}$$

$$k(t_1) = g, \tag{77}$$

$$\dot{c} = -\frac{1}{2}\sigma^T K\sigma + \frac{1}{2}k^T BR^{-1}B^T k - b^T k, \tag{78}$$

$$c(t_1) = 0\ . \tag{79}$$

4 Applications in Financial Engineering

4.1 Introduction

In this section, several stochastic optimal control problems in financial engineering in the area of optimal asset allocation are stated and solved.

In Section 4.2, two viable objective functionals are presented. It is shown that the exponential utility function has a constant absolute risk aversion (CARA) coefficient, whereas the power utility function has a constant relative risk aversion (CRRA) coefficient.

In Section 4.3, the dynamics of a portfolio are given which consists of investments in a risk-free money market account and in n risky investments in stock market indices or even shares of individual companies. The return rate of the risk-free account and both the drift term and the volatility term in the differential equations of the risky assets are allowed to be functions of m economic influence factors $x_1(t), \ldots, x_m(t)$. These economic factors are assumed to be stochastic as well. The increments of the Brownian motions $dZ_p(t)$ driving the prices $P(t)$ of the risky assets and $dZ_q(t)$ driving the economic factors $x(t)$ are allowed to be correlated.

In Sections 4.4 and 4.5, some problems of optimal asset management are solved. It is shown that these continuous-time problems can analytically be solved as long as short selling and borrowing money are unlimited. In the restricted cases, the relevant equations have to be solved numerically.

4.2 Utility Functions

Utility Functions

In financial engineering, the objective functional for the stochastic optimal control problem for asset allocation is taken as the expected value $E[V]$ of the so-called utility function $V(W(t_1))$. In such a problem, the simplest utility function would be maximizing the wealth $W(t_1)$ or value of the considered portfolio at the considered final time t_1:

$$V = W(t_1) \ . \tag{80}$$

Unfortunately, this is not a good utility function as it has experimentally been shown in the first decade of the twenty-first century (resulting in trillions of US dollars of losses to investors and even tax payers). Why? Because we are living on a sample path rather than on the mean (or expected) path of an optimally (or sub-optimally) controlled stochastic (and often poorly modeled) process and, therefore, we do not accept maximal risk (as measured by the variance of $W(t_1)$).

Therefore, in practice, some measure of risk aversion has to be built into the utility function of a considered optimal asset allocation problem.

In order for the utility function $V(W(t_1))$ to admit a unique optimal value, it must be strictly concave in $W(t_1)$.

The following two utility functions are rather popular in financial engineering in risk-averting optimal asset allocation problems:

Exponential utility function:

$$V = -\frac{1}{\gamma}\, e^{-\gamma W(t_1)} \quad \text{for } \gamma > 0 \ , \tag{81}$$

Power utility function:

$$V = \frac{1}{\gamma}\, W^{\gamma}(t_1) \quad \text{for } \gamma < 1, \ \gamma \neq 0 \ . \tag{82}$$

Risk Aversion Coefficients

For measuring the risk aversion, the so-called Arrow-Pratt risk aversion coefficients have been defined [39]:

Absolute risk aversion coefficient:

$$a(W) = -\frac{\frac{\partial^2 V}{dW^2}}{\frac{\partial V}{\partial W}}. \tag{83}$$

Relative risk aversion coefficient:

$$r(W) = -W\frac{\frac{\partial^2 V}{dW^2}}{\frac{\partial V}{\partial W}} \ . \tag{84}$$

For the exponential utility function (81),

$$a(W) = \gamma \quad \text{and} \quad r(W) = \gamma W \qquad (\gamma > 0) \tag{85}$$

and for the power utility function (82),

$$a(W) = \frac{1-\gamma}{W} \quad \text{and} \quad r(W) = 1-\gamma \qquad (\gamma < 1,\ \gamma \neq 0) \tag{86}$$

are obtained, respectively.

4.3 Wealth Dynamics

4.3.1 Risk-Free Money Market Account

The value $P_0(t)$ invested in a risk-free money market account evolves according to the following differential equation:

$$dP_0(t) = r(x(t))P_0(t)dt, \tag{87}$$

$$P_0(t_0) = P_{00}\ . \tag{88}$$

4.3.2 Risky Investments

The value $P(t) \in R^n$ of the vector of risky investments evolves according to the following stochastic differential equation:

$$dP(t) = \mathrm{diag}[P(t)]\left[\mu(x(t))dt + \Sigma_p^{1/2}(x(t))dZ_p(t)\right], \tag{89}$$

$$P(t_0) = P_0\ . \tag{90}$$

Here, $\mu \in R^n$ is the drift vector and $\Sigma_p^{1/2} \in R^{n\times n}$ the positive-definite volatility matrix. Both of them are functions of the measurable, instantaneous vector x of economic influence factors. And $dZ_p \in R^n$ is the increment of a (normalized) Brownian motion.

4.3.3 Economic Influence Factors

The value $x(t) \in R^m$ of the economic influence vector is modeled to evolve according to the following differential equation:

$$dx(t) = [Ax(t)+a]\,dt + \Sigma_q^{1/2}\,dZ_q(t), \tag{91}$$

$$x(t_0) = x_0, \tag{92}$$

with $A \in R^{m\times m}$, $a \in R^m$, $\Sigma_q^{1/2} > 0 \in R^{m\times m}$, and the (normalized) Brownian motion $dZ_q(t) \in R^m$. Often, A is postulated to be a diagonal matrix.

In some of the problems, the Brownian motions Z_p and Z_q are allowed to be correlated, i.e.,

$$\mathrm{Cov}\left(\begin{bmatrix} dZ_p(t) \\ dZ_q(t) \end{bmatrix}\right) = \begin{bmatrix} I & \rho \\ \rho^T & I \end{bmatrix} dt \ > \ 0 \tag{93}$$

with a suitable correlation matrix $\rho \in R^{n\times m}$.

The economic influence factors $x_i(t)$ may include the following: at the macroeconomic level: GDP growth rate, long-term interest rate, inflation rate, etc.; at the industry-specific level: sector growth rate, industry rate of returns, etc.; and at the company-specific level: dividends, cash flow, etc.

4.3.4 Wealth Dynamics

At all times, we are fully invested in risky investments and/or the risk-free money market account. The relative levels of investment are $u_1(t), \ldots, u_n(t)$ for the risky investments and $u_0(t)$ for the risk-free investment.

Being fully invested at all times means

$$\sum_{i=0}^{n} u_i(t) \equiv 1 \ . \tag{94}$$

In classical investment practice, the inequalities

$$u_i(t) \geq 0 \quad \text{for} i = 0, \ldots, n \tag{95}$$

apply. However, if unlimited borrowing and unlimited short selling are permissible, these inequalities do not apply (provided the savings rate and the borrowing rate on the risk-free account are identical).

The wealth $W(t)$, i.e., the value of the portfolio, satisfies the following stochastic differential equation:

$$\begin{aligned} dW(t) = \ & W(t)u_0(t)r(x(t))dt + h(x(t),t)dt \\ & + W(t)u^T(t)\left[\mu(x(t))dt + \Sigma_p^{1/2}(x(t))dZ_p(t)\right] \ , \end{aligned} \tag{96}$$

where $u = [u_1, \ldots, u_n]^T \in R^n$ is the vector of the relative investments in the n risky investment opportunities and h is an additional inflow[4] (for $h > 0$) or outflow (or consumption term for $h < 0$). If $h(x(t),t) \equiv 0$ for all $t \in [t_0, t_1]$, the portfolio is called self-financing.

Of course, the restriction (94) applies in (96). The extra degree of freedom u_0 can be removed in the following way:

$$u_0(t) = 1 - \sum_{i=1}^{n} u_i(t) = 1 - e^T u(t) \ , \tag{97}$$

[4] Mnemonic: "h" as in "help".

where

$$e = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in R^n \ . \tag{98}$$

Thus, the intermediate form of the wealth dynamics is

$$\begin{aligned} dW(t) = \ & W(t)r(x(t))dt + h(x(t),t)dt \\ & + W(t)u^T(t)\left[[\mu(x(t)) - er(x(t))]\,dt + \Sigma_p^{1/2}(x(t))dZ_p(t)\right] \ . \end{aligned} \tag{99}$$

So far, the drift terms $r(x(t)) \in R$ and $\mu(x(t)) \in R^n$ and the volatility matrix $\Sigma_p^{1/2}(x(t)) \in R^{n\times n}$ have been allowed to be arbitrary functions of the vector $x(t) \in R^m$ of stochastic economic influence factors.

In order to keep the applications in the subsequent sections sufficiently simple, the drift terms are assumed to be affine in $x(t)$:

$$r(x(t)) = r_1^T x(t) + r_0, \tag{100}$$

$$\mu(x(t)) = \mu_1(t) + \mu_0, \tag{101}$$

with $r_1 \in R^m$, $r_0 \in R$, $\mu_1 \in R^{n\times m}$, and $\mu_0 \in R^n$. Furthermore, the volatility matrix $\Sigma_p^{1/2}$ is assumed to be constant.

Thus, the final form of the wealth dynamics is

$$\begin{aligned} dW(t) = \ & W(t)\left[r_1^T x(t) + r_0\right]dt + h(x(t),t)dt \\ & + W(t)u^T(t)\Big[\Big[\,\mu_1 x(t) + \mu_0 - e\Big[\,r_1^T x(t) + r_0\Big]\Big]dt + \Sigma_p^{1/2}\,dZ_p(t)\Big], \end{aligned} \tag{102}$$

$$W(t_0) = W_0 \ . \tag{103}$$

4.4 CRRA Problems

In the first problem of this section, a self-financing portfolio with unconstrained controls is considered, i.e., where unlimited borrowing and unlimited short selling are allowed. The goal is maximizing the power utility function which has a constant relative risk aversion coefficient. In the second problem, the first problem is reconsidered with constrained control variables.

Problem 1. For the self-financing portfolio

$$\begin{aligned} dW(t) = \ & W(t)\Big[r_1^T x(t) + r_0\Big]dt \\ & + W(t)u^T(t)\Big[\Big[\,\mu_1 x(t) + \mu_0 - e\Big[r_1^T x(t) + r_0\Big]\Big]dt + \Sigma_p^{1/2}\,dZ_p(t)\Big], \end{aligned} \tag{104}$$

$$W(t_0) = W_0, \tag{105}$$

with the stochastic economic influence factors

$$dx(t) = [Ax(t)+a]dt + \Sigma_q^{1/2} dZ_q(t), \tag{106}$$
$$x(t_0) = x_0\ , \tag{107}$$

where the Brownian motions Z_p and Z_q are correlated according to (93), find the unconstrained optimal asset allocation vector $u : [t_0, t_1] \to R^n$, such that the expected value of the power utility function

$$J = \frac{1}{\gamma}\, E[W^\gamma(t_1)] \tag{108}$$

is maximized for a chosen value $\gamma \in (-\infty, 1)$ with $\gamma \neq 0$.

Remark 2. In order to prevent potential confusion, please note that the state x, the increment dW of the normalized Brownian motion, the drift term f, and the volatility term g in Section 3.4 correspond to the following structured quantities in Problem 1:

$$x \longrightarrow \begin{bmatrix} W \\ x \end{bmatrix} \in R^{1+m}, \tag{109}$$
$$dW \longrightarrow \begin{bmatrix} dZ_p \\ dZ_q \end{bmatrix} \in R^{n+m}, \tag{110}$$
$$f \longrightarrow \begin{bmatrix} W\Big\{r_1^T x + r_0 + u^T\Big(\mu_1 x + \mu_0 - e\Big[r_1^T x + r_0\Big]\Big)\Big\} \\ Ax + a \end{bmatrix}, \tag{111}$$
$$g \longrightarrow \begin{bmatrix} W u^T\, \Sigma_p^{1/2} & 0 \\ 0 & \Sigma_x^{1/2} \end{bmatrix}. \tag{112}$$

Furthermore

$$K \longrightarrow \frac{1}{\gamma} W^\gamma, \tag{113}$$
$$L \longrightarrow 0, \tag{114}$$
$$\mathcal{J}_x \longrightarrow \begin{bmatrix} J_w & J_x \end{bmatrix} \in R^{1\times(1+m)}, \tag{115}$$
$$\mathcal{J}_{xx} \longrightarrow \begin{bmatrix} J_{ww} & J_{wx} \\ J_{wx}^T & J_{xx} \end{bmatrix} = \begin{bmatrix} J_{ww} & J_{wx} \\ \nabla_x J_w & J_{xx} \end{bmatrix} \in R^{(1+m)\times(1+m)}\ . \tag{116}$$

Finally, the Itô correction factor gg^T in the Hamilton-Jacobi-Bellman equation (60) turns into[5]

$$gg^T \longrightarrow \begin{bmatrix} W^2 u^T \Sigma_p u & W u^T\, \Sigma_p^{1/2}\, \rho \Sigma_q^{T/2} \\ W\, \Sigma_q^{1/2}\, \rho^T\, \Sigma_p^{T/2}\, u & \Sigma_q \end{bmatrix} \tag{117}$$

since dZ_p and dZ_q are correlated according to (93).

[5] Remember, this factor stems from the sum-of-squares property of a normalized Brownian motion: $g\, dW dW^T g^T = gg^T dt$. For a non-normalized Brownian motion, $dW dW^T = M dt$ with M such as in (93) for example.

Plugging (152), (153), (154), (155), (156), (157), (158), (159), and (160) into (60) yields the Hamilton-Jacobi-Bellman partial differential equation

$$\begin{aligned} -J_t = \max_u \Big\{ & J_w W \Big\{ r_1^T x + r_0 + u^T \Big(\mu_1 x + \mu_0 - e \Big[r_1^T x + r_0 \Big] \Big) \Big\} \\ & + J_x (Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\ & + \frac{1}{2} \operatorname{tr}[J_{xx} \Sigma_q] + W u^T \, \Sigma_p^{1/2} \, \rho \Sigma_q^{T/2} \, \nabla_x J_w \Big\} \end{aligned} \tag{118}$$

with the boundary condition

$$J(W, x, t_1) = \frac{1}{\gamma} W^\gamma(t_1) \; . \tag{119}$$

Provided, $J_{ww} < 0$, the unique maximizing control is

$$u = \; - \frac{1}{J_{ww} W} \; \Sigma_p^{-1} \Big\{ J_w \left(\mu_1 x + \mu_0 - e \left[r_1^T x + r_0 \right] \right) + \Sigma_p^{1/2} \, \rho \Sigma_q^{T/2} \; \nabla_x J_w \Big\} \; . \tag{120}$$

At first glance, the set of equations (118), (119), and (120) looks rather impressive, indeed. However, the authors have found the following successful separation ansatz for the cost-to-go function $J(W, x, t)$:

$$J(W, x, t) = \frac{1}{\gamma} W^\gamma \cdot \ell(x, t) \tag{121}$$

with

$$\ell(x, t) = \exp \left\{ c(t) + k^T(t) x + \frac{1}{2} x^T K(t) x \right\} \; . \tag{122}$$

Since the exponent in (122) must vanish for all $x \in R^n$ at the final time $t = t_1$ in order for $\ell(x, t_1) \equiv 1$ to be satisfied, the following boundary conditions are obtained immediately:

$$c(t_1) = 0 \in R, \tag{123}$$

$$k(t_1) = 0 \in R^n, \tag{124}$$

$$K(t_1) = 0 \in R^{n \times n} \; . \tag{125}$$

The objective functional defined in (121) and (122) has the following relevant partial derivatives:

$$J_t = \frac{1}{\gamma} W^\gamma \ell(x, t) \left[\dot{c}(t) + \dot{k}^T(t) x + \frac{1}{2} x^T \dot{K}(t) x \right], \tag{126}$$

$$J_w = W^{\gamma - 1} \ell(x, t), \tag{127}$$

$$J_x = \frac{1}{\gamma} W^\gamma \ell(x, t) \left[k^T(t) + x^T K(t) \right], \tag{128}$$

$$J_{ww} = (\gamma - 1) W^{\gamma - 2} \ell(x, t), \tag{129}$$

$$J_{wx} = W^{\gamma - 1} \ell(x, t) \left[k^T(t) + x^T K(t) \right], \tag{130}$$

$$J_{xx} = \frac{1}{\gamma} W^\gamma \ell(x, t) \big[[k(t) + K(t) x] \big[k^T(t) + x^T K(t) \big] + K(t) \big] \, . \tag{131}$$

Notice that $J_{ww} < 0$ for all of the admissible values of the risk aversion parameter $\gamma < 1$ (with $\gamma \neq 0$). Therefore, the optimal control u in (120) is indeed maximizing in (118).

Combining (120), (121), (122), (123), (124), (125), (126), (127), (128), (129), (130), and (131) yields the following affine state feedback control law:

$$u(x(t)) = \frac{1}{\gamma - 1}\,\Sigma_p^{-1}\ \left\{\left[\mu_1 - er_1^T + \Sigma_p^{1/2}\rho\Sigma_q^{T/2}\,K(t)\right]x(t) + \mu_0 - er_0 + \Sigma_p^{1/2}\rho\Sigma_q^{T/2}\,k(t)\right\}, \tag{132}$$

where the symmetric matrix $K(t) \in R^{n\times n}$ and the vector function $k(t) \in R^n$ remain to be found for $t \in [t_0, t_1]$.

The optimal control consists of a myopic part and a look-ahead part, the latter of which exploits the fact that the future increments dZ_p and dZ_q are correlated.

Plugging the optimal feedback control law (132), the ansatz (121) and (122) for the cost-to-go function W, and its derivatives (126), (127), (128), (129), (130), and (131) into the Hamilton-Jacobi-Bellman partial differential equation (118), results in a *very* long expression. However, all of the many terms are either quadratic in x, or linear in x, or scalars. Since $x \in R^n$ is an arbitrary vector argument, the differential equations for the unknown functions $c(\cdot)$, $k(\cdot)$, and $K(\cdot)$ can be obtained by comparing the coefficients in each of the three classes of terms, separately.

Rather tedious algebraic manipulations yield the following unilaterally coupled differential equations for$K(\cdot)$, $k(\cdot)$, and $c(\cdot)$, respectively:

$$-\dot{K}(t) = \mathcal{A}^T K(t) + K(t)\mathcal{A} - K(t)\mathcal{S}K(t) + \mathcal{Q}, \tag{133}$$

with

$$\mathcal{A} = A + \frac{\gamma}{1-\gamma}\,\Sigma_q^{1/2}\,\rho^T\,\Sigma_p^{-1/2}\left(\mu_1 - er_1^T\right), \tag{134}$$

$$\mathcal{S} = -\frac{\gamma}{1-\gamma}\,\Sigma_q^{1/2}\,\rho^T\rho\Sigma_q^{T/2} - \Sigma_q, \tag{135}$$

$$\mathcal{Q} = \frac{\gamma}{1-\gamma}\left(\mu_1 - er_1^T\right)^T\,\Sigma_p^{-1}\left(\mu_1 - er_1^T\right), \tag{136}$$

$$\begin{aligned}-\dot{k}(t) = &\left(\mathcal{A}^T + K(t)\left\{\Sigma_q + \frac{\gamma}{1-\gamma}\,\Sigma_q^{1/2}\,\rho^T\rho\Sigma_q^{T/2}\right\}\right)k(t)\\ &+ K(t)\left\{a + \frac{\gamma}{1-\gamma}\,\Sigma_q^{1/2}\,\rho^T\,\Sigma_p^{-1/2}(\mu_0 - er_0)\right\}\\ &+ \frac{\gamma}{1-\gamma}\left(\mu_1 - er_1^T\right)^T\Sigma_p^{-1}(\mu_0 - er_0) + \gamma r_1,\end{aligned} \tag{137}$$

$$
\begin{aligned}
-\dot{c}(t) = {} & \gamma\, r_0 + \frac{\gamma}{2(1-\gamma)} (\mu_0 - e r_0)^T\, \Sigma_p^{-1} (\mu_0 - e r_0) \\
& + k^T(t) \left\{ a + \frac{\gamma}{1-\gamma}\, \Sigma_q^{1/2}\, \rho^T\, \Sigma_p^{-1/2} (\mu_0 - e r_0) \right\} \\
& + \frac{1}{2}\, k^T(t) \left\{ \Sigma_q + \frac{\gamma}{1-\gamma}\, \Sigma_q^{1/2}\, \rho^T \rho \Sigma_q^{T/2} \right\} k(t) \\
& + \frac{1}{2}\, \mathrm{tr}\, [K(t) \Sigma_q] \ . \qquad (138)
\end{aligned}
$$

For the boundary conditions for $c(t_1)$, $k(t_1)$, and $K(t_1)$, see (123), (124), and (125).

Notice that in Problem 1, the differential equation (138), (123) for $c(\cdot)$ need not be solved, because the value $J(W, x, t)$ and its derivatives are not needed in the closed-form state feedback control law (132) and because the instantaneous value of the economic influence vector $x(t)$ can be measured at all times.

The differential equation (133) and (125) is of the type of the so-called matrix Riccati differential equation which is well known in control theory because it appears in the popular LQ regulator problem [43]. Here, the situation is a trifle more intricate. Suffice it to say that for $0 < \gamma < 1$, the symmetric matrix $K(t)$ will be positive-definite for $t < t_1$, whereas for $\gamma < 0$, $K(t)$ will be negatives-definite for $t < t_1$.

The summary of the analysis of Problem 1 is as follows:

Solution of Problem 1

The optimal CRRA investment strategy $u(t)$, $u_0(t)$ for $t \in [t_0, t_1]$ is given by the state feedback control law (132) and (97), where $K(t)$ and $k(t)$ are the solutions of the differential equations (133) and (137) with the boundary conditions (125) and (124), respectively, which can be computed off-line in advance.

Remark 3. The optimal control consists of two parts. The first part is myopic, i.e., independent of the remaining time horizon $[t, t_1]$:

$$
u_{\mathrm{myopic}}(x(t), t) = \frac{1}{\gamma - 1}\, \Sigma_p^{-1} \left\{ \left[\mu_1 - e r_1^T \right] x(t) + \mu_0 - e r_0 \right\} . \qquad (139)
$$

The second part is of the "look-ahead" type, which makes the best out of the fact that the future increments dZ_p and dZ_q will be correlated:

$$
u_{\mathrm{look-ahead}}(x(t), t) = \frac{1}{\gamma - 1}\, \Sigma_p^{-1}\, \Sigma_p^{1/2}\, \rho \Sigma_q^{T/2} \left[K(t) x(t) + k(t) \right] . \qquad (140)
$$

Its influence decreases as the time t approaches the final time t_1 due to the boundary conditions $K(t_1) = 0$ and $k(t_1) = 0$.

Problem 2. The statement of Problem 2 is identical to the statement of Problem 1, except for the additional control constraint

$$u(t) \in U \subset R^n \ , \tag{141}$$

where U is a closed, bounded, and convex subset of R^n.

In a typical example, where no borrowing of money and no short selling of risky investment opportunities are allowed, the constraint set U can be described as follows:

$$0 \le u_i(t) \le c_i \le 1 \quad \text{for } i = 1, \ldots, n, \tag{142}$$

$$\sum_{i=1}^{n} u_i(t) \le 1 \ . \tag{143}$$

Proceeding in the analysis of Problem 2 along the same lines as for Problem 1 leads to the following equations determining the optimal solution:

Solution of Problem 2

$$\begin{aligned} -J_t = \max_{u \in U} \ & \Big\{ J_w W \Big\{ r_1^T x + r_0 + u^T \Big(\mu_1 x + \mu_0 - e \Big[r_1^T x + r_0 \Big] \Big) \Big\} \\ & + J_x (Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\ & + \frac{1}{2} \operatorname{tr}[J_{xx} \Sigma_q] + W u^T \, \Sigma_p^{1/2} \, \rho \Sigma_q^{T/2} \, \nabla_x J_w \Big\} \end{aligned} \tag{144}$$

with the boundary condition

$$J(W, x, t_1) = \frac{1}{\gamma} W^{\gamma}(t_1) \ . \tag{145}$$

Unfortunately, in the restricted case with $U \neq R^n$, there is no analytical solution. Therefore, these equations have to be solved numerically for the cost-to-go function $J(W, x, t)$ and its derivatives, in order to find the optimal control

$$\begin{aligned} u(t) = \ \arg\max_{u \in U} & \Big\{ J_w W \Big\{ r_1^T x + r_0 + u^T \Big(\mu_1 x + \mu_0 - e \Big[r_1^T x + r_0 \Big] \Big) \Big\} \\ & + J_x (Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\ & + \frac{1}{2} \operatorname{tr}[J_{xx} \Sigma_q] + W u^T \, \Sigma_p^{1/2} \, \rho \Sigma_q^{T/2} \, \nabla_x J_w \Big\} \end{aligned} \tag{146}$$

and $u_0(t)$ according to (97) at any time t, where $W(t)$ and $x(t)$ will be the measured values of the instantaneous wealth and the vector of economic influence factors, respectively, at this time t.

Fortunately, solving the stochastic Hamilton-Jacobi-Bellman partial differential equation poses less numerical problems than solving its deterministic counterpart. For more details, consult [24, 38].

Remark 4. As has been noted in Section 1, in the uncorrelated case with $\rho = 0$ in (93), the dynamic stochastic optimal control problem degenerates to a static or myopic optimization problem. Mathematically, this should now become obvious by inspecting (118) and (120) in Problem 1 and (183) and (185) in Problem 2, respectively.

4.5 CARA Problems

As in Section 4.4, the problem of investing part of the wealth $W(t)$ in n risky assets and the balance of the wealth in a risk-free money market account is considered. However, the portfolio is no longer considered as self-financing. Rather, an inflow is allowed which may depend upon the vector $x(t)$ of economic influence factors. (Think of an entrepreneur or a pension fund [13, 14], for instance.) Furthermore, the return of the risk-free money market account is assumed to be independent of the economic influence factors (i.e., $r_1 = 0$ in (100)).

Here, the exponential utility function (81) is used. It has a constant coefficient of absolute risk aversion (85).

In Problem 3, unlimited borrowing and unlimited short selling are allowed. In Problem 4, the control variables constrained.

Problem 3. For the portfolio

$$\begin{aligned} dW(t) &= r_0 W(t)dt + [Hx(t)+h(t)]dt \\ &\quad + W(t)u^T(t)\left[[\mu_1 x(t)+\mu_0 - er_0]\,dt + \Sigma_p^{1/2}\,dZ_p(t)\right], \qquad (147) \\ W(t_0) &= W_0, \qquad (148) \end{aligned}$$

with the stochastic economic influence factors

$$\begin{aligned} dx(t) &= [Ax(t)+a]\,dt + \Sigma_q^{1/2}\,dZ_q(t), \qquad (149) \\ x(t_0) &= x_0\ , \qquad (150) \end{aligned}$$

where the Brownian motions Z_p and Z_q are correlated according to (93), find the unconstrained optimal asset allocation vector $u : [t_0, t_1] \to R^n$, such that the expected value of the exponential utility function

$$J = -\frac{1}{\gamma}\,E[e^{-\gamma W(t_1)}] \qquad (151)$$

is maximized for a chosen value $\gamma > 0$.

Remark 5. In order to prevent potential confusion, please note that the state x, the increment dW of the normalized Brownian motion, the drift term f, and the volatility term g in Section 3.4 correspond to the following structured quantities in Problem 3:

$$x \longrightarrow \begin{bmatrix} W \\ x \end{bmatrix} \in R^{1+m}, \tag{152}$$

$$dW \longrightarrow \begin{bmatrix} dZ_p \\ dZ_q \end{bmatrix} \in R^{n+m}, \tag{153}$$

$$f \longrightarrow \begin{bmatrix} W\left\{r_0 + u^T\left(\mu_1 x + \mu_0 - e r_0\right)\right\} + Hx + h \\ Ax + a \end{bmatrix}, \tag{154}$$

$$g \longrightarrow \begin{bmatrix} W u^T \Sigma_p^{1/2} & 0 \\ 0 & \Sigma_x^{1/2} \end{bmatrix}. \tag{155}$$

Furthermore

$$K \longrightarrow -\frac{1}{\gamma} e^{-\gamma W}, \tag{156}$$

$$L \longrightarrow 0, \tag{157}$$

$$\mathcal{J}_x \longrightarrow \begin{bmatrix} J_w & J_x \end{bmatrix} \in R^{1\times(1+m)}, \tag{158}$$

$$\mathcal{J}_{xx} \longrightarrow \begin{bmatrix} J_{ww} & J_{wx} \\ J_{wx}^T & J_{xx} \end{bmatrix} \begin{bmatrix} J_{ww} & J_{wx} \\ \nabla_x J_w & J_{xx} \end{bmatrix} \in R^{(1+m)\times(1+m)}. \tag{159}$$

Finally, the Itô correction factor gg^T in the Hamilton-Jacobi-Bellman equation (60) turns into

$$gg^T \longrightarrow \begin{bmatrix} W^2 u^T \Sigma_p u & W u^T \Sigma_p^{1/2} \rho \Sigma_q^{T/2} \\ W \Sigma_q^{1/2} \rho^T \Sigma_p^{T/2} u & \Sigma_q \end{bmatrix} \tag{160}$$

since dZ_p and dZ_q are correlated according to (93).

Plugging (152), (153), (154), (155), (156), (157), (158), (159), and (160) into (60) yields the Hamilton-Jacobi-Bellman partial differential equation

$$\begin{aligned} -J_t = \max_u \; & \left\{ J_w \left[W\left\{r_0 + u^T\left(\mu_1 x + \mu_0 - e r_0\right)\right\} + Hx + h \right] \right. \\ & + J_x(Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\ & \left. + \frac{1}{2} \operatorname{tr}[J_{xx}\Sigma_q] + W u^T \Sigma_p^{1/2} \rho \Sigma_q^{T/2} \nabla_x J_w \right\} \end{aligned} \tag{161}$$

with the boundary condition

$$J(W, x, t_1) = -\frac{1}{\gamma} e^{-\gamma W(t_1)}. \tag{162}$$

Provided, $J_{ww} < 0$, the unique maximizing control is

$$u = -\frac{1}{J_{ww} W} \Sigma_p^{-1} \left\{ J_w\left(\mu_1 x + \mu_0 - e r_0\right) + \Sigma_p^{1/2} \rho \Sigma_q^{T/2} \nabla_x J_w \right\}. \tag{163}$$

The following ansatz for the cost-to-go function $J(W,x,t)$ turns out to be successful here:

$$J(W,x,t) = -\frac{1}{\gamma}\exp\left\{c(t) + c_w(t)W + k^T(t)x + \frac{1}{2}x^T K(t)x\right\} \tag{164}$$

with the following obvious boundary conditions at the final time $t = t_1$:

$$c(t_1) = 0 \in R, \tag{165}$$
$$c_w(t_1) = -\gamma \in R, \tag{166}$$
$$k(t_1) = 0 \in R^n, \tag{167}$$
$$K(t_1) = 0 \in R^{n\times n}\ . \tag{168}$$

The objective functional defined in (164) has the following relevant partial derivatives:

$$J_t = J(W,x,t)\left[\dot{c}(t) + \dot{c}_w(t)W + \dot{k}^T(t)x + \frac{1}{2}x^T\dot{K}(t)x\right], \tag{169}$$
$$J_w = J(W,x,t)c_w(t), \tag{170}$$
$$J_x = J(W,x,t)\left[k^T(t) + x^T K(t)\right], \tag{171}$$
$$J_{ww} = J(W,x,t)c_w^2(t), \tag{172}$$
$$J_{wx} = J(W,x,t)c_w(t)\left[k^T(t) + x^T K(t)\right], \tag{173}$$
$$J_{xx} = J(W,x,t)\left[[k(t)+K(t)x]\left[k^T(t)+x^T K(t)\right] + K(t)\right]\ . \tag{174}$$

Notice that $J_{ww} < 0$ for all of the admissible values of the risk aversion parameter $\gamma > 0$, since the value of J is negative by definition. Therefore, the optimal control u in (163) is indeed maximizing in (161).

Combining (163), (164), (165), (166) (167), (168), (169), (170), (171), (172), (173), and (174) yields the following affine state feedback control law:

$$u(x(t)) = -\frac{1}{c_w W}\,\Sigma_p^{-1}\ \left\{\left[\mu_1 + \Sigma_p^{1/2}\,\rho\Sigma_q^{T/2}\,K(t)\right]x(t) + \mu_0 - er_0 + \Sigma_p^{1/2}\,\rho\Sigma_q^{T/2}\,k(t)\right\}, \tag{175}$$

where the symmetric matrix $K(t) \in R^{n\times n}$ and the vector function $k(t) \in R^n$ remain to be found for $t \in [t_0, t_1]$.

As in Problem 1, the optimal control consists of a myopic part and a look-ahead part, the latter of which exploits the fact that the future increments dZ_p and dZ_q are correlated. Notice, that the "courage" to invest into risky assets decreases with increasing wealth (CARA).

Plugging the optimal feedback control law (175), the ansatz (164) for the cost-to-go function W, and its derivatives (169), (170), (171), (172), (173), and (174) into the Hamilton-Jacobi-Bellman partial differential equation (161), results in a *very* long expression. However, all of the many terms are either

quadratic in x, or linear in x, or scalars. Since $x \in R^n$ is an arbitrary vector argument, the differential equations for the unknown functions $c(\cdot)$, $c_w(\cdot)$, $k(\cdot)$, and $K(\cdot)$ can be obtained by comparing the coefficients in each of the three classes of terms, separately.

Rather tedious algebraic manipulations yield the following unilaterally coupled differential equations for $K(\cdot)$, $k(\cdot)$, $c_w(\cdot)$, and $c(\cdot)$, respectively:

$$-\dot{K}(t) = \mathcal{A}^T K(t) + K(t)\mathcal{A} - K(t)\mathcal{S}K(t) + \mathcal{Q} \tag{176}$$

with

$$\mathcal{A} = A - \Sigma_q^{1/2}\,\rho^T\,\Sigma_p^{-1/2}\,\mu_1, \tag{177}$$

$$\mathcal{S} = \Sigma_q^{1/2}\,\rho^T\rho\Sigma_q^{T/2} - \Sigma_q, \tag{178}$$

$$\mathcal{Q} = -\,\mu_1^T\,\Sigma_p^{-1}\,\mu_1, \tag{179}$$

$$\begin{aligned} -\dot{k}(t) = {} & \big[\mathcal{A}^T - K(t)\mathcal{S}\big]k(t) + c_w H^T + K(t)a \\ & - \Big[\mu_1^T\,\Sigma_p^{-1} + K(t)\,\Sigma_q^{1/2}\,\rho^T\,\Sigma_p^{-1/2}\Big](\mu_o - er_o), \end{aligned} \tag{180}$$

$$-\dot{c}_w(t) = r_0 c_w(t), \tag{181}$$

$$\begin{aligned} -\dot{c}(t) = {} & hc_w(t) + a^T k(t) + \frac{1}{2}k^T(t)\Sigma_q k(t) + \frac{1}{2}\mathrm{tr}[K(t)\Sigma_q] \\ & - \frac{1}{2}(\mu_0 - er_0)^T\,\Sigma_p^{-1}(\mu_0 - er_0) - \frac{1}{2}k^T(t)\,\Sigma_q^{1/2}\,\rho\rho^T\,\Sigma_q^{T/2}\,k(t) \\ & - (\mu_0 - er_0)^T\,\Sigma_p^{-T/2}\,\rho\Sigma_q^{1/2}\,k(t)\ . \end{aligned} \tag{182}$$

For the boundary conditions for $c(t_1)$, $c_w(t_1)$, $k(t_1)$, and $K(t_1)$, see (165), (166), (167), and (168).

Notice that in Problem 3, the differential equations (181) and (182) for $c_w(\cdot)$ and $c(\cdot)$, respectively, need not be solved, because the value $J(W, x, t)$ and its derivatives are not needed in the closed-form state feedback control law (175) and because the instantaneous value of the economic influence vector $x(t)$ can be measured at all times.

The summary of the analysis of Problem 3 is as follows:

Solution of Problem 3

The optimal CARA investment strategy $u(t)$, $u_0(t)$ for $t \in [t_0, t_1]$ is given by the state feedback control law (175) and (97), where $K(t)$ and $k(t)$ are the solutions of the differential equations (176) and (180) with the boundary conditions (168) and (167), respectively, which can be computed off-line in advance.

Problem 4. The statement of Problem 4 is identical to the statement of Problem 3, except for the additional control constraint $u(t) \in U \subset R^n$, where U is a closed, bounded, and convex subset of R^n (see Problem 2).

Solution of Problem 4

$$
\begin{aligned}
-J_t = \max_{u \in U} \big\{ & J_w W \left\{ r_0 + u^T \left(\mu_1 x + \mu_0 - e r_0 \right) \right\} \\
& + J_x (Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\
& + \frac{1}{2} \operatorname{tr}[J_{xx} \Sigma_q] + W u^T \, \Sigma_p^{1/2} \, \rho \, \Sigma_q^{T/2} \, \nabla_x J_w \Big\}
\end{aligned} \tag{183}
$$

with the boundary condition

$$
J(W, x, t_1) = -\frac{1}{\gamma} e^{-\gamma W(t_1)} \; . \tag{184}
$$

Unfortunately, in the restricted case with $U \neq R^n$, there is no analytical solution. Therefore, these equations have to be solved numerically for the cost-to-go function $J(W, x, t)$ and its derivatives, in order to find the optimal control

$$
\begin{aligned}
u(t) = \arg \max_{u \in U} \big\{ & J_w W \left\{ r_0 + u^T \left(\mu_1 x + \mu_0 - e r_0 \right) \right\} \\
& + J_x (Ax + a) + \frac{1}{2} J_{ww} W^2 u^T \Sigma_p u \\
& + \frac{1}{2} \operatorname{tr}[J_{xx} \Sigma_q] + W u^T \, \Sigma_p^{1/2} \, \rho \Sigma_q^{T/2} \, \nabla_x J_w \Big\}
\end{aligned} \tag{185}
$$

and $u_0(t)$ according to (97) at any time t, where $W(t)$ and $x(t)$ will be the measured values of the instantaneous wealth and the vector of economic influence factors, respectively, at this time t.

5 Conclusions

In this chapter, it has been shown how the stochastic model-predictive optimal control theory can be used in order to solve problems of optimal asset allocation with sector rotation, under consideration of risk aversion. In the two types of problems (CRRA and CARA) with unlimited controls,[6] analytic feedback solutions of the continuous-time stochastic optimal control problems have been found. In the more realistic versions of the two problems with limited controls, the optimal feedback control must be found with numerical methods.

These methods were successfully tested in several exhaustive Monte Carlo simulation studies at the Measurement and Control Laboratory of ETH Zurich under the supervision of the authors.

[6] i.e., unlimited investing into an investment opportunity, unlimited short selling, and unlimited borrowing from the money market account

In the next phase, the validity of these methods was established in several studies using real data from reliable data banks (such as Bloomberg Finance) for the relevant market data *and* the relevant economic influence factors at SwissQuant Group AG.[7] This led to several proprietary software products of SwissQuant Group AG to be used in the ALM industry.

Areas for future research: Below some open research problems are sketched.

- In addition to the increments dZ of Brownian motions, allow for increments dQ of Poisson processes (creating jump discontinuities in the market data and/or the economic factors). This is relevant for modeling "crashes" (i.e., extraordinarily large changes within a single period of observation) of stock markets.
- Develop monitoring tools for safely detecting and possibly even predicting such extraordinary events.
- Generalize the presented model-predictive stochastic optimal control methods to adaptive control. The possibilities for adaptation include the following: dynamically changing the coefficient γ of risk aversion and/or dynamically changing the length T of the prediction interval and/or even temporarily switching from the CRRA strategy to the CARA strategy in the situation of such an event.
- Exploit the "cyclic nature" of economics in the modeling of the economic influence factors. In this case, the matrix A in (91) cannot be diagonal because it needs at least one pair of conjugate-complex eigenvalues.
- Generalize the presented stochastic optimal control methods to the problem of optimal stock picking. In this case, the economic influence factors used so far need to be complemented by several company-specific influence factors, including the quality of its management, its markets, and more common factors which are generally used in valuation [12].

References

1. Athans, M., Falb, P.L.: Optimal Control, McGraw-Hill, New York (1966)
2. Balduzzi, P., Lynch, A.: Transaction costs and predictability: Some utility cost calculations. J. Finan. Econ. 52, 47–78 (1999)
3. Bielecki, T.R., Pliska, S.R., Sherries, M.: Risk sensitive asset allocation. J. Econ. Dynam. Control 24, 1145–1177 (2000)
4. Breiman, L.: Probability. Addison-Wesley, Reading. (Reprinted 1992 by SIAM, Philadelphia.) (1968)
5. Brennan, M.J., Schwartz, E.S., Lagnado, R.: Strategic asset allocation. J. Econ. Dynam. Control 21, 1377–1403 (1997)

[7] SwissQuant Group AG is a spin-off company of ETH Zurich.

6. Brennan, M.J., Schwartz, E.S.: The use of treasury bill futures in strategic asset allocation programs. In: W.T. Ziemba and J.M. Mulvey (Eds.), Worldwide Asset and Liability Modelling, Chapter 10. Cambridge University Press, Cambridge, UK (1999)
7. Brennan, M.J., Yihong, X.: Stochastic interest rates and bond-stock mix. Europ. Finan. Rev. 4, 197–210 (2000)
8. Campbell, I.Y., Schiller, R.: The dividend-price ratio and expectations of future dividends and discount factors. Rev. Finan. Stud. 1, 195–228 (1988)
9. Campbell, I.Y., Schiller, R.: Yield spreads and interest rates: A bird's eye view. Rev. Econ. Stud. 58, 495–514 (1991)
10. Campbell, J.Y., Chacko, G., Rodriguez, J., Viceira, L.M.: Strategic asset allocation in a continuous-time var model. J. Econ. Dynam. Control 28, 2195–2214 (2003)
11. Canestrelli, E., Pontini, S.: Inquiries on the application of multidimensional processes to financial investments. Econ. Complexity 2, 44–62 (2000)
12. Copeland, T., Koller, T., Murrin, J.: Valuation: Measuring and Managing the Value of Companies (third edn.), Wiley, New York (2000)
13. Dondi, G.A.: Models and Dynamic Optimization for the Asset and Liability Management of Pension Funds. Dissertation Nr. 16257, ETH Zurich (2005)
14. Dondi, G.A., Herzog, F., Schumann, L.M., Geering, H.P.: Dynamic asset and liability management for Swiss pension funds. In: S.A. Zenios and Ziemba W.T. (Eds.) Handbook of Asset and Liability Management: Applications and Case Studies (Vol. 2, pp. 963–1028), Elsevier, Amsterdam (2007)
15. Doob, J.L.: Stochastic Processes, Wiley, New York (1953)
16. Fama, E., Schwert, G.: Asset returns and inflation. J. Finan. Econ. 5, 115–146 (1977)
17. Fama, E., French, K.: Dividend yields and expected stock returns. J. Finan. Econ. 22, 3–25 (1988)
18. Fama, E., French, K.: Business conditions and expected returns on the stocks and bonds. J. Finan. Econ. 25, 23–49 (1989)
19. Geering, H.P.: Optimal Control with Engineering Applications, Springer, Berlin (2007)
20. Glosten, L.R., Jaganathan, R., Runkle, D.E.: On the relation between the expected value and the volatility of nominal excess returns on stocks. J. Finan. 48, 1779–1802 (1993)
21. Halkin, H.: On the necessary conditions for optimal control of nonlinear systems. J. d'Analyse Math. 12, 1–82 (1964)
22. Haugh, M.B., Lo, A.W.: Asset allocation and derivatives. Quant. Finan. 1, 45–72 (2001)
23. Herzog, F.: Strategic Portfolio Management for Long-Term Investments: An Optimal Control Approach. Dissertation Nr. 16137, ETH Zurich (2005)
24. Herzog, F., Peyrl, H., Geering, H.P.: Proof of the convergence of the successive approximation algorithm for numerically solving the Hamilton-Jacobi-Bellman equation. WSEAS Trans. Sys. 4(12), 2238–2245 (2005)
25. Herzog, F., Dondi, G., Geering, H.P.: Stochastic model predictive control and portfolio optimization. Int. J. Theor. Appl. Finan. 10(2), 203–233 (2007)
26. Ilmanen, A.: Forecasting US bond returns. J. Fix. Income 7, 22–37 (1997)
27. Kalman, R.E., Falb, P.L., Arbib, M.: Topics in Mathematical System Theory, McGraw-Hill, New York (1969)

28. Kim, T.S., Omberg, E.: Dynamic nonmyopic portfolio behavior. Rev. Finan. Stud. 9, 141–161 (1996)
29. Korn, R., Kraft, H.: A stochastic control approach to portfolio problems with stochastic interest rates. SIAM J. Contr. Optim. 40, 1250–1269 (2001)
30. Lynch, A.: Portfolio choice and equity characteristics: Characterizing the hedging demands induced by return predictability. J. Finan. Econ. 62, 67–130 (2001)
31. Merton, R.C.: Lifetime portfolio selection under uncertainty: The continuous case. Rev. Econ. Stat. 51, 247–257 (1969)
32. Merton, R.C.: Optimum consumption and portfolio rules in a continuous-time model. J. Econ. Theory 3, 373–413 (1971)
33. Merton, R.C.: An intertemporal capital asset pricing model. Econometrica 41, 867–887 (1973)
34. Munk, C., Sørensen, C., Vinther, T.N.: Dynamic asset allocation under mean reverting returns, stochastic interest rates, and inflation uncertainty. Proceedings of the 30th Annual Meeting of the European Finance Association, Glasgow (2003)
35. Øksendal, B.: Stochastic Differential Equations (5th edn.), corrected second printing. Springer, New York (2000)
36. Patelis, A.D.: Stock return predictability and the role of monetary policy. J. Finan. 52, 1951–1972 (1997)
37. Pesaran, M.H., Timmermann, A.: Predictability of stock returns: Robustness and economic significance. J. Finan. 50, 1201–1228 (1995)
38. Peyrl, H., Herzog, F., Geering, H.P.: Numerical solution of the Hamilton-Jacobi-Bellman equation for stochastic optimal control problems. Proceedings of the 2005 WSEAS International Conference on Dynamical Systems and Control. 489–497, Venice (2005)
39. Pratt, J.W.: Risk aversion in the small and in the large. Econometrica 32, 122–136 (1964)
40. Samuelson, P.A.: Lifetime portfolio selection by dynamic stochastic programming. Rev. Econ. Stat. 51, 239–246 (1969)
41. Shen, P.: Market timing strategies that worked – Based on the e/p ratio of the S&P 500 and interest rates. J. Portf. Manag. 29, 57–68 (2003)
42. Steele, J.M.: Stochastic Calculus and Financial Applications, Springer, New York (2001)
43. Willems, J.C.: Least squares stationary optimal control and the algebraic Riccati equation. IEEE Trans. Automat. Contr. 16, 621–634 (1971)
44. Yong, J., Zhou, X.Y.: Stochastic Controls, Hamiltonian Systems, and HJB Equations, Springer, New York (1999)

Appendix A: Notation

In order to improve the readability of this chapter, some operator notation is collected in this appendix.

Linear Algebra

Transposing a matrix:

The transpose of the matrix $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$ is $A^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix}$.

In particular, the transpose of a column vector is a row vector and vice versa.

The operator diag*:*

The operator diag maps the vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ or its transpose $\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$ into the diagonal matrix $\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}$.

The trace operator:

The trace operator produces the sum of the diagonal terms of a square matrix:

$$\mathrm{tr} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11} + a_{22} + a_{33}.$$

For matrices A and B of suitable dimensions, the trace operator has the following property: $\mathrm{tr}(AB) = \mathrm{tr}(BA) = \mathrm{tr}(A^T B^T) = \mathrm{tr}(B^T A^T)$. In the special case of two column two-vectors a and b:

$$\mathrm{tr}(ab^T) = \mathrm{tr} \begin{bmatrix} a_1 b_1 & a_1 b_2 \\ a_2 b_1 & a_2 b_2 \end{bmatrix} = \mathrm{tr}(b^T a) = b^T a = a_1 b_1 + a_2 b_2 \,.$$

The square root of a symmetric, positive-definite matrix:

For a symmetric, n by n, positive-definite matrix Σ, its square root is an n by n matrix denoted by $\Sigma^{1/2}$ such that the relation $\Sigma = \Sigma^{1/2}\Sigma^{T/2}$ holds (where the second factor is the transpose of the first). The square root is not unique, unless it is required to be a symmetric matrix as well. Throughout this chapter, terms of the form $\Sigma^{1/2}$ appear in stochastic differential equations as volatility parameters.

Differential Calculus

The Jacobian:

The differentiable function $f : R^3 \to R^2$ has the following Jacobian matrix (of partial derivatives):

$$f_x = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \end{bmatrix}.$$

If the function f is scalar-valued, its Jacobian f_x is a row vector.

The gradient:

The differentiable function $f : R^3 \to R$ has the gradient

$$\nabla_x f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = f_x^T.$$

The Hessian:

The Hessian of a twice differentiable function $f : R^2 \to R$ is the symmetric matrix

$$J_{xx} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}.$$

Stochastics

The expected value of a random quantity x is denoted by $E[x]$.

Appendix B: Controllability

Consider the function $f : R^n \times R^m \times R \to R^n$ which is continuously differentiable with respect to its first and second arguments and piecewise continuous with respect to its last argument.

Definition 3. Controllability [1, 27]
The nonlinear dynamic system

$$\dot{x}(t) = f(x(t), u(t), t)$$

with the state vector $x(t) \in R^n$ and the control vector $u(t) \in R^m$ is completely controllable over the finite time interval $[t_0, t_1] \subset R$, if for arbitrary vectors $x_0 \in R^n$ and $x_1 \in R^n$, there exists a piecewise continuous control $u(., x_0, x_1) : [t_0, t_1] \to R^m$, such that the state vector x is transferred from the initial state

$$x(t_0) = x_0$$

to the final state

$$x(t_1) = x_1 \ .$$

Consider the special case $f(x, u, t) = A(t)x + B(t)u$.

Theorem 3. Controllability of a linear time – varying system [27].
The linear time-varying dynamic system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$

is completely controllable over the finite time interval $[t_0, t_1] \subset R$, if and only if the control Gramian matrix $W(t_0, t_1) \in R^{n\times n}$ is positive-definite:

$$W(t_0, t_1) = \int_{t_0}^{t_1} \Phi(t_1, t)B(t)B^{\mathrm{T}}(t)\Phi(t_1, \sigma)\, dt \;\succ\; 0\ .$$

Here, $\Phi(.,.) \in R^{n\times n}$ is the transition matrix of the dynamics matrix $A(\cdot)$ satisfying the differential equation

$$\frac{d}{dt}\Phi(t, \tau) = A(t)\Phi(t, \tau)$$

with the boundary condition

$$\Phi(\tau, \tau) = I$$

at an arbitrary time $\tau \in [t_0, t_1]$.

Corollary 1. Controllability of a linear time – invariant system [1], [27]
For constant matrices $A \in R^{n\times n}$ and $B \in R^{n\times m}$, we have

$$\mathrm{rank}\{W(t_0, t_1)\} = \mathrm{rank}\{[B, AB, A^2B, \ldots, A^{n-1}B]\}\ ,$$

i.e., the linear time-invariant system

$$\dot{x}(t) = Ax(t) + Bu(t)$$

is completely controllable over every finite time interval $[t_0, t_1]$ if and only if the controllability matrix $\mathcal{U} = [B, AB, A^2B, \ldots, A^{n-1}B] \in R^{n\times n\cdot m}$ has full rank n.

A Nonlinear Optimal Control Approach to Process Scheduling

Fabio D. Fagundez[1], João Lauro D. Facó[2], and Adilson E. Xavier[3]

[1] Graduate School of Engineering, Federal University of Rio de Janeiro, Brazil
fabio.fagundez@ufrj.br
[2] Department of Computer Science, Federal University of Rio de Janeiro, Brazil
jldfaco@ufrj.br
[3] Graduate School of Engineering, Federal University of Rio de Janeiro, Brazil
adilson@cos.ufrj.br

Summary. Scheduling problems in the process industry feature combinatorial and nonlinear aspects arising from task sequencing and product blending. In this chapter, we present an optimal control approach, recognizing that process scheduling problems can be modeled as dynamic systems, where flows are control variables and volumes and composition are state variables. This approach yields a nonlinear optimal control model with continuous state and control variables, bounded by lower and upper limits, avoiding the use of discrete variables. In this optimal control model, mixed-integer constraints are replaced by complementarity constraints. Moreover, we present a hybrid procedure which combines mixed-integer and nonlinear models. Numerical test instances are presented and solved by well-known optimization solvers.

Key words: optimal control, nonlinear programming, mixed-integer programming, scheduling

1 Introduction

Scheduling problems can be modeled as discrete optimization problems, as they feature two general types of constraints: discrete constraints and continuous constraints. The first group relates to enumerative or logical decisions like "choose source A to send cargo B to destination C at time t," whereas the second relates to more general limitations like "the maximum storage capacity of store A is 30,000 m^3." Constraints on discrete variables stand for assignment and sequencing decisions, and continuous equations model mass, volume, energy, or component balances.

Floudas and Lin [5] recently presented a survey on process scheduling, where they emphasize the importance of mixed-integer linear programming

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_19,

(MILP) in this field. The guarantee of global optimality is considered as the highlight of this approach. However, due to scheduling's NP-completeness, such models suffer from the curse of dimensionality (the number of variables is exponential to number of time periods), and MILP solving procedures reach unacceptable computational times to find a solution for a real-world problem. Therefore, the research community has been constantly working on formulations to reduce models' dimensions, particularly within nonuniform timediscretization frameworks (see [4] for a thorough discussion on this subject). Moreover, nonlinear phenomena are dealt with linear approximations, relaxed or removed from the models. In this chapter, a nonlinear programming (NLP) formulation based on continuous variables is proposed, trying to achieve reasonably small models and to converge to local optimal solutions in affordable computing time.

The main idea herein discussed is to reduce the problem's dimensions by avoiding discrete variables. The proposed formulation employs complementarity constraints to handle assignment and sequence decisions, applied on continuous variables. A NLP feasible point is equivalent to a MILP feasible point and vice versa. Therefore, a NLP local solution is equivalent to an integral MILP feasible point, defining an upper bound (if solving a minimization problem) on the correspondent MILP, which can improve the pruning in a branch-and-bound procedure. In fact, the NLP solution is a valid solution for the scheduling problem, and may be kept as the solution in a real-world situation or may be used as an incumbent solution for the MILP.

In addition to the complementarity approach, we can also consider scheduling as dynamic systems, where one action (decision) at a given instant impacts the future states of the system. Dynamic systems are classically made up of control variables (the decisions one can take), state variables (the system features one can measure), and state equations (how a state is affected by past states and decisions). In industries such as the oil and gas industry and the water and wastewater industry, control systems are built upon optimal control dynamic models, where one tries to maintain the system operating safely and efficiently. An optimal control problem features a highly separable Jacobian matrix of the constraints, with a block-diagonal structure, which may result in convergence with lower computational costs [3]. Common NLP solvers can take advantage of this particular structure, as a discrete optimal control problem is equivalent to a NLP problem [1]. In particular, a scheduling problem can be exactly described with this approach: the transfer operations are represented by control variables, while inventories are mapped to state variables. The state at a given time instant is computed from previous state and control variables, by means of the state equations.

In this chapter, we combine both nonlinear approaches: namely, complementarity and optimal control to avoid mixed-integer formulations. We use the scheduling of crude oil and derivatives in ports as an example for the proposed nonlinear optimal control model. The chapter is organized as follows: Section 2 discusses the crude oil problem and its models; Section 3 presents

some numerical examples; and Section 4 closes this work with our final remarks.

2 Crude Oil Scheduling Models

The scheduling of crude oil and derivatives in ports is the problem to determine (i) ship allocation within the port; (ii) transfer operations between ships, tanks, process units, and pipelines; (iii) sequence of pipeline parcels (end products and crude oil), in such a manner that an objective cost function is minimized and operational constraints are respected. It is a complex task, featuring nonlinear (due to crude blending) and combinatorial (due to assignment and sequencing) aspects.

The logistic system can be divided into three main subsystems (Fig. 1): port, distribution center, and refinery, all of them connected by pipelines [9]. It is also possible to consider a single system, when the port tanks are directly connected to refinery charging tanks [10, 12]. In addition to these three systems, one may possibly consider a fourth system: the tanker fleet, whose schedule updates the estimated times of arrival (ETA) for each ship. According to Shah [12], a reasonable approach is to solve the systems hierarchically. We follow this approach in this chapter, considering two systems: the port (tankers, jetties, tanks, and pipelines) and the refinery crude area (pipelines, tanks, and distillation crude unit). However, it is important to mention that the equations presented herein could be employed in other arrangements as well.

Portside tanks serve as a buffer to keep the pipelines in continuous operation, even when tankers are late. In general, a (refinery or portside) tank stores a certain class of crude (e.g., heavy oil tanks cannot store light oil). Ideally, a good schedule will use a small number of tanks, but it is important to notice that inventory costs are secondary when compared to the cost of not

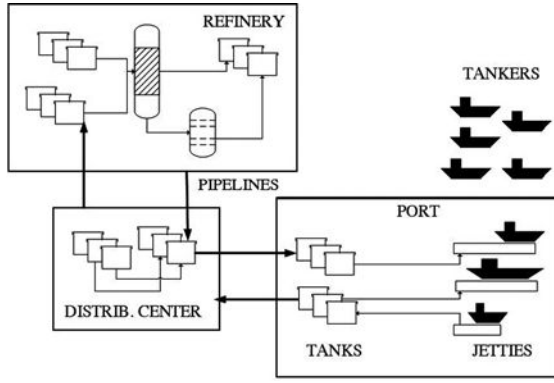


Fig. 1. Logistic subsystems

meeting the refinery production plan or delaying the ships. The refinery's demand for crude oil (as well as derivatives production) must be met by the port scheduling. Jetties can be restrictive on what tankers and cargoes to handle, to according to their dimensions (draught and length) and pumping capacity. A ship must berth, unload, and leave the port within a time window defined by contract, otherwise the oil company will pay heavy demurrage fees. For instance, Brazilian demurrage costs amounted to USD 1.5 billion in 2006 [2].

Therefore, the port schedule's main objective is to minimize demurrage costs, while keeping the refinery plan. A jetty is available for berthing only after the previous ship had enough time to leave the port. In the refinery side, the crude distillation unit operates continuously, around an operational feed flow. Blending is not allowed in the lines, i.e., each transfer operation has only one source equipment and one destination equipment at a given time. Running tanks are not allowed either, i.e., a tank cannot receive and send crude simultaneously. In fact, a tank can make a delivery to another equipment (e.g., pipeline or crude distillation unit) only if the necessary "idle time" has been observed (e.g., to separate brine from crude oil or to assure a lab analysis).

In the recent literature, Shah [12] proposed a MILP formulation for crude oil scheduling from tanker vessels to CDUs, based on two models: (i) a refinery problem (called the downstream problem) and (ii) a port model (called the upstream problem), constrained by the pipeline parcels defined by the solution of the first problem. Magalhães and Shah [10] revisited the problem, extending Shah's original MILP formulation to consider a real-world port–pipeline–refinery infrastructure and additional operational constraints. The authors pointed out that some optimal solutions of the MILP model, if applied to a real-world schedule, could be considered by a human scheduler as non-optimal, or even unfeasible, because certain real-world decisions are sometimes very hard to be mathematically modeled. Más and Pinto [9] modeled another real-world infrastructure, dividing the crude oil logistic system into three subsystems: (1) port, (2) distribution centers (intermediate storage), and (3) refineries. They also presented an exponential equation to calculate an upper bound of binary variables with the number of time intervals in order to illustrate that real-world instances are hard to be solved.

2.1 Modeling the Transfer Operation

The fundamental scheduling activity is the transfer operation, which is made up of a pair of equipments (source–destination) connected by an arc and a flow from the source to the destination. The control vector $u(t_i)$ is the vector where each entry $u_j(t_i)$ stands for a nonnegative flow on arc j at time t_i. The optimization problem is to define a feasible sequence of $u(t_i)$, for all instants t_i, which minimizes the objective function J. All control variables $u_j(t_i)$ are bounded.

The infrastructure of a logistic system can be seen as a graph, defined during the problem's formulation, featuring equipments as nodes, connected by flow arcs. A system graph is built by checking which equipments are connected by pump lines and which are compatible in terms of crude oil and physical dimensions. Figure 2 illustrates a port infrastructure with three jetties, five tanks (three for end products, two for crude oils), two pipelines connecting the port to a refinery (one to receive end products, the other to send crude oil), and three tankers that must be scheduled. In this example, tanker N3 can berth on jetties P3 and P2, but cannot berth on jetty P1. Moreover, N3's cargo is a crude oil that can be pumped to tank T5. Therefore, there is a flow arc (represented by the lower traced line) between tanker N3 and tank T5, through jetty P3.

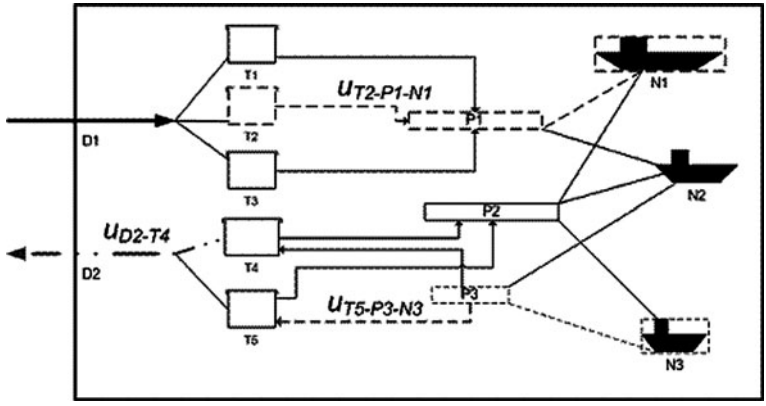


Fig. 2. System as a graph

In summary, the schedule is basically to define a nonnegative flow $u_j(t_i)$ for each arc j at each time instant t_i. If $u_j(t_i) = 0$, there is no transfer operation at arc j at time t_i, otherwise, a transfer operation is occurring at this arc.

Operational constraints, such as "one equipment N cannot be the destination of two transfer operations at the same time t_i, in order to avoid inline blending" can be modeled by different manners. For instance, let us examine the case of a certain tank N which is being fed by other equipments: it can be the destination of at most one transfer operation at a certain time t_i, as inline blending is forbidden. We present two different modeling possibilities (Table 1), with A_N as the set of indexes for all arcs whose destination is N:

(a) an MILP formulation, as commonly found in the literature [9, 12] and
(b) the novel NLP formulation proposed here.

The volume of tank N is calculated by a volumetric balance equation, which is equal in both models. The models differ on how to enforce the upper bounds on the flows and how to guarantee that only one source will be

Table 1. Modeling possibilities for transfer operations

Model	Equations
(a) MILP	$\text{vol}_N(t_i) = \text{vol}_N(t_{i-1}) + \sum_{j \in A_N} u_j(t_i)\Delta t$
	$\sum_{j \in A_N} b_j(t_i) <= 1$
	$0 <= u_j(t_i) <= b_j(t_i) * u_j^{\text{MAX}}(t_i), j \in A_N$
	$b_j(t_i)$ is binary, $u_j(t_i) \in R$
(b) NLP	$\text{vol}_N(t_i) = \text{vol}_N(t_{i-1}) + \sum_{j \in A_N} u_j(t_i)\Delta t$
	$\sum_{j \in A_N} \sum_{k > j \in A_N} u_j(t_i) * u_k(t_i) = 0$
	$0 <= u_j(t_i) <= u_j^{\text{MAX}}(t_i), j \in A_N$
	$u_j(t_i) \in R$

employed to feed the tank. Notice that, as N can participate in at most one transfer operation at time t_i, either all flows in the A_N arcs are zero at t_i (no transfer happens with destination N at time t_i) or only one flow is greater than zero at t_i (N is the destination of only one transfer operation at time t_i). Both formulations enforce this behavior.

Model (a) requires an additional control vector $b(t_i)$ of binary variables, where entry $b_j(t_i)$ is associated with the $u_j(t_i)$ entry. If $b_j(t_i)$ is set to 1, then a positive flow is allowed on arc j; otherwise (if set to zero), no flow is allowed on arc j. This is assured by the manipulation of the bounds on $u_j(t_i)$: if $b_j(t_i) = 1$, the bounds are preserved; otherwise, they are set to zero. The summation constraint on the binary constraints guarantees that at most one $b_j(t_i)$ can be evaluated as 1 at t_i, $j \in A_N$. All other binary variables associated with A_N must be evaluated as zero.

Model (b) relies on the control vector $u(t_i)$ only. There is no need for additional binary variables. The summation of the products of all A_N flows two by two is equal to zero if and only if all flows are equal to zero or only one flow is greater than zero, making N as the destination of at most one transfer operation, as required. The main disadvantage of this formulation is that it defines a nonconvex model. In the next sections, "idle time" and "berthing" constraints are formulated in a similar fashion.

2.2 Optimal Control Nonlinear Model

In the previous section, we have implictly defined the schedule as a dynamic model within the optimal control framework. An optimal control problem is defined mathematically as the following mathematical programming problem.

$$
\begin{aligned}
\min \quad & J = f(u, x, t) \\
\text{s.t.} \quad & u^{\text{MIN}} <= u(t_i) <= u^{\text{MAX}}, \quad t_0 <= t_i < t_F \\
& x^{\text{MIN}} <= x(t_i) <= x^{\text{MAX}}, \quad t_0 <= t_i <= t_F \\
& x(t_i) = g(x(t_{i-1}), u(t_{i-1})), \quad t_0 < t_i <= t_F \\
& x \in R^n, \quad u \in R^m.
\end{aligned}
$$

The control variables are u, the state variables are x, and the time horizon spreads from t_0 to t_F.

The transfer operation is composed of two equipments–source and destination–and a flow u_j from the source to the destination (through arc j). A storage is filled by flows from other equipments, changing its state (volume and composition), and this equipment may later perform an outlet transfer operation performing changes on other equipments. We modeled flow rates as control variables, bounded by upper and lower bounds: each entry of the control vector $u(t_i)$ stands for a flow in an arc between two equipments during the interval $[t_i, t_i + \Delta t]$. The control's upper bounds may not be the same for all intervals, depending on port unavailability because of tides, limited operation of some equipments during certain periods of the day, scheduled maintenance, and the fact that a ship can only berth after its ETA (estimated time of arrival). However, it is important to note that all flow bounds are known a priori during the problem formulation phase.

The proposed nonlinear optimal control model features the flow rates u as control variables (2), and volumes v and qualities p as subsets of the state variables x (3), all bounded by upper and lower limits. The state equations are developed from volume balance (4) and product blending in storage equipments (5). The objective function (1) is a summation of different costs, which can be prioritized with the use of weights (w_{cost})

$$\min \quad J = \sum_{\text{cost}} w_{\text{cost}} * C_{\text{cost}}, \tag{1}$$

$$\text{s.t.} \quad u_{\min} <= u(t_i) <= u_{\max}(t_i), \tag{2}$$

$$x_{\min} <= x(t_i) = [v(t_i)p(t_i)]^T <= x_{\max}, \tag{3}$$

$$v(t_i) = v(t_{i-1}) + Uu(t_{i-1})\Delta t \tag{4}$$

$$p_{N,q}(t_i) = (v_N(t_{i-1})^* p_{N,q}(t_{i-1}) + \sum_{j \in A_N} u_j(t_{i-1}) * p_{j,q}(t_{i-1}))/v_N(t_i) : \tag{5}$$

Equation (4) features U as an incidence square matrix with entries in 0, 1, −1. Equation (5) calculates density, sulfur concentration, and product composition: each $u_j(t_{i-1})$ is an inlet flow at N at time t_{i-1} and $p_{j,q}(t_{i-1})$ is the value of this inlet flow's property q. The complementarity equations will force that at most one source equipment is actually feeding N, i.e., at most one $u_j(t_{i-1})$ will be greater than zero.

The following equations model scheduling decisions: unique definition of source and destination in a transfer operation (6), idle time to segregate impurities (7), berthing time (8), and constant flow constraints (9), and all of them must equal to zero. We define new state variables r_N, z_N, s_N, and q_N for all equipments N, each one referring to a complementarity equation. In the next section, these variables will be employed to relax and penalize the problem.

$$r_N(t_i) = \sum_{j \in A_N} \sum_{k > j \in A_N} u_j(t_{i-1}) u_k(t_{i-1}) = 0, \tag{6}$$

$$z_N(t_i) = \sum_{t'=t_{i-1}-\Delta t_N^{\mathrm{IDLE}}}^{t_{i-1}} \sum_{j \in A_N} \sum_{k \in A_N^{\mathrm{OUTLET}}} u_j(t') u_k(t_{i-1}) = 0, \tag{7}$$

$$s_N(t_i) = \sum_{t'=t_{i-1}-\Delta t_N^{\mathrm{BERTH}}}^{t_{i-1}} \sum_{j \in A_N} \sum_{k \in A_{K<>N}} u_j(t') u_k(t_{i-1}) = 0, \tag{8}$$

$$q_N(t_i) = u_{N,0} - \sum_{j \in A_N} u_j(t_i) = 0. \tag{9}$$

Equation (6) enforces that only one flow can feed N at time t_{i-1}, therefore, any transfer operation has only one source and only one destination at time t_{i-1}. Equation (7) enforces that N will be able to feed another equipment only after its idle time $\Delta t_N^{\mathrm{IDLE}}$ was respected. Equation (8) enforces the necessary berthing time $\Delta t_N^{\mathrm{BERTH}}$ for ships. Equation (9) forces a constant flow $u_{N,0}$ feeding a given equipment (usually a process unit or a pipeline) N–this constraint can be easily changed to force a variable flow, if needed.

In the case of crude oil scheduling, we considered the following costs: demurrage (10–12), unattained demand (13), and inventory (14):

$$C_{\mathrm{demurrage}}^{\mathrm{unload}} = \sum_{N \in \mathrm{Ships}^{\mathrm{unload}}} \sum_{t_i > t_N^{\mathrm{depart}}} c_N^{\mathrm{demur}} v_N(t_i), \tag{10}$$

$$C_{\mathrm{demurrage}}^{\mathrm{load}} = \sum_{N \in \mathrm{Ships}^{\mathrm{load}}} \sum_{t_i > t_N^{\mathrm{depart}}} c_N^{\mathrm{demur}} (\mathrm{Cargo}_N - v_N(t_i)), \tag{11}$$

$$C_{\mathrm{demurrage}} = C_{\mathrm{demurrage}}^{\mathrm{load}} + C_{\mathrm{demurrage}}^{\mathrm{unload}}, \tag{12}$$

$$C_{\mathrm{demand}} = \sum_{t_i} \sum_{N \in \mathrm{Pipelines}} \sum_{P \in \mathrm{Products}} c_N^{\mathrm{demand}} v_{N,P}(t_i), \tag{13}$$

$$C_{\mathrm{inventory}} = \sum_{t_i} \sum_{N \in \mathrm{Storages}} c_N^{\mathrm{inv}} v_N(t_i). \tag{14}$$

Equations (10), (11), and (12) deal with demurrage cost: here we do not employ the classic demurrage formulation, but one that is also proportional to the remaining volume to be transferred that is delayed. Notice that demurrage costs are accounted for ship N only if it has departured after the maximum acceptable time of departure (t_N^{depart}). If ship N is to be loaded, its volume at the departure time must be Cargo_N. If N is to be unloaded, its volume at the departure time must be zero.

In all cost equations, c^{cost} is a different arbitrary unitary cost.

2.3 Solving the Problem

The model is solved as follows: the nonlinear constraints 6, 7, and 8 are relaxed and added to the objective function as penalties, creating the merit function J' (15). This merit function will be minimized instead of the original objective function J. This approach removes most of the difficult constraints, generating a broader search region for nonlinear optimization methods, with fewer constraints. In addition, if only linear state equations are present, the search region becomes a polyhedra. Within the feasible region of the original formulation, all penalties are cancelled. The parameter μ can be determined iteratively by solving successive relaxations of the original problems or fixed a priori as a large enough number.

$$J' = J + \mu \sum_t e^T r(t) + e^T s(t) + e^T z(t), \tag{15}$$

where e is the unitary vector.

Trivial points are points where the control vector u is zero for all time intervals. These points are very easy to be constructed, but they are not feasible for the original problem formulation. At trivial points, the demurrage and demand costs are maximal. The norms of the additional states are $||z|| = ||r|| = ||s|| = 0$ and $||q|| >> 0$. However, such points are feasible in the relaxed formulation and define a descent direction that leads to the minimization of the penalties–moving the points to the feasible region. Therefore, we use these points as starting points.

2.4 Mixed-Integer Linear Model

In order to compare the NLP approach with the MILP approach, we present a MILP model following the crude oil scheduling literature [9, 12]. Equations (2), (6), (7), and (8) from the NLP model are replaced by (16, 17, 18, 19). The blending equation (5) is dropped, as it is nonlinear. Complementarity constraints are replaced by mixed-integer constraints, with the addition of the binary variables vector (b_j):

$$0 <= u_N(t_i) <= \mathrm{Diag}(b_{j \in A_N}(t_i)) u_N^{\mathrm{MAX}}(t_i), \tag{16}$$

$$\sum_{j \in A_N} b_j(t_i) <= 1, \tag{17}$$

$$\sum_{j \in A_N^{\mathrm{OUTLET}}} b_j(t_i) + \sum_{t'=t_i-\Delta t^{\mathrm{BERTH}}}^{t_i} \sum_{j \in A_N} b_j(t') <= 1, N \in \mathrm{Ships}, \tag{18}$$

$$\sum_{j \in A_N^{\mathrm{OUTLET}}} b_j(t_i) + \sum_{t'=t_i-\Delta t^{\mathrm{IDLE}}}^{t_i} \sum_{j \in A_N} b_j(t') <= 1, N \in \mathrm{Storages}. \tag{19}$$

Equation (16) features a diagonal matrix Diag($b_j(t_i)$), composed of the binary variables b_j, which are added to the model in the MILP formulation. These variables represent scheduling decisions: there is no flow u_j at time t_i if $b_j = 0$ at time t_i, and there is a flow u_j if $b_j = 1$. The consecutive equations represent the following constraints: only one flow can be used by an equipment N at time t_{i-1}, idle time and berthing time must be respected before any outlet transfer.

The MILP can be solved with usual mixed-integer procedures and is larger than the NLP model, as shown in the next section.

3 Results

The NLP and MILP models were compared in five preliminary test instances (Tables 2 and 3), coded in AMPL [6], and solved with standard commercial solvers: CPLEX (v. 10.1.0) [8], SNOPT (v. 6.1) [7], and MINOS (v. 5.5) [11]. Case 1 is composed of an infrastructure with two crude tanks and one pipeline connected to a refinery, whose crude demand has to be fulfilled. Case 1 has two configurations: (A) allows the pipeline to be idle in certain periods and (B) keeps the pipeline with a constant flow during the entire schedule. The MINOS run converged to a local minimum in (B) configuration. Case 2 has two crude tanks, one jetty, and two tankers, whose cargo had to be unloaded. Case 3 has three crude tanks, one jetty three tankers, whose cargo had to be unloaded, and one pipeline, whose demand has to be fulfilled. Case 3 has two configurations: (A) allows the pipeline to be idle in certain periods and (B) keeps the pipeline with a constant flow. The SNOPT run converged to a local minimum with demurrag costs in (B) configuration. The number of variables is shown as determined after AMPL's presolve procedure. As both MILP and NLP models have linear objective functions, it is possible to compare them in regard to the global optimality of their solutions. All cases were solved in a workstation with the following configuration: Intel Core Duo T2250 1.73 GHz, RAM 1 GB, Linux OpenSUSE 10.1. The running times were around 1 s.

As the complementarity model is nonconvex, a nonlinear programming method, such as MINOS and SNOPT, may converge to local optima, differently from what happens with the mixed-integer model when solved by a typical branch-and-bound method, such as CPLEX. On the other hand, the complementarity model is more compact, featuring less variables and constraints than the MILP one. Noticing that one NLP solution is equivalent to an MILP feasible point, we propose a hybrid scheme: Solve the continuous NLP problem and then transform its solution into an initial point for the MILP. If needed, call NLP runs at some difficult nodes of the MILP branch-and-bound tree. This scheme may be able to reduce the total number of branches and simplex iterations in the MILP optimization, as the NLP point is an integral MILP good solution. At the current state of our research, we employed the NLP solutions to initialize the MILP previous examples and compared

Table 2. Dimensions of the test cases

Model	Case	Binary variables	Continuous variables	Constraints
NLP	1(A)	0	31	25
	1(B)	0	31	30
	2	0	111	87
	3(A)	0	169	103
	3(B)	0	135	97
MILP	1(A)	12	25	31
	1(B)	12	25	36
	2	34	82	111
	3(A)	93	158	265
	3(B)	93	158	275

Table 3. Results

Model	Case	Solution	Iterations	Global optimum
NLP (SNOPT)	1(A)	1460	51	Yes
	1(B)	1600	13	Yes
	2	0.33	12	Yes
	3(A)	0	812	Yes
	3(B)	18.27	544	No
NLP (MINOS)	1(A)	1460	17	Yes
	1(B)	1625	5	No
	2	0.33	191	Yes
	3(A)	0	411	Yes
	3(B)	0	472	Yes
MILP (CPLEX)	1(A)	1460	14	Yes
	1(B)	1600	13	Yes
	2	0.33	63	Yes
	3(A)	0	324 (8 BB nodes)	Yes
	3(B)	0	397 (25 BB nodes)	Yes

the number of iterations and branched nodes. A substantial reduction in the number of iterations in the MILP optimization run is detected (Table 4). All cases had similar CPU times of approximately 1 s.

Table 4 shows the MILP iterations when the NLP solutions were employed as initial incumbent solutions to the MILP formulation. A solution computed by the NLP formulation is transformed into a MILP point by simply adding the binary variables and replacing the complementarity constraints by the mixed-integer ones. For each positive flow, the corresponding binary variable is set to 1 (one), while for each null flow, the corresponding binary variable is set to 0 (zero). The number of branch-and-bound iterations and visited nodes is significantly reduced, even for these preliminary test cases.

Table 4. MILP results with different initializations

Case	Initial point (x, u)	Iterations
1(A)	$(x^0, 0)$	14
	$(x, u)^{\text{SNOPT}}$	13
	$(x, u)^{\text{MINOS}}$	13
1(B)	$(x^0, 0)$	13
	$(x, u)^{\text{SNOPT}}$	4
	$(x, u)^{\text{MINOS}}$	4
2	$(x^0, 0)$	63
	$(x, u)^{\text{SNOPT}}$	47
	$(x, u)^{\text{MINOS}}$	40
3(A)	$(x^0, 0)$	324 (8 BB nodes)
	$(x, u)^{\text{SNOPT}}$	215
	$(x, u)^{\text{MINOS}}$	215
3(B)	$(x^0, 0)$	397 (25 BB nodes)
	$(x, u)^{\text{SNOPT}}$	265 (6 BB nodes)
	$(x, u)^{\text{MINOS}}$	215

4 Conclusion

A nonlinear optimal control model for process scheduling–based on flow variables–was introduced. This way all constraints are modeled without discrete variables, achieving continuous models that are smaller than their MILP counterparts. Although being able to generate efficient solutions, the NLP formulation is nonconvex in general. The NLP can be employed as an auxiliary problem to traditional MILP formulations. In fact preliminary numerical results showed a significative reduction of MILP iterations when initialized by a NLP solution.

References

1. Abadie, J.: Application of the GRG Algorithm to Optimal Control Problems. In: J. Abadie (Ed.), Integer and Nonlinear Programming, North-Holland, Amsterdam (1970)
2. Collyer, W.: Sobreestadia de navios: A regra once on demurrage, always on demurrage. Jus Navigandi, 1166. Web: http://jus2.uol.com.br/ doutrina/texto.asp? id=8889 (in Portuguese) (Last access on July, 2008) (2006)
3. Facó, J.L.D.: A Generalized Reduced Gradient Algorithm for Solving Large-scale Discrete-Time Nonlinear Optimal Control Problems. In: H.B. Siguerdidjane, P. Bernhard (Ed.), Control Applications of Nonlinear programming and Optimization, Pergamon, Oxford (1990)

4. Floudas, C.A., Lin, X.: Continuous-time versus discrete-time approaches for scheduling of chemical processes: a review. Comput. Chem. Eng. 28, 2109–2129 (2004)
5. Floudas, C.A., Lin, X.: Mixed integer linear programming in process scheduling: modeling, algorithms, and applications. Ann. Oper. Res. 139, 131–162 (2005)
6. Fourer, R., Gay, D.M., Kernighan, B.W.: AMPL: A Modeling Language for Mathematical Programming, Duxbury, Belmont, CA (2003)
7. Gill, P.E., Murray, W., Saunders, M.A.: SNOPT: An SQP algorithm for large-scale constrained optimization. SIAM J. Optim. 12(4), 979–1006 (2002)
8. ILOG AMPL CPLEX System Version 11.0 User's Guide, ILOG Inc. (2008) http://www.netlib.org/ampl/solvers/cplex/ampl110.pdf
9. Más, R., Pinto, J.M.: A mixed-integer optimization strategy for oil supply in distribution complexes. Optim. Eng. 4(1), 23–64 (2003)
10. Magalhães, M.V., Shah, N.: Crude oil scheduling. *Proceedings of the 4th Conference on Foundations of Computer-Aided Process Operations*, Coral Springs, Florida, 323–326 (2003)
11. Murtagh, B.A., Saunders, M.A.: A projected Lagrangian algorithm and its implementation for sparse non-linear constraints. Math. Program. Stud. 16, 84–117 (1982)
12. Shah, N.: Mathematical Programming Techniques for Crude Oil Scheduling. Comput. Chem. Eng. 20 Suppl. S1227–S1232 (1996)

Hadamard's Matrices, Grothendieck's Constant, and Root Two

Dominique Fortin

Inria, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France
dominique.fortin@inria.fr

Summary. In this chapter, we start by a non-cooperative quantum game model for multiknapsack to give a flavor of quantum computing strength. Then, we show that many rank-deficient correlation matrices have Grothendieck's constant that goes beyond $\sqrt{2}$ for sufficiently large size. It suggests that cooperative quantum games relate powerset entanglement with Grothendieck's constant.

Key words: non-cooperative quantum game, multiknapsack, entanglement, grothendieck's constant

1 Introduction

For a long time, quantum computing has been concerning physicists and the experimental refutation of Bell's inequalities [1, 10, 38]. Since the polynomial time prime factorization on a *quantum computer* [36], it intrudes on many different fields in applied mathematics and *inspires* heuristics to tackle real-life applications in optimization. It is widely believed that quantum computing breakthrough comes from *entanglement*: in classical computing, states are well separated while in quantum computing all the states are combined at the same time in a possibly non-separable way (entanglement). However, the objectives widely differ among fields, e.g., theoretical physicists aim at a taxonomy of states when they try to *distillate* pure entangled states, while combinatorists use q-analogue as a tool to prove many old and new identities [25]. Here, we narrow the scope back to Bell's inequalities violation and entanglement modeling in combinatorial optimization problems.

In Section 2 we recall quantum issues on entanglement and narrow the focus to amplitude amplification that promises faster results by *quantum computer* offspring, as well as to the relationship between classical and quantum correlation matrices. In Section 3, we address a multiplayer quantum game model for the multiknapsack problem and show its limitations in sections that

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_20,

follow. In Section 6, we revisit the scaling factor between classical and quantum correlation matrices for Hadamard's matrices and extend the question to almost Hadamard's matrices.

Even under these very strong restrictions, we will see that difficulties remain tough.

2 Quantum Background

2.1 Computational State Space

Definition 1. *A binary quantum digit or qubit is a binary quantum system over the Hilbert space whose canonical basis is denoted* $\{|\,0\rangle, |\,1\rangle\} \equiv \left\{ e_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, e_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$.

This definition extends to ternary qutrit, ... and v-ary quvit quantum digits

Definition 2. *A v-valued quantum digit or quvit is a v-adic quantum system over the Hilbert space whose canonical basis is denoted* $\{|0\rangle, |1\rangle, \ldots, |v\rangle\} \equiv \{e_0, e_1, \ldots, e_v\}$.

Definition 3. *A quantum computational state is a complex-valued combination of quantum basis* $\{|\psi\rangle = \sum \alpha_i \, |i\rangle$ *normalized under* $\sum |\,\alpha_i\,|^2 = 1$.

It stands for a probabilistic combination of basic states where $p_i = |\,\alpha_i\,|^2$ is the probability to *observe* state $|i\rangle$.

Definition 4. *A quantum register of size n is an array of n quvits which can be in any of the individual states of its quvits at any instant or at all of the states (in probabilistic sense) at the same time, e.g.,* $r = \{|1032\rangle = |1\rangle \otimes |0\rangle \otimes |3\rangle \otimes |2\rangle$*, where quantum notations actually shrink notations from* $\mathbf{R}^{v \times n}$ *to a string of length n along with separators.*

It is customary in quantum usage to introduce notations $\langle\psi\,| = |\,\psi\rangle^\dagger$ for transposed conjugate, $\langle\phi\,|\,\psi\rangle = \langle|\phi\rangle, |\psi\rangle\rangle = |\,\psi\rangle^\dagger \;|\,\phi\rangle$ for inner product, or $\langle\phi|\; A \,|\psi\rangle = \langle|\phi\rangle, A\,|\psi\rangle\rangle$ for quadratic form.

2.2 Entanglement

Computational state above extends to computational register state as a combination $|r\rangle = \sum \alpha_{i_1 \ldots i_n} \, |x_1 \ldots x_n\rangle$, and according to the principles of quantum mechanics, the register state is in either *separable* or *entangled* state.

Definition 5 (entanglement vs separability). *A quantum register state is separable if* $\sum \alpha_{i_1 \ldots i_n} \, |x_1 \ldots x_n\rangle = \sum \beta_{i_1} \, |x_1\rangle \otimes \ldots \sum \beta_{i_n} \, |x_n\rangle$*, otherwise, the state is entangled.*

For instance, the ternary state $|02\rangle + |10\rangle = (0,0,1,1,0,0,0,0,0)^t$ is entangled while $\sum_{i=0}^{2}\sum_{j=0}^{2} |ij\rangle = \sum_{i=0}^{2} |i\rangle \otimes \sum_{j=0}^{2} |j\rangle$ is separable. Maximally (unnormalized) entangled state $|0\rangle + |1\rangle$ is related to Hadamard's matrix $H = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ by $|0\rangle + |1\rangle = H\,|0\rangle$. In fact, quantum register computing acts on a vector of complex variables $x_1 \ldots x_n$ so that register states are better seen as labeled structures where a computation is nothing but a generating function where the probability to observe a given register state is associated with the coefficient of $\frac{x_1^{i_1}\ldots x_n^{i_n}}{i_1!\ldots i_n!}$ in the expansion of the generating function. Notice that labeled structures impose an exponential generating function since we are interested in counting states under relabeling. Another striking feature of the exponential generating function viewpoint lies in the non-homogeneous character since each register bit could have its own set of values while we restrict the above, as most authors do, to the homogeneous case where each bit is v-valued. But the main feature of exponential generating function lies in the ability to capture entanglement.

Definition 6 (powerset entanglement). *Given an operator $\mathcal{A}$ then the labeled powerset of $\mathcal{A}$ is $\mathcal{P}(\mathcal{A}) = \emptyset + \mathcal{A} + \frac{\mathcal{A}^2}{2!} + \ldots + \frac{\mathcal{A}^k}{k!} + \ldots = \exp^{\mathcal{A}}$ for all k-products formed from $\mathcal{A}$ up to relabeling.*

Starting from any state $|r\rangle$, clearly $\exp^{\mathcal{A}} |r\rangle$ is maximally entangled w.r.t. powerset; for that reason, it is customary to weight this entanglement as $\exp^{\gamma\mathcal{A}}$ where γ smoothly evolves from 1 (maximally entangled case) to 0 identity (separable case).

Definition 7 (sequence entanglement). *Given an operator $\mathcal{A}$ then the labeled sequence of $\mathcal{A}$ is $\Sigma(\mathcal{A}) = \emptyset + \mathcal{A} + \mathcal{A}^2 + \ldots + \mathcal{A}^k + \ldots = \frac{I}{I-\mathcal{A}} = (I-\mathcal{A})^{-1}$.*

Definition 8 (cycle entanglement). *Given an operator $\mathcal{A}$ then the labeled cycle of $\mathcal{A}$ is $\mathcal{C}(\mathcal{A}) = \cup_{k\geq 0}\mathcal{C}(\mathcal{A},k) = \log\frac{I}{I-\mathcal{A}} = -\log(I-\mathcal{A})$, the union of sequences taken up to circular shifts of their elements where $\mathcal{C}(\mathcal{A},k) = \frac{\mathcal{A}^k}{k}$ denotes number of k-sequences under all possible circular shifts.*

However, whenever 1 is eigenvalue of $\mathcal{A}$ both sequence and cycle entanglements run into troubles associated with singularity. This will ever be the case for big n.

Definition 9 (braid entanglement). *Given a SU(2) presentation of Artin braid's group R, links entanglement is written as a braid word in I, R, R^{-1}.*

For instance, $B = (R\otimes I\otimes R^{-1})(I\otimes I\otimes R\otimes I)(I\otimes R^{-1}\otimes I\otimes I)(R\otimes R\otimes I)$ is unitary operator associated with five strands of Fig. 1. See Kaufmann and Lomonaco's articles [27, 28] for a neat and complete presentation of topological versus quantum entanglement.

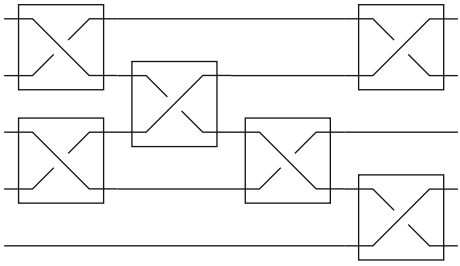


Fig. 1. A braid on five strands from left to right

2.3 Observable State Space

Due to normalization requirement 3, computations are operators preserving norm. We require further *causality*, i.e, norm is unitary similarity invariant: $\|UAU^*\|=\|A\|$ for any state A and unitary U. Unitary invariance through $\|UAV^*\|=\|A\|$ for any state A and unitary U, V does not provide *reversibility* of computation; therefore, it is discarded from the computation model.

Last, quantum computing is supposed to be *safe*; as pointed out in [36], if it were unitary transform, reducible (having a shape $\begin{bmatrix} AB \\ 0C \end{bmatrix}$, equivalently whose supporting graph is strongly connected) then observation of states may collapse since states could remain in a single connected component. It is customary, once more, to deal with the special unitary group ($\det = \pm 1$) instead of unitary group ($\det = \exp^{i\theta}$) to remove phase blurring effects. To summarize, irreducible causal quantum register systems mainly deal with

$$\underbrace{SU(v) \times \cdots \times SU(v)}_{n},$$

the direct product of n times the special unitary group over v-values; unless, the system is closed, any operator in this set is in order.

Let commuting subalgebras $\mathcal{A}_i$ be extracted from the algebra of quantum observables, then the classical/quantum correspondence as illustrated by Khalfin and Tsirelson [29] follows a kind of arithmetic–geometric mean on correlations. Let us denote correlations as standard inner products; for two subsystems $\mathcal{A}_1, \mathcal{A}_2$ with commuting observables A_{1i}, A_{2j}, then $\frac{\langle A_1\rangle^2+\langle A_2\rangle^2}{2} - \langle A_1, UA_2\rangle$ where entries of U belongs to $\{-1, 0, 1\}$ leads to so-called Bell-type inequalities where scaling of right-hand side depends on which classical/quantum case applies. For two observables, Pauli's matrices $P_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, P_2 = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, P_3 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ are the stubs for the original Bell's inequalities through

$$
\begin{aligned}
l\frac{\langle A_1\rangle^2+\langle A_2\rangle^2}{2}-\langle A_1, iP_2A_2\rangle &= \frac{1}{2}\left(A_{11}^2+A_{12}^2+A_{21}^2+A_{22}^2\right)-A_{11}A_{22}+A_{12}A_{21}\\
&= \frac{1}{2}(A_{11}-A_{22})^2+\frac{1}{2}(A_{11}+A_{21})^2\geq 0\\
\frac{\langle A_1\rangle^2+\langle A_2\rangle^2}{2}-\langle A_1, (P_1+P_3)A_2\rangle &= \frac{1}{2}(A_{11}-(A_{21}+A_{22}))^2-\frac{1}{2}(A_{21}+A_{22})^2+\frac{1}{2}(A_{12}-(A_{21}-A_{22}))^2\\
&\quad -\frac{1}{2}(A_{21}-A_{22})^2+\frac{1}{2}\left(A_{21}^2+A_{22}^2\right)\geq 0
\end{aligned}
$$

since all non-commuting terms $A_{21}A_{22}, A_{22}A_{21}$ cancel. From $A_{ij}^2=1$ in last inequality, we retrieve CHSH generic inequality

$$|\,A_{11}(A_{21}+A_{22})+A_{12}(A_{21}-A_{22})\,|\leq 2; \tag{CHSH}$$

in quantum case, Hadamard's matrix $H=(P_1+P_3)$ has to be unitary, hence the well-known scaling by $\sqrt{2}$. For two observables related by a group structure, we have in the same way:

- braid group: use $R=I\otimes I+P_1\otimes iP_2$ as a presentation of the group [14] $\langle A_1, RA_2\rangle\leq 2$ since $\langle A_1\rangle^2=\langle A_2\rangle^2=2$ with a scaling factor $\sqrt{2}$ in quantum case ($RR^t=2I$).
- quaternion group: use $Q=I\otimes I+iP_2\otimes P_3+I\otimes iP_2+iP_2\otimes P_1$ as a presentation [15] $\langle A_1, QA_2\rangle\leq 6$ with a scaling factor 2 in quantum case ($QQ^t=4I$).

It is easily verified that in both cases, $G=R,Q$; first, the 0 mean is preserved in classical case (observables look like $[\pm1,0,\pm1,0]$, $[\pm1,0,0,\pm1]$, $[0,\pm1,\pm1,0]$, $[0,\pm1,0,\pm1]$), and second, non-commuting terms cancel in expansion of $\frac{\langle A_1\rangle^2+\langle A_2\rangle^2}{2}-\langle A_1, GA_2\rangle$. While braiding remains tight (like CHSH) in passing to unitary case, quaternion could not be tight due to the scaling factor 2; we say that the former is facet defining and the latter only defines a valid inequality.

2.4 $SU(v)$ Representation

Fact 1 ($SU(2)$ representation). *If all the special unitary groups over binary values apply then we can use the matrix representation*

$$U_2(\theta,\alpha,\beta)=\begin{bmatrix} e^{i\alpha}\cos(\frac{\theta}{2}) & ie^{i\beta}\sin(\frac{\theta}{2})\\ ie^{-i\beta}\sin(\frac{\theta}{2}) & e^{-i\alpha}\cos(\frac{\theta}{2})\end{bmatrix}, \tag{1}$$

where ranges are in $[-\pi,\pi]$,

with $I=U_2(0,0,0)$ and $F=U_2(\pi,0,0)=\begin{bmatrix}0 & i\\ i & 0\end{bmatrix}$ since $F\,|0\rangle=i\,|1\rangle, F\,|1\rangle=i\,|0\rangle$. Notice $U_2(\pi/2,\pi/2,0)$ is Hadamard's matrix scaled by i to get once more a positive determinant.

For higher dimensional valued system instead of binary, we have to face with the problem of representation of $SU(v)$ w.r.t. entanglement issue mentioned above.

The challenge in quantized problem formulation remains to select the kind of entanglement and to restrict the set of applicable operators; most problems from physics have a clear understanding of both issues as sketched below, but for combinatorial problems, it could be premature to conclude in favor of quantum version over classical one without weighing these issues.

2.5 Amplitude Amplification

Since Grover's algorithm [21] for accelerating search in unstructured data, the notion of amplifying first estimate of successful retrieval has been developed along the following framework [5]. Let $\mathcal{H}$ be decomposed into *good* subspace and *bad* subspace, meaning that every pure state $|x_0, ...x_n\rangle$ is a direct sum of *good* states and *bad* states; let a quantum (unitary) algorithm $\mathcal{A}$ starting from initial state $|0, ...0\rangle$ such that probability of getting a good answer at first step is given by a where $\psi = \mathcal{A}\,|0, ...0\rangle$. $\mathcal{H}$ decomposes into $\mathcal{H}_\psi + \mathcal{H}_\psi^\perp$, the subspace spanned by first answer and its orthogonal complement. Define $Q = H_\psi H_{\psi_0}$ as the product of Householder's reflections through $|\psi\rangle$ and its projection on bad subspace $|\psi_0\rangle$ so that $|\psi\rangle = |\psi_1\rangle + |\psi_0\rangle$. If $a = 1$ then nothing has to be done since we get a positive answer at first trial.

Lemma 1. *Let $\psi = \mathcal{A}\,|0, ...0\rangle = |\psi_1\rangle + |\psi_0\rangle$ and $Q = H_\psi H_{\psi_0}$, then Q acts as the identity on orthogonal complement $\mathcal{H}_\psi^\perp$.*

Proof.

$$H_\psi = I - 2\,|\psi\rangle\langle\psi|, \quad H_{\psi_0} = I - \frac{2}{\langle\psi_0|\psi_0\rangle}\,|\psi_0\rangle\langle\psi_0|\,.$$

Notice that if $a = 0$ then $H_\psi = H_{\psi_0}$ is 1-D and the result is trivial from idempotence of reflection. Let $|\phi\rangle \in \mathcal{H}_\psi^\perp$, i.e., $\langle\psi_0|\phi\rangle = 0, \langle\psi_1|\phi\rangle = 0, \langle\psi|\phi\rangle = 0$.

$$H_\psi H_{\psi_0}\,|\phi\rangle = H_\psi\left(|\phi\rangle - \frac{2}{\langle\psi_0|\psi_0\rangle}\langle\psi_0|\phi\rangle\,|\psi_0\rangle\right) = |\phi\rangle - 2\langle\psi|\phi\rangle\,|\psi\rangle = |\phi\rangle. \qquad \square$$

Lemma 2. *If $0 < a = \langle\psi_1|\psi_1\rangle < 1$, then let us define $\sin(\theta_a) = \sqrt{\langle\psi_1|\psi_1\rangle}$, $\cos(\theta_a) = \sqrt{\langle\psi_0|\psi_0\rangle}$, then*

$$|\psi_\pm\rangle = \frac{1}{\sqrt{2}\sin\theta_a}\,|\psi_1\rangle \pm \frac{i}{\sqrt{2}\cos\theta_a}\,|\psi_0\rangle$$

are unit eigenvectors of Q in $\mathcal{H}_\psi$ associated with eigenvalues $e^{\pm i2\theta_a}$.

Proof. Straightforward computations give

$$\begin{aligned}
|\psi_0\rangle &= \frac{-i\cos\theta_a}{\sqrt{2}}\left(|\psi_+\rangle - |\psi_-\rangle\right), \quad \langle\psi_0| = \frac{i\cos\theta_a}{\sqrt{2}}\left(\langle\psi_+| - \langle\psi_-|\right)\\
|\psi\rangle &= \frac{-i}{\sqrt{2}}\left(e^{i\theta_a}\,|\psi_+\rangle - e^{-i\theta_a}\,|\psi_-\rangle\right), \quad \langle\psi| = \frac{i}{\sqrt{2}}\left(e^{-i\theta_a}\langle\psi_+| - e^{i\theta_a}\langle\psi_-|\right)\\
H_{\psi_0} &= I - (|\psi_+\rangle - |\psi_-\rangle)(\langle\psi_+| - \langle\psi_-|)\\
H_\psi &= I - (e^{i\theta_a}\,|\psi_+\rangle - e^{-i\theta_a}\,|\psi_-\rangle)(e^{-i\theta_a}\langle\psi_+| - e^{i\theta_a}\langle\psi_-|)\\
H_{\psi_0}\,|\psi_\pm\rangle &= |\psi_\mp\rangle, \quad H_\psi\,|\psi_\pm\rangle = e^{\mp i2\theta_a}\,|\psi_\mp\rangle, \quad Q\,|\psi_\pm\rangle = e^{\pm i2\theta_a}\,|\psi_\pm\rangle,
\end{aligned}$$

the result follows from two dimensionality of $\mathcal{H}_\psi$ when $0 < a < 1$. □

Using the decomposition of $|\psi\rangle$ in above eigenvector basis, we get

Corollary 1.

$$Q^j\ |\psi\rangle = \frac{-i}{\sqrt{2}}\left(e^{i(2j+1)\theta_a}\ |\psi_+\rangle - e^{-i(2j+1)\theta_a}\ |\psi_-\rangle\right)$$
$$= \frac{\sin(2j+1)\theta_a}{\sin\theta_a}\ |\psi_1\rangle + \frac{\cos(2j+1)\theta_a}{\cos\theta_a}\ |\psi_0\rangle.$$

While initial amplitude could be amplified in any case, Fig. 2 shows that uncertainty in its knowledge prevents to compute how many diffusion loops have to be done before measurement could take place. A more reliable amplification has been proposed by [42] to partly spread information among qubits. Grover et al. [20] notice that diffusion operator used for searching in a database might be improved by using phase Householder's reflections instead $Q = (I - (1 - e^{i\frac{\pi}{3}})\ |\psi\rangle\langle\psi|)(I - \frac{1-e^{i\frac{\pi}{3}}}{\langle\psi_0|\psi_0\rangle}\ |\psi_0\rangle\langle\psi_0|)$; compared with standard reflections (phase equal to π), it prevents overshooting target state and Li et al. [32] extend it to different phases. Finally, [7] improves ordered search by recoursing to semidefinite programming.

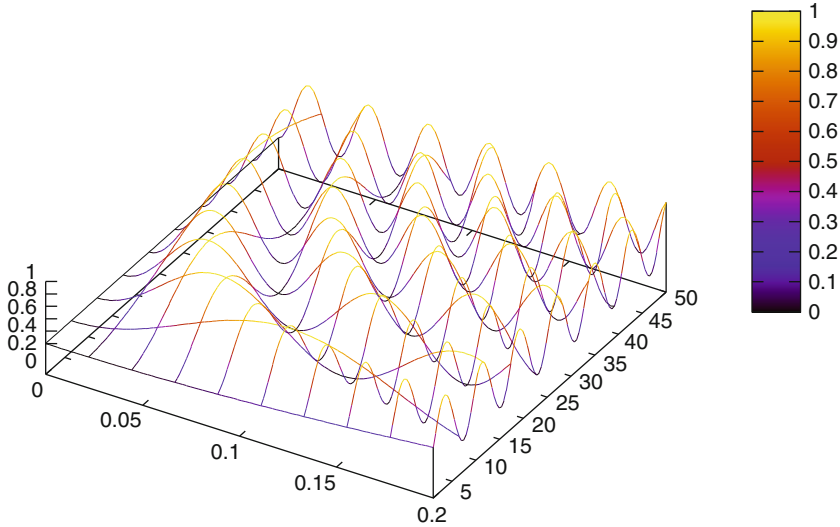


Fig. 2. Amplitude amplification: $\sin^2(2j+1)\theta_a, j = 1\ldots50, \theta_a = 0\ldots0.2$

2.6 Correlation Matrices

Let $\mathcal{C}_n$ (resp. $\mathcal{Q}_n$) denotes the set of classical (resp. quantum) correlation matrices; Tsirelson introduced a Grothendieck's constant K_n to characterize the inclusion

$$\mathcal{Q}_n \subseteq K_n \mathcal{C}_n,$$

see [9] for a detailed presentation. Both sets are centrally symmetric, but the former is a polytope while the latter is smooth, justifying a discrepancy between classical facet defining inequalities and quantum tangent hyperplanes. This gives rise to the discussion about Bell's inequalities and $\sqrt{2}$ scaling in quantum case. In their article, Fishburn and Reeds [16] describe an instance where K_n increases above $\sqrt{2}$ and they report how difficult it could be to exhibit more such instances; in combinatorial optimization, it would mean that the smooth quantum correlation set could be far enough from the classical polytope so that hard problems would be amenable to simple quantum solvers while classical approach fails to escape from a local optimum.

3 The Multiknapsack Multiplayer Game Model

In this section, we carry the prisoner's dilemma quantum game [13] modeling approach over the multiknapsack case. Under multiknapsack constraints, I, F gates model the tendency for an item to keep or revert its status w.r.t. knapsack capacity according to a sequence of trials; therefore, in a maximal diversification approach, each item would try the status opposite to the previous trial to improve the overall profit, so powerset entanglement is the $2^n \times 2^n$ matrix $\mathcal{E} = \exp(\gamma F^{\otimes n})$ having non-null entries on main diagonal while antidiagonal depends on the number of items (see (1)). It is unitary iff $n \equiv \pm 1 \bmod 4$ so that we have to recourse to phase powerset entanglement $\mathcal{E} = \exp(i\gamma J^{\otimes n})$ to deal with an even number of items. Furthermore, it affords to treat even and odd numbers of items the same way, i.e., phase powerset entanglement where $0 \leq \gamma \leq \frac{\pi}{2}$. Let us denote by J_m the $m \times n$ matrix filled with 1's on antidiagonal, where subscript is omitted if no confusion arises; then $\mathcal{E} = \cos(\gamma) I^{\otimes n} \pm i \sin(\gamma) J^{\otimes n}$ and $\mathcal{E} \left|0 \ldots 0\right\rangle = \cos(\gamma) e_0 \pm i \sin(\gamma) e_{2^n-1}$.

Table 1. Powerset entanglement for flipping operator

	Powerset		Phase powerset	
n	diagonal	antidiagonal	diagonal	antidiagonal
$4p$	$\cosh(\gamma)$	$\sinh(\gamma)$	$\cos(\gamma)$	$i\sin(\gamma)$
$4p+1$	$\cos(\gamma)$	$i\sin(\gamma)$	$\cos(\gamma)$	$i\sin(\gamma)$
$4p+2$	$\cosh(\gamma)$	$-\sinh(\gamma)$	$\cos(\gamma)$	$-i\sin(\gamma)$
$4p+3$	$\cos(\gamma)$	$-i\sin(\gamma)$	$\cos(\gamma)$	$-i\sin(\gamma)$

In actual quantum computation, all $SU(2)$ are possible instead; so, let us consider two opponent items, say first two items, with *mixed* strategies $A = U_2(\theta, a, b)$ and $B = U_2(\phi, u, v)$ while each player assumep the remaining items will not change their statuses (identity interaction assumption). Then, starting from an entangled state a quantum computation gives

$$\mathcal{E}^*\left(A\otimes I^{\otimes(n-1)}\right)\mathcal{E}=\cos^2(\gamma)\left(A\otimes I^{\otimes(n-1)}\right)+\sin^2(\gamma)\left(JAJ\otimes I^{\otimes(n-1)}\right)$$
$$\pm i\cos(\gamma)\sin(\gamma)\left(JA\otimes J^{\otimes(n-1)}-AJ\otimes J^{\otimes(p-1)}\right)$$
$$\mathcal{E}^*\left(I\otimes B\otimes I^{\otimes(n-2)}\right)\mathcal{E}=\cos^2(\gamma)\left(I\otimes B\otimes I^{\otimes(n-2)}\right)+\sin^2(\gamma)\left(I\otimes JBJ\otimes I^{\otimes(n-2)}\right)$$
$$\pm i\cos(\gamma)\sin(\gamma)\left(J\otimes JB\otimes J^{\otimes(n-2)}-J\otimes BJ\otimes J^{\otimes(n-2)}\right)$$

with respective measured states

$$|AI\rangle=\cos\left(\frac{\theta}{2}\right)(\cos(a)+i\sin(a)\cos(2\gamma))e_0\pm i\sin\left(\frac{\theta}{2}\right)\sin(b)\sin(2\gamma)e_{2^{n-1}-1}$$
$$+i\sin\left(\frac{\theta}{2}\right)(\cos(b)-i\sin(b)\cos(2\gamma))e_{2^{n-1}}\mp\cos\left(\frac{\theta}{2}\right)\sin(a)\sin(2\gamma)e_{2^n-1}$$
$$|IB\rangle=\cos\left(\frac{\phi}{2}\right)(\cos(u)+i\sin(u)\cos(2\gamma))e_0+i\sin\left(\frac{\phi}{2}\right)(\cos(v)$$
$$-i\sin(v)\cos(2\gamma))e_{2^n-2}\pm i\sin\left(\frac{\phi}{2}\right)\sin(v)\sin(2\gamma)e_{32^{n-2}-1}$$
$$\mp\cos\left(\frac{\phi}{2}\right)\sin(u)\sin(2\gamma)e_{2^n-1},$$

which involve $|011\ldots1\rangle, |111\ldots1\rangle, |101\ldots1\rangle$ that almost surely violate knapsack constraints, $|100\ldots0\rangle, |010\ldots0\rangle$ that trivially fulfill the constraints, and $|110\ldots0\rangle$ we could assume to satisfy the constraints too. Precisely, expected profit of this pair of mixed strategies are

$$c(|AI\rangle)=\ c_{|000\ldots0\rangle}\cos^2\left(\frac{\theta}{2}\right)(\cos^2(a)+\sin^2(a)\cos^2(2\gamma))$$
$$+c_{|011\ldots1\rangle}\sin^2\left(\frac{\theta}{2}\right)\sin^2(b)\sin^2(2\gamma)+c_{|110\ldots0\rangle}\sin^2\left(\frac{\theta}{2}\right)(\cos^2(b)$$
$$+\sin^2(b)\cos^2(2\gamma))+c_{|111\ldots1\rangle}\cos^2\left(\frac{\theta}{2}\right)\sin^2(a)\sin^2(2\gamma)$$
$$c(|IB\rangle)=\ c_{|000\ldots0\rangle}\cos^2\left(\frac{\phi}{2}\right)(\cos^2(u)+\sin^2(u)\cos^2(2\gamma))$$
$$+c_{|010\ldots0\rangle}\sin^2\left(\frac{\phi}{2}\right)(\cos^2(v)+\sin^2(v)\cos^2(2\gamma))$$
$$+c_{|101\ldots1\rangle}\sin^2\left(\frac{\phi}{2}\right)\sin^2(v)\sin^2(2\gamma)+c_{|111\ldots1\rangle}\cos^2\left(\frac{\phi}{2}\right)\sin^2(u)\sin^2(2\gamma).$$

In a similar way, if we consider flipping interaction of remaining items, our two opponent items *mixed* strategies lead to

$$\mathcal{E}^*\left(A\otimes J^{\otimes(n-1)}\right)\mathcal{E}=\cos^2(\gamma)\left(A\otimes J^{\otimes(n-1)}\right)+\sin^2(\gamma)\left(JAJ\otimes J^{\otimes(n-1)}\right)$$
$$\pm i\cos(\gamma)\sin(\gamma)\left(AJ\otimes I^{\otimes(n-1)}-JA\otimes I^{\otimes(n-1)}\right)$$
$$\mathcal{E}^*\left(J\otimes B\otimes J^{\otimes(n-2)}\right)\mathcal{E}=\cos^2(\gamma)\left(J\otimes B\otimes J^{\otimes(n-2)}\right)+\sin^2(\gamma)\left(J\otimes JBJ\otimes J^{\otimes(n-2)}\right)$$
$$\pm i\cos(\gamma)\sin(\gamma)\left(I\otimes BJ\otimes I^{\otimes(n-2)}-I\otimes JB\otimes I^{\otimes(n-2)}\right)$$

with measured states

$$
\begin{aligned}
|AJ\rangle = & \mp i \sin\left(\frac{\theta}{2}\right) \sin(b) \sin(2\gamma) e_0 \cos\left(\frac{\theta}{2}\right) (\cos(a) + i \sin(a) \cos(2\gamma)) e_{2^{n-1}-1} \\
& \pm \cos\left(\frac{\theta}{2}\right) \sin(a) \sin(2\gamma) e_{2^{n-1}} + i \sin\left(\frac{\theta}{2}\right) (\cos(b) - i \sin(b) \cos(2\gamma)) e_{2^n-1} \\
|JB\rangle = & \mp i \sin\left(\frac{\phi}{2}\right) \sin(v) \sin(2\gamma) e_0 \pm \cos\left(\frac{\phi}{2}\right) \sin(u) \sin(2\gamma) e_{2^{n-2}} \\
& + \cos\left(\frac{\phi}{2}\right) (\cos(u) + i \sin(u) \cos(2\gamma)) e_{32^{n-2}-1} \\
& + i \sin\left(\frac{\phi}{2}\right) (\cos(v) - i \sin(v) \cos(2\gamma)) e_{2^n-1}
\end{aligned}
$$

and expected cost functions

$$
\begin{aligned}
c(|AJ\rangle) = & \; c_{|000\ldots0\rangle} \sin^2\left(\frac{\theta}{2}\right) \sin^2(b) \sin^2(2\gamma) + c_{|011\ldots1\rangle} \cos^2\left(\frac{\theta}{2}\right) (\cos^2(a) \\
& + \sin^2(a) \cos^2(2\gamma)) + c_{|100\ldots0\rangle} \cos^2\left(\frac{\theta}{2}\right) \sin^2(a) \sin^2(2\gamma) \\
& + c_{|111\ldots1\rangle} \sin^2\left(\frac{\theta}{2}\right) (\cos^2(b) + \sin^2(b) \cos^2(2\gamma)) \\
c(|JB\rangle) = & \; c_{|000\ldots0\rangle} \sin^2\left(\frac{\phi}{2}\right) \sin^2(v) \sin^2(2\gamma) + c_{|010\ldots0\rangle} \cos^2\left(\frac{\phi}{2}\right) \sin^2(u) \sin^2(2\gamma) \\
& + c_{|101\ldots1\rangle} \cos^2\left(\frac{\phi}{2}\right) (\cos^2(u) + \sin^2(u) \cos^2(2\gamma)) \\
& + c_{|111\ldots1\rangle} \sin^2\left(\frac{\phi}{2}\right) (\cos^2(v) + \sin^2(v) \cos^2(2\gamma)).
\end{aligned}
$$

Neglecting costs of very unlikely feasible states, then $\Pr(x = 1|AI) = \sin^2(\frac{\theta}{2})(\cos^2(b)+\sin^2(b)\cos^2(2\gamma))$ and $\Pr(x = 1|AJ) = \cos^2(\frac{\theta}{2})\sin^2(a)\sin^2(2\gamma)$. Furthermore, $\Pr(x = 1|IB) = \sin^2(\frac{\phi}{2})(\cos^2(v) + \sin^2(v)\cos^2(2\gamma))$ and $\Pr(x = 1|JB) = \cos^2(\frac{\phi}{2})\sin^2(u)\sin^2(2\gamma)$ show it does not change by location interchange. Using rotation gates $a = 0, b = \frac{\Pi}{2}$ as in [23], there is no way to make invariant $\Pr(x = 1|AI) = \Pr(x = 1|AJ)$ under either identity or flipping strategies, for the rest of items. On the contrary, for $a = b = 0$, we keep invariance for $\theta = 4\gamma \in [0, \pi]$ and the gate

$$
Q(\theta) = \begin{bmatrix} \cos(\frac{\theta}{2}) & i\sin(\frac{\theta}{2}) \\ i\sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{bmatrix}
$$

is referred to Q-rotation.

4 The Multiknapsack Quantum Simulation

Of course, we could not aim at simulating a quantum computer since it is intractable to store all 2^n complex coefficients associated to a quantum state;

we could only expect to simulate unentangled states as $\bigotimes_n(\alpha_n\,|0\rangle + \beta_n\,|1\rangle)$. On the other hand, given an unentangled state with α's and β's, there is no numerical difficulties in rotating a single qbit as in the (non-cooperative) game entanglement model from unitary property, but for braiding entanglement the presentation matrix could give rise to the well-known expression swell for coefficients obtained by multiplying α_i, β_j for qbits i, j; the same is true for amplitude amplification, so we have to scale α's and β's and what else apart from $\sum_n |\alpha_n|^2 + |\beta_n|^2$ associated to simple unentangled states? As a consequence, using necessary and sufficient condition for a state to be unentangled (5) [26] coefficient of non-trivial combination of items is, almost surely, 0 and trivial combination (item n) may lead to $|\alpha_n|^2 + |\beta_n|^2$ quite small so that its probability to be 0 or 1 is undetermined 0/0. By no means, we could beat classical heuristics, but a comparison between game and braiding entanglement as well as diffusion benefits could be studied under the quantum scheme shown in Fig. 3 where each module is selected according to the comparison purpose.

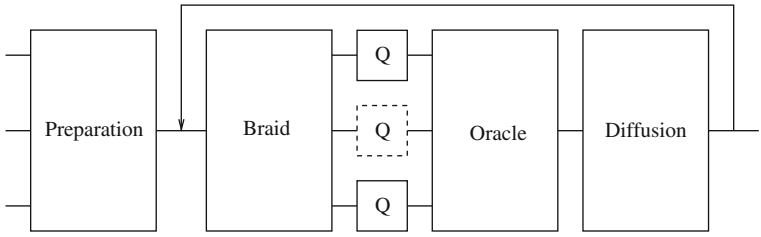


Fig. 3. Multiknapsack quantum simulation

In classical case, Sethi's greedy algorithm is known to perform well for singleknapsack; it simply orders the ratio profit/weight of items to provide a good starting solution for two or more exchange heuristics. This decreasing ordering of items is the basis for a braiding of the associated strands; yet, braiding beyond the capacity constraint does not make sense since those items could not fit the constraint.

In the same way, given the $\Pr(x_i = 1)$ for each qubit, a *greedy* oracle consists in assigning items along decreasing $\Pr(x_i = 1)$ while capacity constraints are fulfilled. Furthermore, this assignment splits items into good and bad components which allows amplitude amplification as described in Section 2.5.

We carry this structural property of single knapsack over the multiknapsack case, braiding each capacity constraint in turn to give one (among many) topological description of the combinatorial problem. To summarize the overall scheme of multiknapsack quantum simulation is drawn in Fig. 3 where Q's, which depend on phase powerset angle θ, play the role of adversary Flipping game strategy as shown in Fig. 3. Notice that braiding and oracle differs only by the comparison between two items, the former relies on the ratios profit/weight and the latter on the probabilities for qubit to be 1.

Preparation may be parameterized in two ways; first, *maximally entangled* case assumes $|0\rangle + |1\rangle$ and second *average probabilities*: use Sethi's greedy algorithm for each constraint to assign items until all capacities are maximally satisfied and take the average (number of times assigned to 1/number of constraints) for the initial $\Pr(x_i = 1)^2$ for item i. Finally, diffusion equal to $H_\psi H_{\psi_0}$ relies on what *good* component does mean; in fact, Householder's reflection reduces a subset of components to its most significant contribution, so diffusion tends to spread the most significant contribution to the objective. In this respect, if the number of items in an optimal solution is less than half the number of items then H_{ψ_0} should reflect the components assigned to 1 in a counterintuitive fashion to Grover's algorithm.

A simulation lasts for a given number of oracles and a given number of angle ticks $\theta = [\pi/2, 0]$ in place of counting down entanglement parameter γ of 3.

5 The Multiknapsack Entanglement Issues

Different simulations on the basis of Fig. 3 have been made on standard benchmarks for multiknapsack and are available from the author. Despite, all of them are far worse than classical heuristics on the problem, what are the issues of studying entanglement for given combinatorial problems? This case study was primarily intended to fight against the pervasive idea that classical evolutionary algorithms might be successfully inspired by the quantum computing paradigm [23]. To this purpose, we put further shed on the role of rotation in entanglement simulation and show that Q-rotation is underlied by a non-cooperative game entanglement modeling. We also observe in our simulations that amplitude amplification effect is rather unlikely unless a global definition for good and bad components remains invariant; under this circumstance, it gives an interpretation of its efficiency in terms of non-cooperative game entanglement by means of Q-rotations. Braiding entanglement simulations are more efficient than non-cooperative game modeling and there are rational arguments for such behavior: from the one hand, it is tightly related to greedy algorithm on single knapsack and from the other hand a presentation of the braiding group is likely to induce facet defining inequalities. However, using quaternion group instead leads to very close objective values, albeit solutions are completely different. Therefore, it is not strong evidence that actual quantum computers would prove effectiveness of one model over the other; this case study suggests new questions about entanglement: Is oddity an artefact we have to workaround? with phase powerset entanglement as we did above or with any specific trick as in [7] with unitary $\begin{bmatrix} I & 0 & f \\ 0 & \sqrt{2}0 & \\ J & 0 & J \end{bmatrix}$ for odd cases. We argue that in combinatorial optimization problems, powerset entanglement could not render the combinatorial explosion underlied by the constraints structure that imposes a cooperative interaction on the contrary to previous model. Let us sketch how to deal with cooperation for the multiknapsack; the

best classical heuristics known, work around the number of items in a feasible solution; it is easy to bound this number by an interval $[k_l, k_u]$; and for each value to specialize search space by adding a constraint $\langle e, x \rangle = k$ fixing this number. Despite combinations are still blowing up, it opens up tracks for modeling entanglement at a cooperative level. Let $\mathcal{A}_k^1$ be a non-cooperative game/braiding modeling for a combination of k items then $\exp(\mathcal{A}_k^1)$ stands for powerset entanglement of this coalition as we have done along Fig. 3. Applying it to all such combinations $i \in \binom{n}{k}$ we get a cooperative game entanglement $\prod_i \exp(\mathcal{A}_k^i) = \exp\left(\sum_i \mathcal{A}_k^i\right)$. Summing over all $k \in [k_l, k_u]$ we arrive at a Dirichlet series $\mathcal{D}(\mathcal{A}, s) = \sum_{k_l}^{k_u} \frac{\exp\left(\sum_i \mathcal{A}_k^i\right)}{k^s}$. For instance, with two items we directly get $\mathcal{D}(\mathcal{A}, s) = I + \frac{1}{2^s} \begin{bmatrix} \cos\gamma & i\sin\gamma) \\ i\sin\gamma & \cos\gamma \end{bmatrix}$; unitary condition implies $\cos\gamma = -2^{s-1}$ for $s < 1$, bounding γ within $[\pi/2, \pi]$. But, to deal with only three items, we have to add identity with (phase) powerset entanglement of

- 2-coalitions

$$a_{12} = \begin{bmatrix} \cos\gamma & 0 & 0 & i\sin\gamma & 0 & 0 \\ 0 & \cos\gamma & i\sin\gamma & 0 & 0 & 0 \\ 0 & i\sin\gamma & \cos\gamma & 0 & 0 & 0 \\ i\sin\gamma & 0 & 0 & \cos\gamma & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad a_{13} = \begin{bmatrix} \cos\gamma & 0 & 0 & 0 & 0 & i\sin\gamma \\ 0 & \cos\gamma & 0 & 0 & i\sin\gamma & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & i\sin\gamma & 0 & 0 & \cos\gamma & 0 \\ i\sin\gamma & 0 & 0 & 0 & 0 & \cos\gamma \end{bmatrix}$$

$$a_{23} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos\gamma & 0 & 0 & i\sin\gamma \\ 0 & 0 & 0 & \cos\gamma & i\sin\gamma & 0 \\ 0 & 0 & 0 & i\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & i\sin\gamma & 0 & 0 & \cos\gamma \end{bmatrix}$$

- 3-coalitions

$$a_{123} = \begin{bmatrix} \cos\gamma & 0 & 0 & 0 & 0 & i\sin\gamma \\ 0 & \cos\gamma & 0 & 0 & i\sin\gamma & 0 \\ 0 & 0 & \cos\gamma & i\sin\gamma & 0 & 0 \\ 0 & \imath\sin\gamma & \cos\gamma & 0 & 0 & 0 \\ 0 & i\sin\gamma & 0 & 0 & \cos\gamma & 0 \\ i\sin\gamma & 0 & 0 & 0 & 0 & \cos\gamma \end{bmatrix}$$

scaled by $1, 1/2^s, 1/3^s$, respectively, so that $\left(I + \frac{a_{12}+a_{13}+a_{23}}{2^s} + \frac{a_{123}}{3^s}\right)$ is unitary.

Though feasible in principle, cooperative game entanglement will remain intractable for real-life problems we are addressing in combinatorial optimization. Moreover, braiding entanglement whose relevance to multiknapsack is sound does not directly enter this Dirichlet approach.

6 Hadamard Matrices and Fishburn and Reeds Formulation

The key to understand Fishburn and Reeds' elegant work comes after noticing the relationship between their formulation and quantum states; given a qbit in state $\alpha_i \ |0\rangle + \beta_i \ |1\rangle$ interesting cases in regard to K_n arise for highly degenerate interaction between 2 qbits i, j:

$$\begin{bmatrix} 1\,0 & 1\,0 \\ 0\,0 & 0\,0 \\ 1\,0 & -1\,0 \\ 0\,0 & 0\,0 \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \\ \alpha_j \\ \beta_j \end{bmatrix}.$$

Discarding meaningless rows and columns of 0s, we recognize the 2×2 Hadamard matrix H_2. It plays a prominent role in quantum computing from its unitary similar property up to normalization ($H_2 H_2^t = 2I$). Let us define v_1^{ij}, v_2^{ij} as the column vectors of interaction between i, j (understanding, no interaction with the remaining qbits), then Fishburn and Reeds' sample is nothing else than $F = H\binom{n}{2}$ for all combinations of 2 among n qbits. Denoting column vectors in F by superscripts and components of vectors by subscripts, we get the formulation for computing the Grothendieck-like constant, in concise notations

$$K\binom{n}{2} = \frac{N\binom{n}{2}}{D\binom{n}{2}} = \frac{\max\limits_{\|\eta^j\|=1} \sum\limits_{i=1}^{n(n-1)/2} \left\| \sum\limits_{j=1}^{n(n-1)/2} \left\langle \bar{f}^i, \bar{f}^j \right\rangle \eta^j \right\|}{\max\limits_{\epsilon_j \in \{-1,+1\}} \sum\limits_{i=1}^{n(n-1)/2} \sum\limits_{j=1}^{n(n-1)/2} \left\langle f^i, f^j \right\rangle \epsilon_i \epsilon_j},$$

where $\bar{f}^i = \bar{\epsilon}_i f^i$ stands for solution values of maximized denominator.

The numerator happens to be a convex maximization under the sphere of unit vectors in some dimensional space $\mathbf{R}^p$ while the denominator is a convex maximization on the non-convex box domain since in matrix notations it amounts to $\max_\epsilon \langle F^t F \epsilon, \epsilon \rangle$ for ϵ in box. As for the numerator, from the positivity of norm, the maximum is achieved at the same point as squared norms instead of norms, i.e.,

$$\begin{aligned} N'\binom{n}{2} &= \sum_{i=1}^{n(n-1)/2} \left\langle \sum_{j=1}^{n(n-1)/2} \left\langle \bar{f}^i, \bar{f}^j \right\rangle \eta^j, \sum_{k=1}^{n(n-1)/2} \left\langle \bar{f}^i, \bar{f}^k \right\rangle \eta^k \right\rangle \\ &= \sum_{i=1}^{n(n-1)/2} \sum_{j=1}^{n(n-1)/2} \sum_{k=1}^{n(n-1)/2} \left\langle \bar{f}^i, \bar{f}^j \right\rangle \left\langle \bar{f}^i, \bar{f}^k \right\rangle \left\langle \eta^j, \eta^k \right\rangle. \end{aligned}$$

Fishburn and Reeds' trick consists in using Cauchy–Schwarz inequality for bounding the value together with the knowledge that the maximum is achieved when vectors are colinear; it applies to $p = 2n$ and vectors $\eta^j = \begin{bmatrix} \bar{f}^j \\ -\bar{f}^j \end{bmatrix}$, say colinear to original vectors by abuse of language, and it yields the maximum value $\frac{1}{\sqrt{2}} \sum_{i=1}^{n(n-1)/2} \left\| \sum_{j=1}^{n(n-1)/2} \bar{\epsilon}_i \left\langle f^i, f^j \right\rangle f^j \right\|$ after canceling the two factors from p dimension.

In other words, all the difficulty comes from maximizing the denominator in a general setting. In the special case with Hadamard's matrices H_2

Theorem 1 ([16]).

$$K\left(\tbinom{n}{2}\right) \geq \frac{3n-3}{2n-1}$$

Fishburn and Reeds introduce another improvement from the flat spectrum of Hadamard matrices. Clearly $(F^tF - \lambda I)$ remains semidefinite positive for any $0 \leq \lambda \leq 2$, so let us consider $\langle f^i, f^j \rangle_\lambda = \langle f^i, f^j \rangle$ for all $j \neq i$ and $\langle f^i, f^i \rangle_\lambda = \langle f^i, f^j \rangle - \lambda$ then $\max_\epsilon \langle (F^tF - \lambda I)\epsilon, \epsilon \rangle = \max_\epsilon \langle F^tF\epsilon, \epsilon \rangle - \lambda \binom{n}{2}$ is attained at the same point. The numerator is shifted accordingly as

$$\begin{aligned}
N_\lambda\left(\tbinom{n}{2}\right) &= \sum_{i=1}^{n(n-1)/2} \left\| \sum_{j=1}^{n(n-1)/2} \langle \bar{f}^i, \bar{f}^j \rangle_\lambda \eta^j \right\| \\
&= \sum_{i=1}^{n(n-1)/2} \sum_{j=1}^{n(n-1)/2} \left\| \sum_{j=1}^{n(n-1)/2} \langle \bar{f}^i, \bar{f}^j \rangle \eta^j - \lambda \eta^i \right\| \\
&\geq N\left(\tbinom{n}{2}\right) - \lambda \sum_{i=1}^{n(n-1)/2} \|\eta^i\| \\
&= N\left(\tbinom{n}{2}\right) - \lambda \tbinom{n}{2}.
\end{aligned}$$

Whence an improved lower bound

$$K_\lambda\left(\tbinom{n}{2}\right) \geq \frac{N\left(\tbinom{n}{2}\right) - \lambda\tbinom{n}{2}}{D\left(\tbinom{n}{2}\right) - \lambda\tbinom{n}{2}}.$$

For instance, $K\left(\tbinom{10}{2}\right) = 1.4210, K\left(\tbinom{8}{4}\right) = 1.4064$ lower bounds surround $\sqrt{2}$ in Table 3 without deploying much effort compared to Fishburn and Reeds' refined analysis used to extract a better constant for $\lambda = 4/3$ and $\binom{5}{2}$.

Fishburn and Reeds' sample of increasing Grothendieck's constant relies on the rank deficiency of unitary similar matrix transform on pairs of 2 qbits; in this sense, the rank deficiency is maximally propagated to all combinations of two among n qbits. Therefore, we could ask for the constant under maximally rank-deficient Hadamard matrices of size $m \equiv 0 \bmod 4$ and all combinations $\binom{n}{m}$. Nothing is changed for maximizing the numerator but the scaling; $\eta^j = \begin{bmatrix} \bar{f}^j \\ -\bar{f}^j \end{bmatrix}$, *colinear* to original vectors yields the maximum value at

$$\begin{aligned}
N\left(\tbinom{n}{m}\right) &= \frac{1}{\sqrt{2m}} \sum_{i=1}^{n} \left\| \sum_{j=1}^{n} 2\bar{\epsilon}_i \langle f^i, f^j \rangle f^j \right\| \\
&= \frac{1}{\sqrt{m}} \sum_{i=1}^{n} \left\| \sum_{j=1}^{n} \bar{\epsilon}_i \langle f^i, f^j \rangle f^j \right\|.
\end{aligned}$$

Table 2. Hadamard representatives of size 4

$$\begin{bmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix}$$

Unfortunately, as could be seen in Table 3, the largest the combinations, the hardest the maximization problems to solve so that we are not able to produce in practice so many instances overtaking $\sqrt{2}$; the trends confirm Fishburn and Reed's instance as the easiest one. The technique used to solve the maximization of the denominator is beyond the scope of this chapter, and we could not prove optimality was reached since we abort enumeration of Branch and Bound tree and we could wonder whether a rational form in n exists for all Hadamard matrices as Fishburn and Reeds found for H_2. However, we conjecture that for all Hadamard matrices, there exists sufficiently large n such that Grothendieck constant $K\binom{n}{m} > \sqrt{2}$.

Table 3. Grothendieck's constant $K\binom{n}{m}$ for Hadamard matrices

m	$K\binom{m+1}{m}$	$K\binom{m+2}{m}$	$K\binom{m+3}{m}$
2	1.2000	1.2857	1.3333
4	1.1765	1.2821	1.3540
8	1.1077	1.1903	1.2563
12	1.0759	1.1395	
16	1.0584		
20	1.0474		
24	1.0399		

7 Lehman Matrices and Grothendieck's Constant

A pair of $n \times n$ square $0,1$ matrices (A, B) such that $AB^t = E + kI$ where E, I are respectively the all 1's and identity matrices is called Lehman and B is called the dual of A. Matrices can be thought as vertex–vertex adjacency matrix of graphs, then it is known.

Theorem 2 (Bridges and Ryser). *Let (A, B) be a Lehman pair. Then there exist integers $r \geq 2, s \geq 2$ such that A graph is r-regular, B graph is s-regular, and $rs = n + k$; moreover (A^t, B^t) is a Lehman pair too.*

Normalizing by $1/\sqrt{k}$, a Lehman pair $\frac{1}{k}AB^t = I + \frac{1}{k}E$ acts as a *minimally* perturbed unitary invariant operator, in the sense that perturbation has rank 1. Clearly, applying transformation $A \mapsto 2A - E$, we get under ± 1 domain, $(2A - E)(2B^t - E) = 4(E + kI) - 2AE - 2EB^t + nE = (rs - 2(r + s) + 4 - k)E + 4kI$; by abuse of language we call the transformed pair a Lehman pair too. In other words, Lehman pairs provide the minimally perturbed unitary invariant operators. In [11], Cornuéjols et al. fully characterize level one Lehman pairs, those which fulfill $AB^t = E + I$, as well as nearly self-dual Lehman pairs $(A, A + I)$ while it is known that point-line incidence matrices A of finite projective planes of order k define self-dual pairs

$AA^t = E + kI$. In Fano plane (order 2) we arrive at a unitary similar *nearly* invariant operator $\frac{1}{8}F_7F_7^t = I + \frac{1}{8}E$. On the contrary to Cornuéjols et al. we do not focus on the global interaction between n qbits (Lehman pairs whose entries belong to $0, \pm 1$) but highly degenerated cases like Fishburn and Reeds did, i.e., we study homogeneous (fixed size) coalitions of qbits that interact among themselves with high rank deficiency unitary-based operator while all coalitions are dumped in a total of n qbits.

In Tables 4 and 5 our maximization program seems to retrieve a global solution for smallest m (even though Branch and Bound tree is too large to be completely enumerated) and simply a local solution for next m.

Table 4. Grothendieck's constant $K\binom{n}{m}$ for projective planes of order k

k	m	$K\binom{m+1}{m}$	$K\binom{m+2}{m}$	$K\binom{m+3}{m}$
2	7	1.1634	1.3224	1.3249
3	13	0.5820		

Table 5. Grothendieck's constant $K\binom{n}{m}$ for conference matrices

m	$K\binom{m+1}{m}$	$K\binom{m+2}{m}$	$K\binom{m+3}{m}$
6	0.9748	1.01404	1.0416
10	0.7389	1.3805	

Self-dual pairs clearly follow Fishburn and Reeds' formulation since, let $F[H_m]\binom{n}{m}$ be the sample generated from a self-dual Lehman pair (H_m, H_m) and R_m, C_m be any permutation matrices acting on rows and columns of H_m then

$$F[R_m H_m C_m]\binom{n}{m} = R\binom{n}{m} F[H_m]\binom{n}{m} C\binom{n}{m}$$

for some row and column permutations $R\binom{n}{m}, C\binom{n}{m}$ acting on the whole sample. Clearly, the denominator is unchanged under row or column permutations since $\langle C^t F^t R^t RFC\epsilon, \epsilon\rangle$ amounts to permute the columns (and components of ϵ accordingly) for any F.

However, general Lehman pairs (A, B) lead to the formulation

$$K\binom{n}{2} = \frac{N\binom{n}{2}}{D\binom{n}{2}} = \frac{\max\limits_{\|\eta^j\|=1} \sum\limits_{i=1}^{n(n-1)/2} \left\| \sum\limits_{j=1}^{n(n-1)/2} \langle \bar{a}^i, \bar{b}^j \rangle \eta^j \right\|}{\max\limits_{\epsilon_j \in \{-1,+1\}} \sum\limits_{i=1}^{n(n-1)/2} \sum\limits_{j=1}^{n(n-1)/2} \langle a^i, b^j \rangle \epsilon_i \epsilon_j},$$

where $\bar{a}^i = \bar{\epsilon}_i a^i, \bar{b}^i = \bar{\epsilon}_i b^i$ stand for solution values of maximized denominator.

Denominator is no longer convex and no Cauchy–Schwarz inequality may relate optimal solution of the numerator to optimal solution of the denominator even though positivity of norm guarantees the same solution for squared norm

$$N'\binom{n}{2} = \sum_{i=1}^{n(n-1)/2} \left\langle \sum_{j=1}^{n(n-1)/2} \langle \bar{a}^i, \bar{b}^j \rangle \eta^j, \sum_{k=1}^{n(n-1)/2} \langle \bar{a}^i, \bar{b}^k \rangle \eta^k \right\rangle$$
$$= \sum_{i=1}^{n(n-1)/2} \sum_{j=1}^{n(n-1)/2} \sum_{k=1}^{n(n-1)/2} \langle \bar{a}^i, \bar{b}^j \rangle \langle \bar{a}^i, \bar{b}^k \rangle \langle \eta^j, \eta^k \rangle.$$

As in Hadamard case, the $\sqrt{2}$ looks feasible for many different instances of self-dual Lehman pairs while general pairs are much more difficult to deal with.

8 Grothendieck's Constant and Root 2 Issues

8.1 $2-(v,k,\lambda)2$ Designs

Lehman's matrices suggest that many new classes of matrices are good candidates for $\sqrt{2}$ violation.

Definition 10 (t-design). *Given a v-set $\mathcal{V}$ (called points), a $t-(v,k,\lambda)$-design $T_{t,v,k,\lambda} = (\mathcal{V}, \mathcal{B})$ (resp. packing $P_{t,v,k,\lambda}$, covering $C_{t,v,k,\lambda}$) is a collection $\mathcal{B}$ of k-subsets (called blocks) of $\mathcal{V}$ such that every t-subset of $\mathcal{V}$ is contained in exactly (resp. at most, at least) λ blocks.*

Let $b = |\mathcal{B}|$, then each point belongs to $r = \frac{bk}{v}$ blocks. Existence of design is known for very long.

Theorem 3 (Fisher's inequality). *Let $0 < t \leq k \leq v - t$; if $T_{t,v,k,\lambda} = (\mathcal{V}, \mathcal{B})$ is a $t-(v,k,\lambda)$-design then $|\mathcal{B}| \geq \binom{v}{\lfloor t/2 \rfloor}$.*

Theorem 4 (Wilson). *Let $0 < t \leq k \leq v$, there exists some λ^0 such that for any $\lambda > \lambda^0$ some $T_{t,v,k,\lambda} = (\mathcal{V}, \mathcal{B})$ admissible exists (possibly with repeated blocks).*

An incidence matrix $\mathcal{T}^v_{t,k}$ of size $\binom{v}{t} \times \binom{v}{k}$ is defined between t-subsets and k-subsets such that $\mathcal{T}^v_{t,k}[T,K] = 1$ iff $T \subseteq K$. Then t-design, maximum t-packing, and minimum t-covering are solutions of

- set partitioning $\mathcal{T}^v_{t,k} x = \lambda e$
- set packing $\max \langle e, x \rangle$, s.t. $\mathcal{T}^v_{t,k} x \leq \lambda e$
- set covering $\min \langle e, x \rangle$, s.t. $\mathcal{T}^v_{t,k} x \geq \lambda e$

for $x \in \{0,1\}^{\binom{v}{k}}$ and where e is the all 1's vector. A constructive proof of designs is therefore tightly connected to assignment problems in combinatorial optimization. Among all designs, those with $t = 2$ are, once more, closest to unitary 0/1 matrices in the sense of rank deficiency, since

Theorem 5 ($2-(v,k,\lambda)$ design). *let X be the point-block incidence matrix of a $2-(v,k,\lambda)$ design*

$$XX^t = (r-\lambda)I_v + \lambda E_v$$
$$E_v X = kE_{vb}$$

under admissibility conditions $b = \lambda \binom{v}{2} / \binom{k}{2}$, $r = bk/v = \lambda(v-1)/(k-1)$, where I_v, E_v, E_{vb} denote the identity matrix of size v, the all 1's square matrix of size v and the all 1's matrix of size $v \times b$, respectively.

A design is symmetric (square) if $b = v$ and two designs arise in self-dual pairs (X, X^t) through their point-block incidence matrices.

This opens a wide range of non-convex maximization problems to compute the denominator in Grothendieck's constant fraction after the variable change $X \mapsto 2X - E_v$ and $\binom{n}{v}$ combinations; on the contrary to Lehman's case, the transform does not necessarily keep the rank-deficiency property, so the $\sqrt{2}$ violation is likely to be harder to find.

Apart from the challenging task to solve these maximization problems for themselves, we stress, in next sections, that entanglement and Grothendieck's constant computation are tightly coupled.

8.2 Classical/Quantum Metaheuristics Issues

It is known that most efficient metaheuristics [19, 34] for combinatorial optimization, play with adaptive memory (to prevent examination of the same subspace repeatedly) and variable neighborhood search (to intensify/diversify search). The key point is that neither the input nor the output sample in basic steps of such methods are required feasible; it is even observed that zigzaging around the borderline between feasible and non-feasible solution sets, e.g., taking convex combinations of samples in either set, provides very good solutions at end. On the other hand, *quantum-inspired evolutionary* algorithms claim they could improve classical metaheuristics by generating complex samples through Q-rotation gate whose legitimacy in multiknapsack case suffers from oddity restrictions. Instead of mimicking quantum computers by classical ones, we would ask whether classical metaheuristics are able to approximate a Grothendieck's constant as some distance between feasible and non-feasible sets? After all, a non-feasible sample in classical case is related to a measured probability of a larger set of solutions in complex space, what else? Unless a realistic model with unitary complex gates is runnable on an actual quantum computer, classical simulation of quantum computing is certainly not viable.

Grothendieck's constant has been settled for the $\sqrt{2}$ violation issue, whence the minimal rank (almost unitary) matrices and the uniform combinations of

coalitions among quantum bits; however, hypergraphs with hyperedges labeled with dedicated *almost unitary* matrices are Grothendieck's constant defining (on a very speculative level).

8.3 Quantum Oddities

In Fig. 4, braiding entanglement was claimed to take into account the topological constraints within items; a major difficulty arises with the ordering in the rigid representation of the corresponding braid so that, even on a quantum computer, there is no unique program to get a solution with corresponding R-gates. It lacks of *parallelism* in links entanglement; it is provided by invariance under Reidemeister moves in $\mathbf{R}^3$ and so-called

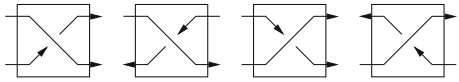


Fig. 4. Oriented representation of link crossings: $W_i, i = 1 \dots 4$

Definition 11 (four-weight spin model). *A 4-tuple* (W_1, W_2, W_3, W_4) *of* $n \times n$ *complex matrices is a four-weight spin model iff the following conditions are satisfied:*

$$W_1(i,j)W_3(j,i) = 1, \quad W_2(i,j)W_4(j,i) = 1, \quad \textit{for all } i,j = 1,n$$

$$\sum_r W_1(i,r)W_3(r,j) = n\delta(i,j), \quad \sum_r W_2(i,r)W_4(r,j) = n\delta(i,j), \quad \textit{for all } i,j = 1,n$$

$$\textstyle\sum_r W_2(i,r)W_2(j,r)W_4(r,k) = \sqrt{n}W_1(j,i)W_3(i,k)W_3(k,j), \quad \textit{for all } i,j,k = 1,n$$

$$\textstyle\sum_r W_2(r,i)W_2(r,j)W_4(k,r) = \sqrt{n}W_1(i,j)W_3(k,i)W_3(j,k), \quad \textit{for all } i,j,k = 1,n,$$

where first two equations are associated with Reidemeister moves of type II (Fig. 5) and last two with Reidemeister moves of type III (Fig. 6) [3, 24]. Notice, for type III, the symmetric role is played by i, j so that all handsides are equal. Setting $j = k$ in type III equations and using type II equations, it implies $\sum_r W_2(i,r) = \sum_r W_2(r,i) = \sqrt{n}W_3(j,j)$, $\sum_r W_4(i,r) = \sum_r W_4(r,i) = \sqrt{n}W_1(j,j)$ for all i, j, whence existence of a complex *modulus* $\mu \in \mathbf{C}$ of the spin model and $W_1(i,i) = \mu$, $W_3(i,i) = 1/\mu$; these last equations provide invariance under Reidemeister move of type I. All equations could be written in concise form in the algebra of standard matrix product, matrix transpose, and Hadamard product $\circ$ (matrix entrywise product). In particular, $W_1W_3 = nI$, $W_2W_4 = nI$ enlarge the candidates for Grothendieck constants in the spirit of

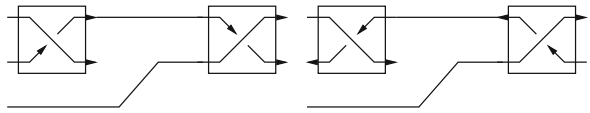


Fig. 5. Oriented Reidemeister moves of type II

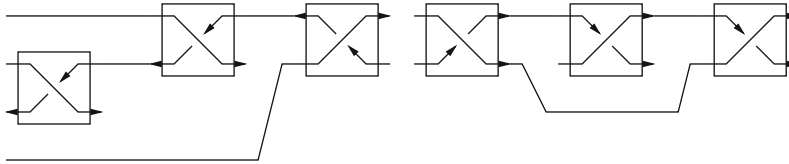


Fig. 6. Oriented Reidemeister move of type III

Hadamard, Lehman, design incidence of previous constructions. Notice also that orientation generalizes the Yang–Baxter equation

$$(R \otimes I)(I \otimes R)(R \otimes I) = (I \otimes R)(R \otimes I)(I \otimes R) \quad (Yang{-}Baxter)$$

used to find a presentation of the braiding group in direct entanglement modeling.

Definition 12 (quasi design). *A* 2*-design is quasi-2 (resp. quasi-3) if the intersection of any 2 (resp. 3) blocks could take only two values.*

Quasi-2 is referred to quasi for short. In a similar way,

Definition 13 (quasi-3 Hadamard). *An Hadamard matrix is quasi-3 if for any three distinct rows, the number of columns where all three rows have* -1 *takes only two values.*

Definition 14 (row/column regularity). *A matrix A is row (resp. column) k-regular iff rows (resp columns) sum to k, i.e.,* $Ae = k$ *(resp.* $A^t e = k$*).*

Four-weight spin models are connected to quasi-3 designs by

Theorem 6 (Bannai-Sawano [3]). *Let* $W_2 = (\alpha - \beta)A + \beta E$ *for complex numbers* α, β *and 0-1 matrix A row and column k-regular with* $2 \leq k \leq n-2$*, then* W_2 *defines a four-weight spin model if and only if A is the incidence matrix of quasi-3 symmetric design* $2-(n,k,\lambda)$ *where the intersection of three distinct blocks takes the values* $(k\lambda - (k-\lambda) \pm (k-\lambda)^{3/2})/n$*.*

Since, a row regular Hadamard matrix H of size $4u^2$ maps to an incidence matrix $(E-H)/2$ of a symmetric $2-(4u^2, 2u^2-u, u^2-u)$ design, the quasi designs are in turn connected to Hadamard matrices.

- **Theorem 7 (Bracken-MacGuire-Ward [4]).** *Let u be an even positive number. Suppose there exists a $2u \times 2u$ Hadamard matrix and $u-2$ (resp. $u-1$) mutually orthogonal $2u \times 2u$ Latin squares, then there exists a quasi-2 symmetric $2-(2u^2-u, u^2-u, u^2-u-1)$ (resp. $2-(2u^2+u, u^2, u^2-u)$) design with double intersection sizes $u(u-1)/2, u(u-2)/2$ (resp. $u^2/2, u(u-1)/2$).*
  Notice the hidden oddity in this result that reminds the difficulty we faced on direct entanglement modeling.
- **Theorem 8 (Broughton-MacGuire [6]).** *Let H, K be quasi-3 regular Hadamard matrices with respective sizes $4u^2, 4w^2$ and triple intersection sizes $u(u-1)/1, u(u-2)/2$, $w(w-1)/1, w(w-2)/2$, then $H \otimes K$ is a quasi-3 regular matrix of size $4(2uw)^2$ and triple intersection sizes $uw(2uw-1), 2uw(uw-1)$.*
  In Table 2, leftmost matrix is row regular quasi-3, so that $\otimes^m H$ are too.

On the contrary to Fishburn and Reeds' thinking, the spectrum of candidates for $\sqrt{2}$ violation seems quite large, even though proving the violation becomes harder and harder as the sizes increase.

9 Concluding Remarks

The aim of this study is twofold: first, we give an operational, although non-cooperative, quantum game model for binary programming; second, we argue for many instances of rank-deficient correlation matrices whose Grothendieck's constant go beyond $\sqrt{2}$ for sufficiently large size. However, combinatorial optimization highly involves cooperative interaction with subsets of variables; simple tracks to move in this direction reveals major difficulties:

- Dirichlet's entanglement in place of (unconstrained) powerset entanglement becomes intractable beyond very small sizes;
- braiding entanglement could handle specific (with non-negative coefficients) constraints but it lacks of parallelism within the many possible representations affordable;
- Grothendieck's constant computation is NP-hard unless we discover analytical solutions (as Fishburn and Reeds did) for the different classes related to Hadamard's matrices.

Within the spectrum of cooperative quantum games, powerset entanglement and Grothendieck's constant appear as extreme cases since the former assumes independent interaction and the latter fixed-size interaction among binary variables. Braiding entanglement intends to enrich the first with, indirect, partial ordering and on the other hand Grothendieck's constant could be defined on non-uniform hypergraphs of interaction. Notice that for multiknapsack, fixing the number of items in optimal cooperative interaction makes sense in regards with efficient heuristics known; second, it is also known that

most efficient heuristics travel around the borderline between feasible and non-feasible sets, the constant answers the question how far from feasible a candidate has to be considered?

Somehow, both our attempts drastically lack of theoretical foundation for constrained binary programs; it suggests to address for future study, simpler binary programs, like assignments, in the spirit of covering coalitions and quantum calculus [17, 25] to put further shed on quantum constrained modeling.

References

1. Adenier, G.: Refutation of Bell's Theorem. In: Foundations of Probability and Physics (Vol. 13, pp. 29–38) *QP–PQ: Quantum Probability and White Noise Analysis*. World Science, Publishing, River Edge, NJ (2001)
2. Avis, D., Fukuda, K.: A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. Discrete Comput. Geom. 8(3), 295–313 (1992) ACM Symposium on Computational Geometry (North Conway, NH, 1991)
3. Bannai, E., Sawano, M.: Symmetric designs attached to four-weight spin models. Des. Codes Cryptogr. 25(1), 73–90 (2002)
4. Bracken, C., McGuire, G., Ward, H.: New quasi-symmetric designs constructed using mutually orthogonal Latin squares and Hadamard matrices. Des. Codes Cryptogr. 41(2), 195–198 (2006)
5. Brassard, G., Hoyer, P., Mosca, M., Tapp, A.: Quantum amplitude amplification and estimation. In Quantum Computation and Information (Washington, DC, 2000) (Vol. 305, pp. 53–74), Contemporary Mathematics. Amer. Math. Soc., Providence, RI (2002)
6. Broughton, W., McGuire, G.: Some observations on quasi-3 designs and Hadamard matrices. Des. Codes Cryptogr. 18(1–3), 55–61 (1999) Designs and codes – A memorial tribute to Ed Assmus.
7. Childs, A.M., Landahl, A.J., Parrilo, P.A.: Improved quantum algorithms for the ordered search problem via semidefinite programming (2006)
8. Chu, P.C., Beasley, J.E.: A genetic algorithm for the multidimensional knapsack problem. J. Heuristics 4, 63–86 (1998)
9. Cirel'son, B.S.: Quantum generalizations of Bell's inequality. Lett. Math. Phys. 4(2), 93–100 (1980)
10. Collins, D., Gisin,N., Linden, N., Massar, S., Popescu, S.: Bell inequalities for arbitrarily high-dimensional systems. Phys. Rev. Lett. 88(4), 040404, 4 (2002).
11. Cornuéjols, G., Guenin, B., Tunçel, L.: Lehman matrices (2006) http://integer.tepper.cmu.edu/webpub/Lehman-v06.pdf
12. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algor. 22(1), 60–65 (2003)
13. Du, J.F., Xu, X., Li, H., Zhou, X., Han, R.: Playing prisoner's dilemma with quantum rules. Fluct. Noise Lett. 2(4), R189–R203 (2002)
14. Dye, H.A.: Unitary solutions to the Yang-Baxter equation in dimension four. Quant. Inf. Process. 2(1-2), 117–151 (2003)

15. Farebrother, R.W., Groß, J., Troschke, S.-O.: Matrix representation of quaternions. Linear Algebra Appl. 362, 251–255 (2003)
16. Fishburn, P.C., Reeds, J.A.: Bell inequalities, Grothendieck's constant, and root two. SIAM J. Discrete Math. 7(1), 48–56 (1994)
17. Fortin, D., Rudolf, R.: Weak monge arrays in higher dimensions. Discrete Math. 189(1–3), 105–115 (1998)
18. Fukuda, K., Prodon, A.: Double Description Method Revisited. In: Combinatorics and Computer Science (Brest, 1995) (Vol. 1120, pp. 91–111) Lecture Notes in Comput. Sci., Springer: Berlin (1996)
19. Glover, F., Rego, C.: Ejection chain and filter-and-fan methods in combinatorial optimization. 4OR 4(4), 263–296 (2006)
20. Grover, L., Patel, A., Tulsi, T.: A new algorithm for fixed point quantum search (2005)
21. Grover, L.K.: A Fast Quantum Mechanical Algorithm for Database Search, ACM, New York (1996)
22. Han, K.-H., Kim, J.-H.: Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans. Evol. Comput. 6(6), 580–593 (2002)
23. Han, K.-H., Kim, J.-H.: Quantum-inspired evolutionary algorithms with a new termination criterion, h_ϵ gate, and two phase scheme. IEEE Trans. Evol. Comput. 8(2), 580–593 (2004)
24. Jaeger, F.: On four-weight spin models and their gauge transformations. J. Algebraic Combin. 11(3), 241–268 (2000)
25. Kac, V., Cheung, P.: Quantum Calculus. Universitext, Springer, New York (2002)
26. Kauffman, L.H., Lomonaco, S.J.: Entanglement criteria – Quantum and topological. New J. Phys. 4, 73.1–73.18 (electronic) (2002)
27. Kauffman, L.H., Lomonaco, S.J.: Braiding operators are universal quantum gates (2004)
28. Kauffman, L.H., Lomonaco, S.J.: Quantum knots (2004)
29. Khalfin, L.A., Tsirelson, B.S.: Quantum/classical correspondence in the light of Bell's inequalities. Found. Phys. 22(7), 879–948 (1992)
30. Kitaev, A.Yu., Shen, A.H., Vyalyi, M.N.: Classical and Quantum Computation. Graduate Studies in Mathematics (Vol. 47). American Mathematical Society, Providence, RI (2002) Translated from the 1999 Russian original by Lester J. Senechal.
31. Kitaev, A.Y.: Quantum measurements and the abelian stabilizer problem (1995)
32. Li, D., Huang, H., Li, X.: The fixed-point quantum search for different phase shifts (2006)
33. Marinescu, D., Marinescu, G.D.: Approaching Quantum Computing. Prentice Hall (2004)
34. Martí, R., Laguna, M., Glover, F.: Principles of scatter search. Eur. J. Oper. Res. 169(2), 359–372 (2006)
35. Monge, G.: déblai et remblai. Mémoires de l'Académie des sciences, Paris (1781)
36. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Rev. 41(2), 303–332 (electronic) (1999)
37. Sica, L.: Bell's inequalities i: An explanation for their experimental violation (2001)

38. Sica, L.: Correlations for a new Bell's inequality experiment. Found. Phys. Lett. 15(5), 473–486 (2002)
39. Vasquez, M., Hao, J.-K.: A hybrid approach for the 0–1 multidimensional knapsack problem (2001)
40. Werner, R.F., Wolf, M.M.: Bell inequalities and entanglement (2001)
41. Yost, D.: The Johnson-Lindenstrauss space. Extracta Math. 12(2), 185–192 1997. II Congress on Examples and Counterexamples in Banach Spaces (Badajoz, 1996)
42. Younes, A., Rowe, J., Miller, J.: Quantum search algorithm with more reliable behaviour using partial diffusion (2003)

On the Pasture Territories Covering Maximal Grass

Haltar Damba[1], Vladimir M. Tikhomirov[2], and Konstantin Y. Osipenko[3]

[1] Institute of Mathematics, National University of Mongolia
ubu@mongol.net
[2] Department of General Problems of Control, Faculty of Mechanics and Mathematics, Moscow state University
vmtikh@googlemail.com
[3] Department of Mathematics, "MATI" - Russian State Technological University
Kosipenko@yahoo.com

Summary. In this chapter we consider the so-called pasture territory problem, its basic elements, and some related extremal problems. We describe the pasture territory as a graph of a piecewise smooth and continuous function $f(x, y)$ defined on a closed, connected domain of a plane. Considering extremal problems is related with finding the location of the nomadic residence, when the exploiting pasture territory has maximum grass mass, and finding the bound of the territory, when the place of the residence is fixed [1, 2, 5].

Key words: pasture territory, herbage density, piecewise smoothness, non-negative measure, watering place, closure of a set, upper semi-continuity, convexity

1 Main Concepts and the Problem Definition

Let $K \subseteq \Re^2$ be a closure of an open and connected set with a piecewise smooth boundary. Suppose that K consists of a union of a finite number of domains K_i with piecewise smooth boundaries. Then the pasture surface is defined as a graph of a continuous function $f\colon K \to R$ such that $f(x, y)$ is twice differentiable on the interior of K_i for any i.

We define the watering place for the herd as a closure of a set $W \subseteq f(K)$ with an empty interior. That means pasture surface does not contain the interior of the water resource [1, 2].

We denote a closed set $Q \subseteq f(K)$ as the possible locations for the nomadic residence.

Theorem 1. *Between any two points in $f(K)$, there exists a curve of minimal length (minimal curve) in $f(K)$ connecting them.*

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_21,

Proof. Suppose $O_1, O_2 \in f(K)$, $O_1 \neq O_2$. Since the connectedness of K, it follows that the points $f^{-1}(O_1)$ and $f^{-1}(O_2)$ can be connected by a rectifiable planar curve l. Then $f(l)$ is also a rectifiable surface curve with a length d. Let us construct a planar disk $B(f^{-1}(O_1), d) := \{z \in \Re^2 |\ \|z - f^{-1}(O_1)\|_2 \leq d\}$ with a center $f^{-1}(O_1)$ and a radius d. Then the graph $f(B(f^{-1}(O_1), d) \cap K)$ is a complete metric space with a surface metric. This space, evidently, contains the curve $f(l)$ and the point O_2. Hence, by Theorem 3 (p. 112) of [3], there exists a minimal surface curve connecting O_1 and O_2.

If the nomadic residence is located at the point $O \in Q$, we define the maximal possible exploiting area $A_r(O, \overline{W}) \subseteq f(K)$ as the union of all points $M \in f(K)$ such that there exists a loop $l \subseteq f(K)$ of length no more than $2r$ passing through the points M, O and some point $N \in \overline{W}$. This means, for a day, while grazing and watering one's livestock, the herdsman must pass the distance no more than $2r$. The $r > 0$ is called the radius of grazing. It is clear that $A_r(O, \overline{W})$ is a connected compact set[1, 2].

Pasture surface $f(K)$ is a complete metric space, where the distance $\rho_1(M, N)$ for the points $M, N \in f(K)$ is equal to the length of a minimal curve connecting them. This metric ρ_1 is called a surface metric. Any minimal curve consists of possible pieces of the boundary $\partial f(K)$ and some geodesics.

Surface ellipse $E_r(O_1, O_2)$ with focuses $O_1, O_2 \in f(K)$ is a compact set satisfying

$$\rho_1(O_1, M) + \rho_1(O_1, O_2) + \rho_1(M, O_2) \leq 2r,\ \forall M \in E_r(O_1, O_2).$$

Each shoot of the boundary $\partial E_r(O_1, O_2)$ is a closed curve.

When $\rho_1(O_1, O_2) = r$, int $E_r(O_1, O_2) = \emptyset$.
When $\rho_1(O_1, O_2) < r$, int $E_r(O_1, O_2) \neq \emptyset$.

We denote by $W_r(O)$ the subset of $\overline{W}$ such that

$$W_r(O) := \{N \in \overline{W} |\ \rho_1(O, N) \leq r\}.$$

Assume that $\rho_1(O, N) < r$ for any $N \in W_r(O)$. Then the next theorem holds.

Theorem 2. *The boundary* $\partial(\overline{\text{int } A_r(O, \overline{W})})$ *is a union of a finite number of closed, rectifiable curves.*

Proof. Since

$$A_r(O, \overline{W}) = \bigcup_{N \in W_r(O)} E_r(O, N),$$

the boundary $\xi(0) = \partial A_r(O, \overline{W})$ consists of $\partial(\overline{\text{int } A_r(O, \overline{W})})$ and some possible shoots. It is clear that $\overline{\text{int } A_r(O, \overline{W})}$ is a union of a family of ellipses $E_r(O, N),\ N \in W_r(O)$, where $\overline{\text{int } E_r(O, N)} = E_r(O, N)$. Since $\overline{\text{int } A_r(O, \overline{W})}$

is a compact set, we can choose some ellipses $E_r(O, N_1), ..., E_r(O, N_k)$ covering $\overline{\text{int } A_r(O, \overline{W})}$ in union. As each $\partial E_r(O, N_i), i = \overline{1, k}$ is a union of a finite number of closed and rectifiable curves, $\partial(\overline{\text{int } A_r(O, \overline{W})})$ also is a union of a finite number of closed and rectifiable curves.

Corollary 1. *When there exists only a finite number of points $N_i \in W_r(O)$ satisfying $\rho_1(O, N_i) = r$ and total length of the shoots of $\partial A_r(O, \overline{W})$ is finite, the boundary $\xi(O) = \partial A_r(O, \overline{W})$ has a finite length.*

Herbage density is a non-negative measure $\mu(K)$ such that for any compact set $M \subseteq K$,

$$\mu(M) < \infty$$

and the charge $Z(A)$ generated by bounded function $g(x, y) = \sqrt{1 + f_x^2 + f_y^2}$:

$$Z(A) = \int\limits_A \sqrt{1 + f_x^2 + f_y^2}\, d\mu$$

is absolutely continuous, where $A \subseteq K$ is any measurable subset with respect to μ [3](p. 331).

The main maximization problem for nomads is to find the best place for the residence, i.e.,

$$G(O) = \int\limits_{f^{-1}(A_r(O,\bar{W}))} \sqrt{1 + f_x^2 + f_y^2}\, d\mu = \int\limits_{A_r(O,\bar{W})} d\mu \to \max;\ O \in Q. \quad (1)$$

This problem is considered very difficult because of defining the boundary of $A_r(O, \overline{W})$.

Suppose that ξ and η are any two continuous curves on $f(K)$. Let us construct a metric space Ξ of all continuous curves on $f(K)$ by defining the distance as

$$\rho(\xi, \eta) = \inf \rho(f_1, f_2).$$

Here, the lower bound is taken by all admissible pairs of parametric representations for ξ and η which are continuous functions $f_1(t)$ and $f_2(t)$ $(0 \le t \le 1)$, and the distance between functions f_1 and f_2 is defined as

$$\rho(f_1, f_2) = \sup_{0 \le t \le 1} \rho(f_1(t), f_2(t)).$$

Lemma 1. *Suppose that $\xi = \bigcup\limits_{i=1}^{k} \xi^i$, where each ξ^k is a closed and continuous curve on $f(K)$, and Π_ξ is a side view of the surface piece defined by ξ. Then the function*

$$S(\xi) = \int\limits_{\Pi_\xi} \sqrt{1 + f_x^2 + f_y^2}\, d\mu$$

is upper semi-continuous in Ξ^k.

Proof. Consider a sequence $\xi_n = \bigcup_{i=1}^{k} \xi_n^i$, where each sequence ξ_n^i converges to ξ^i with respect to the above metric in Ξ. If we denote

$$S(\eta_n) = \sup_{i \geq n} S(\xi_i) \text{ with } \inf_{\xi_j \in \bigcup\limits_{i \geq n} \xi_i} \rho_{\max}(\eta_n, \xi_j) = 0,$$

$$\rho_{\max}(\xi_1, \xi_2) = \max_{1 \leq i \leq k} \rho(\xi_1^i, \xi_2^i),$$

then we have

$$S(\eta_n) = S(\xi) + \int\limits_{\Pi_{\eta_n} \setminus (\Pi_{\eta_n} \cap \Pi_\xi)} \sqrt{1 + f_x^2 + f_y^2}\, d\mu - \int\limits_{\Pi_\xi \setminus (\Pi_{\xi_0} \cap \Pi_\xi)} \sqrt{1 + f_x^2 + f_y^2}\, d\mu.$$

The first integral tends to zero, but the second integral tends to -$\mu(\xi_0)$, where ξ_0 is a piece of the curve ξ. Therefore, $S(\xi) \geq \overline{\lim}_{n \to \infty} S(\xi_n)$ and the lemma is proved.

For any $O \in f(K)$, we introduce a notation $O^p = f^{-1}(O)$.

Theorem 3. *Function* $G(O)$ $(G(O^p))$ *given in* (1) *is upper semi-continuous on* $f(K)$ (K).

Proof. Let a sequence $O_n^p \to O^p$ in K. Then the sequence $O_n = f(O_n^p)$ also tends to O in $f(K)$. Suppose that ξ_n is a boundary of $A_r(O_n)$ consisting of k closed continuous curves. It is clear that $O_n \to O$ $(O_n^p \to O^p)$ implies $\xi_n \to \xi$. By previous lemma, the function $G(O)$ $(G(O^p))$ is also upper semi-continuous.

Corollary 2. *If* $Q(f^{-1}(Q))$ *is compact, then problem* (1) *has a solution on* $Q(f^{-1}(Q))$.

In the next two parts of this chapter, we assume that f is a linear function.

2 On the Forms of Exploiting Pasture Territories in Simple Cases

At first, we assume that the pasture territory is $\overline{\Re^2 \setminus B(O_1, R)}$, where $B(O_1, R)$ is a disk generated by circle $C(O_1, R)$ presenting the watering place $\overline{W}$. Our goal is to define the exploiting area $A_r(O, \bar{W})$ in cases of $R = 0$ (a well), $R = \infty$ (a straight bank of a river or a straight brook), and $0 < R < \infty$ (a bank of a deep lake). In all cases we assume that the nomadic residence is located at distance $k < r$ from the watering place [5].

We study each case, separately.

1. Suppose $R = 0$. In this case, $\overline{W}$ consists of unique point. Let this point be $O_1(0, -\frac{k}{2})$ and the nomadic residence is located at the point $O(0, \frac{k}{2})$ (Fig. 1a). The pasture territory is the whole plane. It is clear that the maximal exploiting area $A_r(O, \bar{W})$ is an ellipse given by the inequality

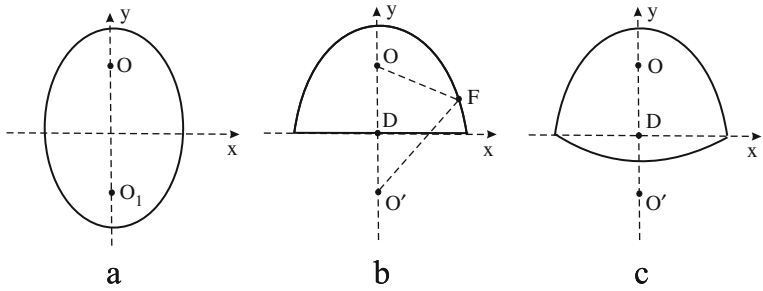


Fig. 1. Pasture territories when R = 0

$$\sqrt{x^2 + \left(\frac{k}{2} - y\right)^2} + \sqrt{x^2 + \left(\frac{k}{2} + y\right)^2} \leq 2r - k.$$

2 Suppose $R = \infty$. Let us consider two cases.
a. A bank of a river. In this case, the pasture territory is a half plane, where the livestock cannot cross the river and the maximal exploiting area $A_r(O, \bar{W})$ is a semi-ellipse bounded by a straight bank of a river. In fact, if we assume that the nomadic residence is located at the point $O(0, k)$, $D(0, 0)$ is the origin of coordinates (Fig. 1b), and denote $O'(0, -k)$, then for any point $F(x, y)$ of the curve $\partial A_r(O, \bar{W})$ the following equality holds:

$$\rho(O, F) + \rho(O', F) = 2r.$$

Hence, we have the following inequalities for the exploiting area $A_r(O, \bar{W})$:

$$\sqrt{x^2 + (k-y)^2} + \sqrt{x^2 + (k+y)^2} \leq 2r, \ y \geq 0 \ .$$

b. A brook. In this case, the livestock can cross the brook and the pasture territory is the whole plane. The part of the maximal exploiting area $A_r(O, \bar{W})$ on the other side of the brook is a segment of a disk with a center $O(0, k)$ and a radius r (Fig. 1c). Therefore, $A_r(O, \bar{W})$ is defined as follows:

$$\begin{cases} \sqrt{x^2 + (k-y)^2} + \sqrt{x^2 + (k+y)^2} \leq 2r, \ y \geq 0, \\ x^2 + (k-y)^2 \leq r^2, \ y < 0 \ . \end{cases}$$

3. Suppose $0 < R < \infty$. In this case, the pasture territory is the closure $\overline{R^2 \setminus B(O_1, R)}$ of the complement of the disk $B(O_1, R)$ on the plane. We suppose that $B(O_1, R)$ is a deep like. Without losing generality, we assume that $R = 1$ and O_1 is the origin of coordinates. Then the nomadic residence $O(0, m)$ is located at distance $m = k+1$ from the origin of coordinates (Fig. 2). Clearly, the boundary curve $\xi(O)$ of the maximal exploiting area $A_r(O, \bar{W})$ is closed and symmetric with respect to the ordinate. Depending on values of m and r, the boundary curve $\xi(O)$ has different forms.

Theorem 4. i. *If $r \leq \sqrt{m^2 - 1}$, then the lower part of $\xi(O)$ is an arc F_1F_2 and its upper part is an envelope Γ of the family of ellipses with focuses O and E:*

$$\rho(O, E) + \rho(O, M) + \rho(E, M) = 2r, \tag{2}$$

where E is a point on the arc F_1F_2 ($OF_1 = OF_2 = r$) and M is a point on the envelope (Fig. 2a).
ii. *If $\sqrt{m^2 - 1} + (\frac{\pi}{2} + \arcsin\frac{1}{m}) \geq r > \sqrt{m^2 - 1}$, then the upper part of $\xi(O)$ is the same as the previous envelope Γ generated by (2). The lower part of $\xi(O)$ is an arc C_1B_1BC, where B and B_1 are the contact points of tangents from O to $C(O_1, 1)$. But the middle two parts of $\xi(O)$ are generated by the endpoints of minimal curves of length r starting from O and without passing the interior of disk $B(O_1, 1)$ (Fig. 2b).*
iii. *If $r > \sqrt{m^2 - 1} + \frac{\pi}{2} + \arcsin\frac{1}{m}$, then the upper part of $\xi(O)$ is the same envelope Γ generated by (2). But the lower part of $\xi(O)$ is generated by the endpoints of minimal curves of length r as in the previous case (Fig. 2c).*

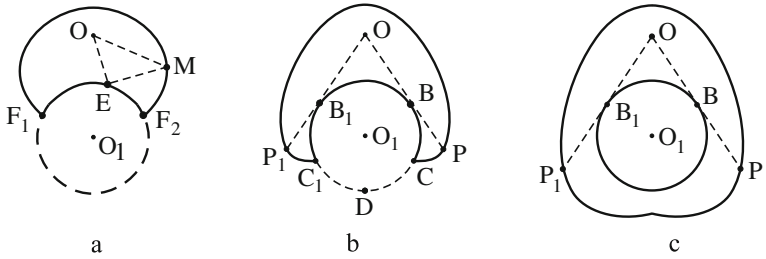


Fig. 2. Pasture territories when $0 < R < \infty$

Proof. The length of tangents OB and OB_1 is equal to $\sqrt{m^2 - 1}$. When $r \leq \sqrt{m^2 - 1}$, the boundary curve $\xi(O)$ of the exploiting area $A_r(O, \bar{W})$ must contain any point M satisfying (2). Therefore, the upper part of $\xi(O)$ must be an envelope Γ, and for any point $M(x, y)$ of Γ, the segments OE and ME have the same reflection angle to the circle $C(O_1, 1)$ at the point E.

Denote $\angle OO_1E = \alpha$. Then after some simple calculations we have the following parametric system for the envelope Γ (Fig. 2a):

$$\begin{cases} (x\cos\alpha - y\sin\alpha)(m\cos\alpha - 1) - (x\sin\alpha + y\cos\alpha - 1)m\sin\alpha = 0, \\ (1 + \frac{x}{m}ctg\alpha - \frac{y}{m})\sqrt{1 + m^2 - 2m\cos\alpha} - 2r - \\ -\sqrt{x^2 + y^2 + 1 - 2y\cos\alpha - 2x\sin\alpha} = 0. \end{cases} \tag{3}$$

The endpoints of envelope Γ are the points F_1 and F_2 of arc BB_1 such that $OF_1 = OF_2 = r$. The first part of the theorem is proved.
The curve consisting of tangent OB (OB_1) and arc BD (B_1D) has a length of

$$\sqrt{m^2 - 1} + \frac{\pi}{2} + \arcsin\frac{1}{m}.$$

Therefore, when

$$\sqrt{m^2 - 1} + \frac{\pi}{2} + \arcsin\frac{1}{m} \geq r > \sqrt{m^2 - 1}\ ,$$

the system (3) expresses only the top part of the boundary curve $\xi(O)$. The endpoints of this part coincide with the ends of tangent OP and OP_1(Fig. 2b). But the lower ends of the boundary curve $\xi(O)$ are located at points C and C_1 of the circle $C(O_1, O)$, where the sum of lengths of arc BC (B_1C_1) and tangent OB (OB_1) is equal to r. Thus, any point of the boundary curve $\xi(O)$ locating between C (C_1) and P (P_1) is defined by the endpoints of the minimal curve of length r. For constructing such curves, we use Cruggs's theorem on the shortest curves with barriers [4]. This theorem claims that the shortest path consists from tangents and geodesics on barrier sets. By this theorem the minimal curve consists of two tangents and an arc. A simple calculation shows that the coordinates of the endpoint satisfy

$$\begin{aligned} &\sqrt{m^2 - 1} + \sqrt{x^2 + y^2 - 1} + \pi - \arccos\tfrac{1}{m} - \arcsin\sqrt{\tfrac{x^2+y^2-1}{x^2+y^2}} \\ &- \arcsin\sqrt{\tfrac{x^2}{x^2+y^2}} = r \end{aligned} \ . \tag{4}$$

The second part of the theorem is proved.
When

$$r > \sqrt{m^2 - 1} + \frac{\pi}{2} + \arcsin\frac{1}{m},$$

the boundary curve $\xi(O)$ consists of an inner part which is the circle $C(O_1, 1)$ (Fig. 2c) and an outer part of which any point satisfies either (3) or (4).

Now let us consider the case where the pasture territory is the upper half plane with a half-disk of radius R ($R \leq r$). Assume that the nomadic residence is located at the origin of coordinates O (Fig. 3). We also assume that the boundary curve of the pasture territory is the watering place.

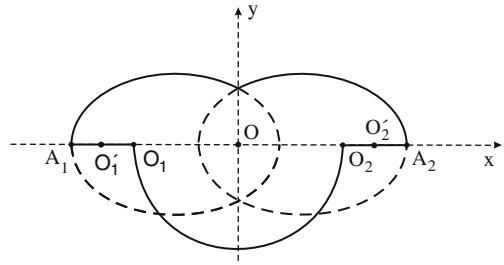


Fig. 3. Pasture territories as the upper half plane with a half-disk of radius R(R ≤ r)

When $R = r$, the maximal exploiting area $A_r(O, \overline{W})$ is a half-disk:

$$x^2 + y^2 \leq R^2,\ y \leq 0.$$

Our goal is to find the useful pasture territory, namely its boundary curve, when $r > R$.

Proposition 1. *Let O_1 and O_2 be the two points of circle $C(O, R)$ on the abscissa. Then for $r > R$ the upper bound of the maximal exploiting area $A_r(O, \bar{W})$ is defined as a union of the upper bounds of ellipses $E_r(O, O_i)$:*

$$| OO_i | + | OM | + | MO_i | \leq 2r,\ M \in \Re^2_+,\ i = 1, 2.$$

But the lower bound consists of two sections O_1A_1, O_2A_2 and the semicircle O_1DO_2, where $| OA_1 | = | OA_2 | = r$.

Proof. It is clear that the ellipse $E_r(O, O_i)$ contains the ellipse $E_r(O, O'_i)$, where O'_i is an arbitrary point on A_iO_i satisfying $|A_iO'_i| < |A_iO_i|$, $i = 1, 2$. Also, the set $(E_r(O, O_1) \cup E_r(O, O_2)) \cap A_r(O, \bar{W})$ contains $E_r(O, N) \cap A_r(O, \bar{W})$ for any point N of the arc O_1DO_2 of circle $C(O, R)$. Hence, the upper part of $\partial(E_r(O, O_1) \cup E_r(O, O_2))$ is also the upper part of $\xi(O)$. The proposition is proved.

3 Some Solution Properties of the Main Maximization Problem on the Plane

We suppose that $f(K) = Q = \overline{W} = \Re^2$ and the herbage density $\mu(\Re^2)$ has a positive Lebesgue measure only for some closed and connected set $M \subseteq \Re^2$ (or $\mu(\Re^2 \setminus M) = 0$) satisfying $\overline{\text{int } M} = M$. In this case, evidently, $A_r(O, \overline{W}) = B(O, r)$. Let Q^*_r be the set of solutions of the problem (1). It is required to define the set $Q^*_r \subseteq Q$ as the area $S(M \cap B(O, r))$ is maximal for any $O \in Q^*_r$. Denote r_M by the maximal radius of the inscribed circles contained in M and R_r by the minimal radius of the described circles containing M.

Lemma 2. *Suppose that M is a closed and convex set and O_1O_2 is a closed interval. Then for any $r > 0$ and for any $O \in O_1O_2$, the following inequality holds* [2]*:*

$$S(B(O,r) \cap M) \geq \min(S(B(O_1,r) \cap M), S(B(O_2,r) \cap M)),$$

where we denote S as the area of a domain.

Proof. Since $O \in O_1O_2$, there exists a number $\alpha \in [0,1]$ such that

$$O = \alpha O_1 + (1-\alpha)O_2 \text{ and } B(O,r) = \alpha B(O_1,r) + (1-\alpha)B(O_2,r).$$

By convexity of M, we have

$$(\alpha B(O_1,r) \cap M) + ((1-\alpha)B(O_2,r) \cap M) \subseteq B(O,r),$$
$$(\alpha B(O_1,r) \cap M) + ((1-\alpha)B(O_2,r) \cap M) \subseteq M,$$

and this implies

$$(\alpha B(O_1,r) \cap M) + ((1-\alpha)B(O_2,r) \cap M) \subseteq (B(O,r) \cap M).$$

Using the Brunn–Minkowski inequality [6], we have

$$\begin{aligned}\sqrt{S(B(O,r) \cap M)} &\geq \sqrt{S(\alpha B(O_1,r) \cap M + (1-\alpha)B(O_2,r) \cap M)} \\ &\geq \alpha\sqrt{S(B(O_1,r) \cap M)} + (1-\alpha)\sqrt{S(B(O_2,r) \cap M)} \\ &\geq \min\left(\sqrt{S(B(O_1,r) \cap M)}, \sqrt{S(B(O_2,r) \cap M)}\right)\end{aligned}$$

and the lemma is proved.

Theorem 5. *Following statements hold*

1a. *If $r < r_M$ or $r > R_M$, then* int $Q_r^* \neq \emptyset$.
1b. *If $r = r_M$ or $r = R_M$, then* int $Q_r^* = \emptyset$ *and Q_r^* consists of unique point.*
2. *If $r > R_M$ or M is convex, then Q_r^* is convex and compact.*
3. *If M is a simply connected set, then* int $Q_r^* = \emptyset$ *for $r_M < r < R_M$.*

Proof. Statements 1a, 1b and statement 2 in case $r > R_M$ are evident. Statement 2 follows from Lemma 2 when M is convex.
Now we consider statement 3. We have

$$\pi r_M^2 < S(B(O,r) \cap M) < \pi R_M^2,$$

for any $O \in Q_r^*$. On the contrary, we assume that int $Q_r^* \neq \emptyset$. Then there exists a small scalar $\varepsilon > 0$ for every point $O \in$ int Q_r^* such that $S(B(O',r) \cap M)$ is constant for any $O' \in B(O,\varepsilon)$.
Hence, it follows that either the ring $(B(O,r+\varepsilon) \setminus B(O,r-\varepsilon))$ consists of points of M or $M \cap (B(O,r+\varepsilon) \setminus B(O,r-\varepsilon)) = \emptyset$. In first case, from the simply connectedness of M it follows that $r < r_M$ which contradicts to $r > r_M$. In second case, from the connectedness of M it follows that $r > R_M$ which contradicts to $r < R_M$. The theorem is proved [1].

Note that for statement 3, the simply connectedness of M is necessary. In fact, if M is a ring $B(O,R) \setminus (\text{int } B(O,R_1))$, where $R > R_1$ and $\frac{R+R_1}{2} < r - R$, then Q_r^* contains $B(O, R-r)$ so that int $Q_r^* \neq \emptyset$.

Now we consider some primary propositions which may be useful. It is clear that the function

$$g(r) = \max_{O \in R^2} S(B(O,r) \cap M)$$

is strongly increasing on the interval $[r_M, R_M]$.

Proposition 2. *If M is convex, then $Q_r^* \subseteq M$ for any $r \in [0, R_M]$.*

Proof. The statement of proposition in case of $r \in [0, r_M]$ is evident. We consider the case when $r \in (r_M, R_M]$. Suppose $O \in Q_r^*$ and $O \notin M$. Then there exists $O_1 \in Q_r^*$ such that $\rho(O, O_1) = \min_{A \in M} \rho(O, A)$. Passing through O_1, we can construct a line separating M and O and perpendicular to the straight line OO_1. It is clear that $B(O,r) \cap M$ is included in int $B(O_1, r)$. Then, there exists $\varepsilon > 0$ such that $(B(O,r) \cap M) \subseteq B(O_1, r-\varepsilon)$. Therefore,

$$g(r-\varepsilon) \geq S(B(O_1, r-\varepsilon) \cap M) \geq S(B(O,r) \cap M) = g(r),$$

which contradicts to the strongly monotonicity of $g(r)$.

When $r \geq R_M$, then the nomadic residence must be located at the point O which is the center of the minimal circle describing M.

Proposition 3. *O is either the center of the describing circle of an acute triangle $\triangle ABC$, where $A, B, C \in C(O, R_M) \cap M$, or the middle point of the diameter of M.*

Proof. If $C(O, R_M) \cap M$ contains some acute triangle, then O indeed coincides with the center of the describing circle of this triangle. Otherwise, there exists a half-disk including $C(O, R_M) \cap M$. If the both ends of the diameter of this half-disk do not belong to M, then by moving O slightly we can obtain another circle $B(O_1, R_1)$ $(R_1 < R_M)$ containing M. This contradicts to the fact that R_M is the radius of the minimal circle describing M.

Now we assume that the possible location Q for the nomadic residence is a line l and $R_M \leq r$. Let O be the center of the describing circle $C(O, R_{M,l})$ $(R_{M,l} \geq R_M)$ of M. We consider a line η which is perpendicular to l and passes through O. This line η separates $C(O, R_{M,l})$ into two parts: $C^+(O, R_{M,l})$ and $C^-(O, R_{M,l})$, none of which contain an end of the separating diameter.

Proposition 4. *Either there exist two points $A \in C^+(O, R_{M,l}) \cap M$ and $B \in C^-(O, R_{M,l}) \cap M$ or there exists a point $C \in M \cap \eta \cap C(O, R_{M,l})$.*

Proof. If neither A and B nor C exists, then all points of the set $C(O, R_{M,l}) \cap M$ are located on either $C^+(O, R_{M,l})$ or $C^-(O, R_{M,l})$. Therefore, by moving O to $O_1 \in l$ slightly, we can construct a disk $B(O_1, R_1)$ satisfying $M \subseteq B(O_1, R_1), R_1 < R_{M,l}$. This contradicts to the fact that $R_{M,l}$ is the radius of the minimal describing circle of M with a center belonging to l.

Now, again we assume that $Q = \Re^2,\ r_M \leq r \leq R_M$.

Theorem 6. *Let $O \in Q_r^*$ and M be a triangle or a diagonally symmetric convex quadrangle or any regular convex polygon. Then there exists a number $r_{\max} \leq R_M$ such that for any r, $r_M < r < r_{\max}$ the ratio of the chord generated by $C(O, r) \cap M$ and the length of the corresponding side is constant.*

Proof. The statement of the theorem for regular convex polygon is evident because $O \in Q_r^*$ is the center of polygon, where $r_{\max} = R_M$.

Let M be a triangle. Assume that a triangle $\triangle ABC$ with edges a, b, and c is given, and its largest angle is $\angle ABC$. Let a circle with radius r is given.

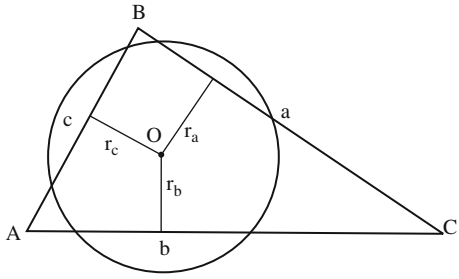


Fig. 4. Pasture territories where M is a triangle

We denote by r_a, r_b, and r_c distances measured from the center O of the circle to edges a, b, and c of the triangle, respectively, where $r \leq \min\{OB, OA, OC\}$ (Fig. 4). We construct the following Lagrange function:

$$L(r_a, r_b, r_c, \lambda) = r^2 \arccos \frac{r_a}{r} + r^2 \arccos \frac{r_b}{r} + r^2 \frac{r_c}{r} - \sqrt{r^2 - r_a^2} r_a$$
$$- \sqrt{r^2 - r_b^2} r_b - \sqrt{r^2 - r_c^2} r_c + \lambda(r_a a + r_b b + r_c c - a - b - c)$$

and consider the maximization problem

$$L(r_a, r_b, r_c, \lambda) \rightarrow \max,$$
$$0 < r_a,\ 0 < r_b,\ 0 < r_c.$$

By Lagrange rule, the partial derivatives of the Lagrange function are equal to zero, we obtain

$$\frac{2\sqrt{r^2-r_a^2}}{a}=\frac{2\sqrt{r^2-r_b^2}}{b}=\frac{2\sqrt{r^2-r_c^2}}{c}=\lambda.$$

If $\angle ABC \le \frac{\pi}{2}$, then $r_{\max}=R_M$, otherwise $r_{max}<R_M$ and $r_{max}=\mathrm{OB}$.

Now, let us consider a diagonally symmetric quadrangle $ABCD$ with edges $AB=AD=a$, $BC=DC=b$. Clearly, the center O of the maximal circle with radius r always lies on the axis of symmetry AC, and the Lagrange function for this circle has the following form:

$$L(r_a,r_b,\lambda)=2r\big(\arccos\frac{r_a}{r}+\arccos\frac{r_b}{r}\big)$$
$$-2\sqrt{r^2-r_a^2}\cdot r_a-2\sqrt{r^2-r_b^2}\cdot r_b+2\lambda(ar_a+br_b-a-b).$$

Corresponding Lagrange problem is

$$L(r_a,r_b,\lambda)\ \to\ \max,$$
$$r_a>0,\ r_b>0.$$

By Lagrange rule, the partial derivatives of the Lagrange function are equal to zero, we obtain

$$\frac{2\sqrt{r^2-r_a^2}}{a}=\frac{2\sqrt{r^2-r_b^2}}{b}=\lambda.$$

If BD < AC, then $r_{\max}=OB$. But, if BD ≥ AC, then

$$r_{\max}=\begin{cases}\mathrm{OC},\ \text{if}\angle BAD\le\angle BCD,\\ \mathrm{OA},\ \text{if}\angle BAD>\angle BCD\end{cases}.$$

The proof is completed.

4 Conclusion

Nowadays, the world civilization is divided into two forms: settled and nomadic. The nomadic civilization is closely connected with the nature, and ecological and economical problems of nomads are regulated simultaneously. Therefore, research activities in this field are increasing more and more, and many international conferences are being organized every year.

Mongolia is one of the few countries where the nomadic civilization still exists in classical form. Fifty percent of the population is involved somehow in stock nomadic breeding. Since Mongolian has extreme climate, it is very important for nomads to determine optimal choices for roaming places, i.e., the location for the nomadic residence depending on the seasons. While the settled civilization is well studied and modeled mathematically, the study of the nomadic civilization is practically ignored and less. Therefore, our work may be regarded as new in mathematical modeling.

In this work, we consider extremal problems on pasture surface, define its basic elements, present and solve the problem of determining optimal locations for the nomadic residence, and prove some related and existence theorems. This research is realized within the Russia–Mongolian joint grant "Economic and geometry extremal problems on equipped surfaces."

We have used mathematical apparatus such as geometry, functional analysis, and theory of extremal problems in our study.

References

1. Khaltar, D.: Some mathematical problems on fasture surface. Sci. Trans. Nat. Univ. Mong, 8(186), 91–105 (2001)
2. Khaltar, D.: The pasrural geometry and it's extremal problems. J. Mong. Math. Soc. 28, 38–43 (1998)
3. Kolmogorov, A.N., Fomin S.V.: Elements of Functions and Functional Analysis Theory (Rus.), (pp. 187) Nauka, Moscow (1972)
4. Craggs, J.M.: Calculus of Variations, Allen and Unwin. London (1943)
5. Haltar, D., Itgel, M.: Boundary curves of exploiting pasture territories. Sci. Trans. Nat. Univ. Mong 7(168), 10–18 (2000)
6. Tikhomirov, V.M.: Stories on Maximimum and Minimum(Rus.) (pp. 188) Nauka, Moscow (1986)

On Solvability of the Rate Control Problem in Wired-cum-Wireless Networks

Won-Joo Hwang[1], Le Cong Loi[2], and Rentsen Enkhbat[3]

[1] Department of Information and Communications Engineering, Inje University
ichwang@inje.ac.kr
[2] Department of Information and Communications Engineering, Inje University
loilc@vnu.edu.vn
[3] School of Mathematics and Computer Sciences, National University of Mongolia
renkhbat46@ses.edu.mn

Summary. In a wired-cum-wireless network, the rate control problem is a difficult optimization problem. This chapter addresses the solvability of the optimization problem, where the optimization variables are both end-to-end session rates and wireless link transmission rates. The convergence of all algorithms on solving the rate control problem in wireless or wired-cum-wireless networks has been shown in [2, 5, 8–11]. But existence of a unique solution in the problem has not been studied so far. Although the problem is a nonconvex optimization problem, the unique solvability of the end-to-end session rates of the problem has been shown. In addition, we also prove that there exist infinitely many corresponding values of the wireless link transmission rates which are optimal solutions of the rate control problem. Simulation results are provided to illustrate our approach.

Key words: wired-cum-wireless networks, rate control problems, convex optimization problems, nonconvex optimization problems, convex functions, concave functions

1 Introduction

We consider the wired-cum-wireless networks with CSMA/CA-based wireless LANs, which extend a wired backbone and provide access to mobile users. Wireless LANs provide sufficient bandwidth for office applications with relatively limited mobility, and typically the users may roam inside a building or campus. Wireless LANs help extend wired networks when it is impractical or expensive to use cabling. In a wired-cum-wireless network, mobile hosts (MHs) can roam in a wireless network, called basic service sets (BSSs), which are attached at the periphery of a wired backbone. The wired infrastructure can be an IEEE 802 style Ethernet LAN or some other IP-based network. The wired and wireless networks are interconnected via access points (APs), which

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_22,

are actually fixed base stations that provide interfaces between the wired and wireless parts of the network and control each BSS. For example, a typical wired-cum-wireless network is shown in Fig. 1.

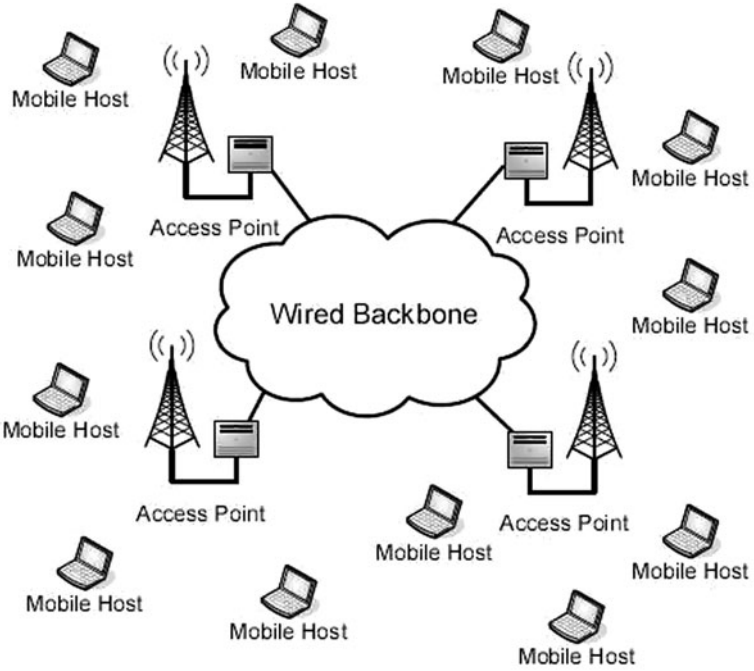


Fig. 1. Architecture of wired-cum-wireless network

Congestion control in the network is an extensively researched topic. The objective of rate control is to provide proportional fairness among the end-to-end sessions in the network. The problem of rate control has been extensively studied, e.g., in [2–7, 9–11]. It is well known that in wired networks [3, 4, 7], based on convex programming, globally fair rates are unique and attainable via distributed approaches.

In wireless networks, the capacity is not a fixed quantity. For example, in code-division multiple-access wireless networks, transmit powers can be controlled to induce different signal-to-interference ratios on the links, changing the attainable throughput on each link [2]. Unlike [2], in [9], authors have formulated the rate control problem in multi-hop wireless networks with random access, where the attainable throughput on each link depends on the attempt probabilities on all links. The rate control problems in [2, 9] are nonconvex optimization problems.

In wired-cum-wireless networks [5, 6, 10, 11], similar in wireless networks, the capacity of a wireless link is not a fixed quantity and depends on wireless link transmission rates. End-to-end session rates are also attainable by solving

a nonlinear programming using the dual-based (DB) or the primal–dual interior-point (PDIP) algorithms. However, both papers [5, 10] have only addressed optimal end-to-end session rates while optimal wireless link transmission rates were not their concern. Recently, in [11], the optimal wireless link transmission rates were examined.

Note that, the solvability of the rate control problems in both wireless networks [2, 9] and wired-cum-wireless networks [5, 10, 11] has not been studied. There exist only global convergent algorithms for the problem. This chapter has been motivated by the papers [10, 11]. In this chapter, we focus on the solvability of the rate control problem introduced in [10, 11] in a wired-cum-wireless network. We show that there is a unique optimal solution for the end-to-end session rates, but there may be many corresponding optimal values of the wireless link transmission rates.

This chapter is organized as follows. In Section 2, we survey recent results on the rate control problems in the wired-cum-wireless networks. In Section 3, we discuss the rate control problem as optimization problem. Section 4 is devoted to the solvability of the rate control problem. In Section 5, we illustrate our theoretical results through a discussion of some numerical examples. Finally, all necessary proofs are presented in the Appendix.

2 Related Works

There were several existing works which addressed the problem of rate control in wired, wireless, and wired-cum-wireless networks. In [4, 7], the rate control problem in wired networks was formulated as a convex optimization problem with a rate vector as optimization variable and the constraints are the source rates and fixed link capacities. Under some assumptions on the objective function, their results showed that the problem has a unique optimal solution. Kelly et al. [4] have decomposed the problem into a user sub-problem and a network sub-problem. Furthermore, they have proposed two classes of decentralized algorithm to implement solution to relaxations of the problems, which are network sub-problem and dual of the network sub-problem. In [7], authors have also presented different flow control algorithms to solve the same optimization problem. Kelly [3] has shown that the problem has a unique optimal rate vector, but there may be corresponding values of the flow rates which are optimal solutions.

Recently in [2, 5, 6, 9–11], the rate control problem in a wireless network and in a wired-cum-wireless network has been studied as a nonconvex optimization problem. Chiang [2] studied the rate control problem in wireless multi-hop networks. He considered the problem with elastic link capacities depending on transmit powers and proposed a jointly optimal congestion control for solving a nonlinear programming problem. In [9], Wang et al. discussed the rate control problem in multi-hop wireless networks with random access,

but unlike [2], they examined whether the attainable throughput of wireless links depends only on transmission probabilities and have proposed both penalty-based and dual-based algorithms to find an optimal solution of the rate control problem. In [10, 11], authors have formulated the rate control problem in a wired-cum-wireless network from end-to-end session rates, wireless link transmission rates, and capacities of both wired and wireless links, where capacity of wireless links is elastic and depends on the wireless link transmission rates. The proportional fair rate in the wired-cum-wireless network can be obtained by solving an equivalent convex optimization problem using the DB-distributed algorithm [10, 11] or PDIP algorithm [5]. In order to solve the rate control problem in wired-cum-wireless networks, we need to find both scheduling rates for the wireless links and end-to-end session rates for the wired links. The papers [5, 10, 11] proposed algorithms which converge to global solutions. However, simulation results in [5, 10] only showed the optimal end-to-end session rates for the wired links, but did not show the scheduling rates on the wireless links (see [5, 10] and the references therein). In [11], which is an extended version of [10], both optimal end-to-end session rates and optimal wireless link transmission rates on the wireless links were shown with a unique optimal wireless link transmission rate.

3 The Rate Control Problem

In this section, we introduce briefly the rate control problem in the wired-cum-wireless network (see [10, 11] and the references therein for more details). Consider the wired-cum-wireless network that consists of a set M of all MHs, a set W of CSMA/CA-based BSSs, a set N of fixed nodes in a wired backbone, and a set L of unidirectional links which connect the fixed nodes in the wired backbone. We assume that each MH belongs to one and only one BSS, and each BSS has one and only one AP denoted as $A(s)$. In BSS w, let us denote N_w, E_w, and A_w as a set of nodes, a set of directed edges in that particular BSS w, and an AP for BSS w, respectively. For any node $s \in N_w$, we denote the set of s's out-neighbors $D_s = \{t : (s,t) \in E_w\}$, which represents the set of neighbors to which s is sending traffic and s's in-neighbors $J_s = \{t : (t,s) \in E_w\}$, which represents the set of neighbors from which s is receiving traffic. In our network model, we assume that each node has a single transceiver. A node cannot transmit and receive simultaneously and cannot receive more than one frame at a time. For ease of exposition, we assume that all end-to-end sessions originate and terminate in MHs, and the source and destination MHs of any session belong to different BSSs. Since end-to-end sessions within a BSS are not allowed according to the assumption, an immediate consequence is that all links in a BSS w are between its MHs and the AP A_w. The transmission rate for a wireless link $(s,t) \in E_w$ is denoted as $\rho_{s,t}$ and let $\rho := (\rho_{s,t} : (s,t) \in E_w, w \in W) \in \mathbb{R}_+^{|M|}$ be a vector of transmission rates for all wireless links, where $|M|$ denotes its cardinality. As shown in [8], the capacity of link

$(s,t) \in E_w$ in BSS w, in which either s or t must be the AP A_w is given as

$$c_{s,t}(\rho) = \frac{\rho_{s,t}}{1 + \sum_{k \in D_{A_w}} \rho_{A_w,k} + \sum_{k \in J_{A_w}} \rho_{k,A_w}}. \tag{1}$$

Note that, the second and the third terms in the denominator of formula (1) are the sum of transmission rates on all downlinks and uplinks, respectively, in BSS w.

The wired backbone connects all the APs using the set L of unidirectional wired links whose capacity is c_l, $l \in L$, where c_l is fixed for all $l \in L$. We denote $L(A_w, A_v)$ as a set of wired links that are used for the communication from A_w to A_v and let $S(l) := \{(A_w, A_v) : w, v \in W, l \in L(A_w, A_v)\}$ be a set of communication pairs consisting of APs that use link $l \in L$.

The wired-cum-wireless network is shared by a set S of end-to-end sessions. Each session in S can be expressed as (i,j), where MHs i and j are source and sink of the session, respectively. Let y_{ij} be a session rate for session $(i,j) \in S$. We denote a vector of the end-to-end session rates by $y := (y_{ij} : (i,j) \in S) \in \mathbb{R}_+^{|S|}$. Due to our assumptions the set M of all MHs and the set S of all end-to-end sessions must satisfy $|M| = 2|S|$.

Now we specify the following rate control problem in the wired-cum-wireless network [5, 10, 11]:

$$\begin{aligned}
\text{maximize} \quad & \sum_{(i,j) \in S} \log(y_{ij}), \\
\text{subject to} \quad & y_{ij} \le c_{i,A(i)}(\rho) \quad \forall (i,j) \in S, \\
& y_{ij} \le c_{A(j),j}(\rho) \quad \forall (i,j) \in S, \\
& \sum_{(A(i),A(j)) \in S(l)} y_{ij} \le c_l \quad \forall l \in L, \\
& y_{ij} \ge 0 \quad \forall (i,j) \in S, \\
& \rho_{s,t} \ge 0 \quad \forall (s,t) \in E_w, \quad \forall w \in W,
\end{aligned} \tag{2}$$

where optimization variables are both vector of end-to-end session rates $y := (y_{ij} : (i,j) \in S)$ and vector of wireless link transmission rates $\rho := (\rho_{s,t} : (s,t) \in E_w, w \in W)$, and the capacities of wireless links $c_{i,A(i)}(\rho)$ and $c_{A(j),j}(\rho)$ are given by formula (1). Each session in the network model runs across both wired links which have fixed link capacities and wireless links whose capacities are elastic and depend on the wireless link transmission rate of MHs in that particular BSS. Therefore, the first and the second sets of constraints of problem (2) ensure that the session rates cannot exceed the attainable throughputs of the two wireless links that are traversed. The third set of constraints states that the total session rates on a wired link cannot exceed the capacity of that link. The fourth and the last sets of constraints ensure, respectively, that all the end-to-end session rates and all the wireless link transmission rates are non-negative.

The capacities of the wireless links are not concave functions of the transmission rates ρ. Thus problem (2) is a nonconvex optimization problem. In order to solve it, we can use the DB-distributed algorithm [10, 11] or the PDIP algorithm [5]. In Section 4, we will address a solvability of the rate control problem (2).

4 Solvability of the Rate Control Problem

We begin this section by a useful lemma. First, we define z_{ij}, $r_{s,t}$, and d_l as the logarithmic values of the end-to-end session rate y_{ij}, wireless link transmission rate $\rho_{s,t}$, and wired link capacity c_l, respectively. It can be easily shown that problem (2) reduces a convex optimization problem by the following lemma.

Lemma 1. *Problem* (2) *is equivalent to the following convex optimization problem:*

$$\begin{aligned}
&\text{minimize} \quad -\sum_{(i,j)\in S} z_{ij} \\
&\text{subject to} \quad z_{ij} + \log\left(1 + \sum_{k\in D_{A(i)}} e^{r_{A(i),k}} + \sum_{k\in J_{A(i)}} e^{r_{k,A(i)}}\right) - r_{i,A(i)} \le 0 \\
&\qquad\qquad \forall (i,j) \in S, \\
&\qquad z_{ij} + \log\left(1 + \sum_{k\in D_{A(j)}} e^{r_{A(j),k}} + \sum_{k\in J_{A(j)}} e^{r_{k,A(j)}}\right) - r_{A(j),j} \le 0 \\
&\qquad\qquad \forall (i,j) \in S, \\
&\qquad \log\left(\sum_{(A(i),A(j))\in S(l)} e^{z_{ij}}\right) - d_l \le 0 \quad \forall l \in L.
\end{aligned} \tag{3}$$

Based on Lemma 1, we can conclude that the vectors

$$y := (y_{ij} : (i,j) \in S) \in \mathbb{R}_+^{|S|}$$

and

$$\rho := \left(\rho_{s,t} : (s,t) \in E_w, w \in W\right) \in \mathbb{R}_+^{|M|}$$

are optimal solutions of problem (2), if and only if the vectors

$$z := (z_{ij} : (i,j) \in S) \in \mathbb{R}^{|S|}$$

and

$$r := (r_{s,t} : (s,t) \in E_w, w \in W) \in \mathbb{R}^{|M|}$$

are optimal solutions of problem (3). Therefore, we will study the solvability of the original problem (2) via its equivalent problem (3). For ease of exposition, we denote functions $f(z,r)$, $g_{ij}^{(1)}(z,r)$, $g_{ij}^{(2)}(z,r)$ $((i,j) \in S)$, and $h_l(z,r)$ $(l \in L)$ as

$$f(z,r) := -\sum_{(i,j)\in S} z_{ij},$$

$$g_{ij}^{(1)}(z,r) := z_{ij} + \log\left(1 + \sum_{k\in D_{A(i)}} e^{r_{A(i),k}} + \sum_{k\in J_{A(i)}} e^{r_{k,A(i)}}\right) - r_{i,A(i)},$$

$$g_{ij}^{(2)}(z,r) := z_{ij} + \log\left(1 + \sum_{k\in D_{A(j)}} e^{r_{A(j),k}} + \sum_{k\in J_{A(j)}} e^{r_{k,A(j)}}\right) - r_{A(j),j},$$

$$h_l(z,r) := \log\left(\sum_{(A(i),A(j))\in S(l)} e^{z_{ij}}\right) - d_l,$$

and gradients of these functions are denoted as $\nabla f(z,r)$, $\nabla g_{ij}^{(1)}(z,r)$, $\nabla g_{ij}^{(2)}(z,r)$ $((i,j) \in S)$, and $\nabla h_l(z,r)$ $(l \in L)$, respectively.

In order to study the solvability of problem (3), we assume that there exist vectors $\bar{z} \in \mathbb{R}^{|S|}$ and $\bar{r} \in \mathbb{R}^{|M|}$ such that $g_{ij}^{(1)}(\bar{z},\bar{r}) < 0$, $g_{ij}^{(2)}(\bar{z},\bar{r}) < 0$, for all $(i,j) \in S$ and $h_l(\bar{z},\bar{r}) < 0$ for all $l \in L$, i.e., Slater's condition of problem (3) holds (see [1, p. 226]). Furthermore, according to Lemma 1, the problem (3) is convex, and it leads to the conclusion that the Karush–Kuhn–Tucker (KKT) conditions provide necessary and sufficient conditions for optimality (see [1, p. 244]). Thus, $(z^*, r^*) \in \mathbb{R}^{|S|+|M|}$ is an optimal solution of problem (3) if and only if there is a dual optimal solution $(\lambda_{ij}^{(1)^*}, \lambda_{ij}^{(2)^*}, \gamma_l^*) \in \mathbb{R}^{2|S|+|L|}$ that, together with (z^*, r^*), satisfies the KKT conditions, see [1, p. 243] as follows:

$$\begin{aligned} g_{ij}^{(1)}(z^*,r^*) &\le 0 \quad \forall (i,j) \in S, \\ g_{ij}^{(2)}(z^*,r^*) &\le 0 \quad \forall (i,j) \in S, \\ h_l(z^*,r^*) &\le 0 \quad \forall l \in L; \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda_{ij}^{(1)^*} &\ge 0 \quad \forall (i,j) \in S, \\ \lambda_{ij}^{(2)^*} &\ge 0 \quad \forall (i,j) \in S, \\ \gamma_l^* &\ge 0 \quad \forall l \in L; \end{aligned} \tag{5}$$

$$\begin{aligned} \lambda_{ij}^{(1)^*} g_{ij}^{(1)}(z^*,r^*) &= 0 \quad \forall (i,j) \in S, \\ \lambda_{ij}^{(2)^*} g_{ij}^{(2)}(z^*,r^*) &= 0 \quad \forall (i,j) \in S, \\ \gamma_l^* h_l(z^*,r^*) &= 0 \quad \forall l \in L; \end{aligned} \tag{6}$$

$$\nabla f(z^*, r^*) + \sum_{(i,j)\in S} \lambda_{ij}^{(1)^*} \nabla g_{ij}^{(1)}(z^*, r^*) + \sum_{(i,j)\in S} \lambda_{ij}^{(2)^*} \nabla g_{ij}^{(2)}(z^*, r^*) + \sum_{l\in L} \gamma_l^* \nabla h_l(z^*, r^*) = 0. \quad (7)$$

It is worth noting that, in BSS $w \in W$, each wireless link $(s,t) \in E_w$ capacity depends on the wireless link transmission rates $\rho_{k,m}, \forall (k,m) \in E_w$. Furthermore, each session $(i,j) \in S$ originates from one wireless network and ends at another such as MHs i and j, respectively, where $(i, A(i)) \in E_w$, $(A(j), j) \in E_v$ $w, v \in W$, and $w \neq v$. Notice that in our network model, we have $\sum_{w\in W} |E_w| = |M|$ and $|M| = 2|S|$. Then, in each BSS $w \in W$, we can restore an index of variables $r_{k,m}, \forall (k,m) \in E_w$ as $r_1^{(w)}, \ldots, r_{|E_w|}^{(w)}$ and variables $\lambda_{ij}^{(1)}$ or $\lambda_{ij}^{(2)}$ such that sessions (i,j) travel across wireless links (k,m), respectively, as $\lambda_1^{(w)}, \ldots, \lambda_{|E_w|}^{(w)}$. Note that system (7) consists of $|S| + |M|$ equations and $3|S| + |M| + |L|$ unknowns. In particular, there are $|S|$ unknowns z_{ij}^*, $|M|$ unknowns $r_{s,t}^*$, $|S|$ unknowns $\lambda_{ij}^{(1)^*}$, $|S|$ unknowns $\lambda_{ij}^{(2)^*}$, and $|L|$ unknowns γ_l^*. On the other hand, since the functions $f(z,r)$ and $h_l(z,r)$ do not depend on variables $r_{s,t}$, we obtain a subsystem equation that consists of $|M|$ equations and $|M|$ unknowns $r_{s,t}^*$, $2|S|$ unknowns $\lambda_{ij}^{(1)^*}$, $\lambda_{ij}^{(2)^*}$; and this subsystem only depends on the functions $g_{ij}^{(1)}(z,r)$ and $g_{ij}^{(2)}(z,r)$. Now, in the subsystem, we only consider $\lambda_{ij}^{(1)^*}$, $\lambda_{ij}^{(2)^*}$ as unknowns. Due to $|M| = 2|S|$, the subsystem is a square linear system equation. The square linear system equation can be separated into $|W|$ square subsystems. We can calculate the gradients $\nabla g_{ij}^{(1)}(z,r)$ and $\nabla g_{ij}^{(2)}(z,r)$, and as mentioned above, for each $w \in W$ from the system equation (7), we get the square linear subsystem equations as follows:

$$A^{(w)} \lambda^{(w)^*} = 0, \quad w \in W, \quad (8)$$

where $\lambda^{(w)^*} := \left(\lambda_1^{(w)^*}, \ldots, \lambda_{|E_w|}^{(w)^*}\right) \in \mathbb{R}^{|E_w|}$ and

$$A^{(w)} := \begin{pmatrix} \frac{e^{r_1^{(w)^*}}}{d^{(w)}} - 1 & \frac{e^{r_1^{(w)^*}}}{d^{(w)}} & \cdots & \frac{e^{r_1^{(w)^*}}}{d^{(w)}} \\ \frac{e^{r_2^{(w)^*}}}{d^{(w)}} & \frac{e^{r_2^{(w)^*}}}{d^{(w)}} - 1 & \cdots & \frac{e^{r_2^{(w)^*}}}{d^{(w)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e^{r_{|E_w|}^{(w)^*}}}{d^{(w)}} & \frac{e^{r_{|E_w|}^{(w)^*}}}{d^{(w)}} & \cdots & \frac{e^{r_{|E_w|}^{(w)^*}}}{d^{(w)}} - 1 \end{pmatrix}.$$

Here, we denote $d^{(w)} := 1 + \sum_{j=1}^{|E_w|} e^{r_j^{(w)^*}}$.

Theorem 1. *The linear system equations* (8) *always have a unique solution* $\lambda^{(w)^*} = 0$ *for any given vectors* $r^{(w)^*} := \left(r_1^{(w)^*}, \ldots, r_{|E_w|}^{(w)^*}\right) \in \mathbb{R}^{|E_w|}$ *and for all* $w \in W$.

Theorem 2. *The rate control problem* (2) *always has a unique optimal solution for the end-to-end session rates* $y^* := \left(y^*_{ij} : (i,j) \in S\right)$ *and has infinitely many optimal solutions for the wireless link transmission rates* $\rho^* := \left(\rho^*_{s,t} : (s,t) \in E_w, w \in W\right)$.

If we take into account that the functions $f(z,r)$ and $h_l(z,r)$ do not depend on variable r, the system of equations (7) can be rewritten as follows:

$$\nabla f(z^*) + \sum_{l \in L} \gamma^*_l \nabla h_l(z^*) = 0. \tag{9}$$

This is a system of nonlinear equations which consists of $|S|$ equations and $|S|$ variables z^*_{ij} $((i,j) \in S)$ and $|L|$ variables γ^*_l $(l \in L)$. If we view the variables γ^*_l $(l \in L)$ as parameters, the system of nonlinear equations (9) has only $|S|$ unknowns z^*_{ij} $(i,j) \in S$ and $|S|$ equations. As a consequence of Theorem 2, we state the following result.

Corollary 1. *The system of nonlinear equations* (9) *always has a unique solution* z^* *for any given vector* $\gamma^* = (r^*_l : l \in L) \in \mathbb{R}^{|L|}$ *provided that* $\gamma^*_l \geq 0$ $\forall l \in L$ *and* $\sum_{l \in L} \gamma^*_l > 0$.

5 Numerical Example

In this section, we investigate a numerical example, which is taken from [5, 10, 11], to illustrate our theoretical results in Section 4. Consider the network which is illustrated in Fig. 2.

The network is composed of four APs which are denoted as 0, 1, 2, and 3 and eight MHs which are labeled as A, B, C, D, E, F, G, and H. There are a total of eight wireless links which are denoted as a, b, c, d, e, f, g, and h. The wired backbone of the network connects the APs through four wired links, denoted as 0, 1, 2, and 3. The capacities of the wired links are 0.5, 0.2, 0.6, and 0.8, respectively. Four end-to-end sessions, namely, f_0, f_1, f_2, and f_3 are set up in this network. The source, the destination, and the path of the four sessions are shown in Table 1.

Table 1. The source, sink, and path of the sessions

Session	Source node	Sink node	Link on the path
f_0	E	A	$e, 0, a$
f_1	B	G	$b, 0, 2, g$
f_2	C	F	$c, 3, 2, f$
f_3	H	D	$h, 2, 1, d$

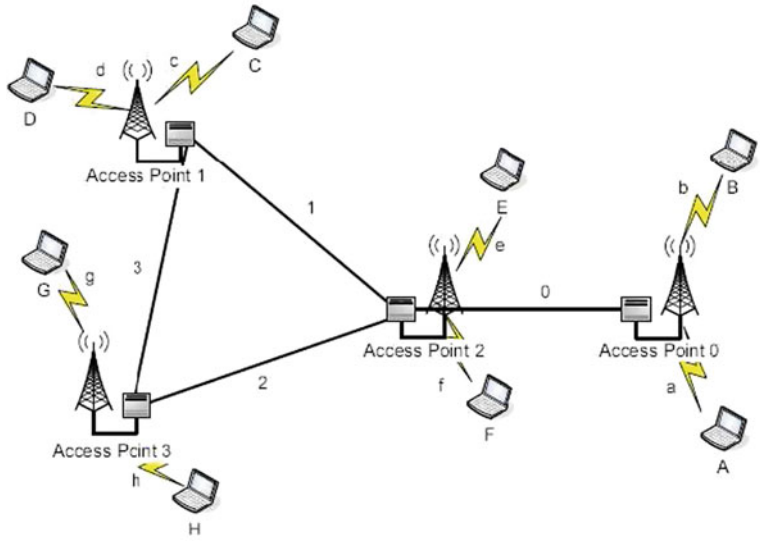


Fig. 2. A wired-cum-wireless network example

We found optimal solutions of the rate control problem for this network by using both the DB-distributed algorithm in [10, 11] and the PDIP algorithm in [1, 5]. Our computations were done using Matlab 7.0 on a machine with 3.00 GHz Pentium processor and 1.00 GB of RAM. In this example, we denote four end-to-end session rates of sessions by f_0, f_1, f_2, and f_3, and eight wireless link transmission rates of wireless links by a, b, c, d, e, f, g, and h as y_0, y_1, y_2, and y_3 and ρ_i $(i = \overline{1,8})$, respectively.

Dual-based algorithm [5, 10, 11]: In [10, 11], the DB algorithm has been proposed to solve the rate control problem (2) iteratively. This algorithm has been reviewed in [5] when we compare one with the PDIP algorithm to find a solution of problem (2). In DB algorithm, there are inner and outer iterations. We compute iteratively the link prices and the end-to-end session rates while the wireless link transmission rates are fixed by inner loop. The wireless link transmission rates are updated by using outer loop.

In this simulation, the step sizes, i.e., step size β and step size δ, for the inner and outer loops are set to $\beta = \delta = 0.15$ and $\beta = 0.15$, $\delta = 5 \times 10^{-4}$. Both inner and outer loops terminate when the norm of the difference of two successive iterative end-to-end session rates, and transmission rates, respectively, is smaller than $\epsilon = 10^{-8}$.

The convergence of the DB algorithm ensures that a numerical solution is only one optimal solution. The problem may have other optimal solutions. As shown in Section 4, problem (2) has a unique optimal solution for the end-to-end session rates and has infinitely many optimal solutions for the wireless

link transmission rates, i.e., problem (2) for this network has many optimal solutions. From experiments, we can see that the optimal solution obtained by the DB algorithm does not depend on an initial value of the link prices λ and γ and the step size β (see [5, 10, 11]). This can be interpreted that a dual problem of problem (2) has a unique optimal solution. Thus, we choose an initial value for the link price vectors $\lambda^{(0)} = e \in \mathbb{R}^8$ and $\gamma^{(0)} = e \in \mathbb{R}^4$ where e denotes the vector of all ones whose dimension is determined by the context. However, the optimal solution depends on an initial value of the wireless link transmission rates and the step-size δ. Tables 2 and 3 show dependence of the optimal solution on the initial transmission rate vectors $\rho^{(0)}$ in three cases with $\delta = 0.15$ and $\delta = 5 \times 10^{-4}$, respectively.

Primal–dual interior–point algorithm [1, 5]: In this simulation, instead of solving problem (2) for this network example directly, we will solve the equivalent problem (3) using the PDIP algorithm in [5]. Similar in DB algorithm, through this simulation example, it can be seen that the optimal numerical solution given by the PDIP algorithm also depends on the choice of initial values of vectors $z^{(0)} \in \mathbb{R}^4$ and $r^{(0)} \in \mathbb{R}^8$, which are logarithms of the end-to-end session rates and the wireless link transmission rates, respectively, and the backtracking parameters α and β in the PDIP algorithm. From Tables 4 and 5, the initial vector $\lambda^{(0)} \in \mathbb{R}^{12}$ is chosen as $\lambda_i^{(0)} = -1/c_i(z^{(0)}, r^{(0)}), \overline{1, 12}$ where $c_i(z, r)$ are the constrained functions $g_{ij}^{(1)}(z, r)$, $g_{ij}^{(2)}(z, r)$, $\forall (i, j) \in S$ and $h_l(z, r)$, $l \in L$ in problem (3). We consider two cases $\alpha = 0.01$, $\beta = 0.5$ and $\alpha = 0.1$, $\beta = 0.8$ corresponding to Tables 4 and 5. Other parameter values that we used for the PDIP algorithm (see [5] for details) are $\epsilon = 10^{-8}$, $\mu = 10$.

From Tables 2, 3, 4, and 5, it can be seen that the rate control problem (2) or the equivalent problem (3) for this network example has a unique optimal solution for the end-to-end session rates. However, it has infinitely many optimal solutions for the wireless link transmission rates.

6 Conclusion

We have discussed the solvability of the rate control problem in wired-cum-wireless networks. The rate control problem is a nonconvex optimization problem. In general, finding an optimal solution for the rate control problem in wired-cum-wireless networks is more difficult than its wired network counterpart. In this chapter, using linear algebra and convex optimization techniques, we have proved existence of a unique solution in the end-to-end session rates. We have also shown that there may exist infinitely many of optimal solutions for the wireless link transmission rates in the rate control problem for the wired-cum-wireless network. Numerical examples have been provided to support our obtained results.

Table 2. The optimal solutions given by the DB algorithm with $\delta = 0.15$

Initial wireless link transmission rates	Optimal end-to-end session rates	Optimal wireless link transmission rates
$\rho_1^{(0)} = 1$	$y_0^* = 0.35275514$	$\rho_1^* = 1.07019867$
$\rho_2^{(0)} = 1$	$y_1^* = 0.14724746$	$\rho_2^* = 0.96361791$
$\rho_3^{(0)} = 1$	$y_2^* = 0.25275111$	$\rho_3^* = 1.00000000$
$\rho_4^{(0)} = 1$	$y_3^* = 0.20000074$	$\rho_4^* = 1.00000000$
$\rho_5^{(0)} = 1$		$\rho_5^* = 1.07019867$
$\rho_6^{(0)} = 1$		$\rho_6^* = 0.96361791$
$\rho_7^{(0)} = 1$		$\rho_7^* = 1.00000000$
$\rho_8^{(0)} = 1$		$\rho_8^* = 1.00000000$
$\rho_1^{(0)} = 0.5$	$y_0^* = 0.35275212$	$\rho_1^* = 0.77061752$
$\rho_2^{(0)} = 0.5$	$y_1^* = 0.14724775$	$\rho_2^* = 0.39055502$
$\rho_3^{(0)} = 0.5$	$y_2^* = 0.25275355$	$\rho_3^* = 0.50749777$
$\rho_4^{(0)} = 0.5$	$y_3^* = 0.19999966$	$\rho_4^* = 0.49750074$
$\rho_5^{(0)} = 0.5$		$\rho_5^* = 0.89419835$
$\rho_6^{(0)} = 0.5$		$\rho_6^* = 0.64071231$
$\rho_7^{(0)} = 0.5$		$\rho_7^* = 0.50000000$
$\rho_8^{(0)} = 0.5$		$\rho_8^* = 0.500000007$
$\rho_1^{(0)} = 0.1$	$y_0^* = 0.35275319$	$\rho_1^* = 0.89667809$
$\rho_2^{(0)} = 0.1$	$y_1^* = 0.14724726$	$\rho_2^* = 0.64524173$
$\rho_3^{(0)} = 0.1$	$y_2^* = 0.25275155$	$\rho_3^* = 0.72498237$
$\rho_4^{(0)} = 0.1$	$y_3^* = 0.20000071$	$\rho_4^* = 0.72498103$
$\rho_5^{(0)} = 0.1$		$\rho_5^* = 0.89668243$
$\rho_6^{(0)} = 0.1$		$\rho_6^* = 0.64523699$
$\rho_7^{(0)} = 0.1$		$\rho_7^* = 0.72498198$
$\rho_8^{(0)} = 0.1$		$\rho_8^* = 0.72498106$

Appendix

A. Proof of Theorem 1

In order to prove that the linear system equations (8) have a unique solution $\lambda^{(w)^*} = 0 \in \mathbb{R}^{|E_w|}$, it is sufficient to show $\det\left(A^{(w)}\right) \neq 0$ for any given vectors $r^{(w)^*} = \left(r_1^{(w)^*}, \ldots, r_{|E_w|}^{(w)^*}\right) \in \mathbb{R}^{|E_w|}$ and for all $w \in W$. First, by the properties of the determinant, adding all $|E_w| - 1$ the last rows of the matrix $A^{(w)}$ to its first row, and then after multiplying the first column by -1 we add it to each column from the second column to the last column of the matrix $A^{(w)}$, it follows that

Table 3. The optimal solutions given by the DB algorithm with $\delta = 5 \times 10^{-4}$

Initial wireless link transmission rates	Optimal end-to-end session rates	Optimal wireless link transmission rates
$\rho_1^{(0)} = 1$	$y_0^* = 0.35275255$	$\rho_1^* = 1.07006794$
$\rho_2^{(0)} = 1$	$y_1^* = 0.14724911$	$\rho_2^* = 0.96340464$
$\rho_3^{(0)} = 1$	$y_2^* = 0.25274991$	$\rho_3^* = 1.00000000$
$\rho_4^{(0)} = 1$	$y_3^* = 0.20000051$	$\rho_4^* = 1.00000000$
$\rho_5^{(0)} = 1$		$\rho_5^* = 1.07006794$
$\rho_6^{(0)} = 1$		$\rho_6^* = 0.96340464$
$\rho_7^{(0)} = 1$		$\rho_7^* = 1.00000000$
$\rho_8^{(0)} = 1$		$\rho_8^* = 1.00000000$
$\rho_1^{(0)} = 0.5$	$y_0^* = 0.35275052$	$\rho_1^* = 0.75665540$
$\rho_2^{(0)} = 0.5$	$y_1^* = 0.14724950$	$\rho_2^* = 0.38835639$
$\rho_3^{(0)} = 0.5$	$y_2^* = 0.25275109$	$\rho_3^* = 0.50661377$
$\rho_4^{(0)} = 0.5$	$y_3^* = 0.19999973$	$\rho_4^* = 0.49777919$
$\rho_5^{(0)} = 0.5$		$\rho_5^* = 0.89417477$
$\rho_6^{(0)} = 0.5$		$\rho_6^* = 0.64068918$
$\rho_7^{(0)} = 0.5$		$\rho_7^* = 0.50000000$
$\rho_8^{(0)} = 0.5$		$\rho_8^* = 0.500000007$
$\rho_1^{(0)} = 0.1$	$y_0^* = 0.35274934$	$\rho_1^* = 0.73135710$
$\rho_2^{(0)} = 0.1$	$y_1^* = 0.14725034$	$\rho_2^* = 0.34194240$
$\rho_3^{(0)} = 0.1$	$y_2^* = 0.25275094$	$\rho_3^* = 0.46185889$
$\rho_4^{(0)} = 0.1$	$y_3^* = 0.19999944$	$\rho_4^* = 0.36546383$
$\rho_5^{(0)} = 0.1$		$\rho_5^* = 0.89416777$
$\rho_6^{(0)} = 0.1$		$\rho_6^* = 0.64068733$
$\rho_7^{(0)} = 0.1$		$\rho_7^* = 0.34225203$
$\rho_8^{(0)} = 0.1$		$\rho_8^* = 0.34130752$

$$\det\left(A^{(w)}\right) = \begin{vmatrix} -\frac{1}{d^{(w)}} & 0 & \cdots & 0 \\ \frac{e^{r_2^{(w)^*}}}{d^{(w)}} & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{e^{r_{|E_w|}^{(w)^*}}}{d^{(w)}} & 0 & \cdots & -1 \end{vmatrix}.$$

Note that the right-hand side of the above equality is a determinant of an $|E_w|$ by $|E_w|$ lower triangular matrix. Thus, we obtain that

$$\det\left(A^{(w)}\right) = \frac{(-1)^{|E_w|}}{d^{(w)}} \neq 0,$$

for any given vectors $r^{(w)^*} = \left(r_1^{(w)^*}, \ldots, r_{|E_w|}^{(w)^*}\right) \in \mathbb{R}^{|E_w|}$, and for all $w \in W$. Q.E.D.

Table 4. The optimal solutions given by the PDIP algorithm with $\alpha = 0.01$, $\beta = 0.5$

Initial wireless link transmission rates		Optimal end-to-end session rates	Optimal wireless link transmission rates
$z_0^{(0)} = -2$,	$r_1^{(0)} = -1$	$y_0^* = 0.35275252$	$\rho_1^* = 2.98323299$
$z_1^{(0)} = -2$,	$r_2^{(0)} = -1$	$y_1^* = 0.14724748$	$\rho_2^* = 1.65669132$
$z_2^{(0)} = -2$,	$r_3^{(0)} = -1$	$y_2^* = 0.25275252$	$\rho_3^* = 2.24011428$
$z_3^{(0)} = -2$,	$r_4^{(0)} = -1$	$y_3^* = 0.20000000$	$\rho_4^* = 1.91004133$
	$r_5^{(0)} = -1$		$\rho_5^* = 3.54833077$
	$r_6^{(0)} = -1$		$\rho_6^* = 2.76138572$
	$r_7^{(0)} = -1$		$\rho_7^* = 1.42793561$
	$r_8^{(0)} = -1$		$\rho_8^* = 1.71710508$
$z_0^{(0)} = -3$,	$r_1^{(0)} = -2$	$y_0^* = 0.35275252$	$\rho_1^* = 3.18691351$
$z_1^{(0)} = -3$,	$r_2^{(0)} = -2$	$y_1^* = 0.14724748$	$\rho_2^* = 1.76955508$
$z_2^{(0)} = -3$,	$r_3^{(0)} = -2$	$y_2^* = 0.25275252$	$\rho_3^* = 2.36056534$
$z_3^{(0)} = -3$,	$r_4^{(0)} = -2$	$y_3^* = 0.20000000$	$\rho_4^* = 2.01155550$
	$r_5^{(0)} = -2$		$\rho_5^* = 3.83139094$
	$r_6^{(0)} = -2$		$\rho_6^* = 3.00129934$
	$r_7^{(0)} = -2$		$\rho_7^* = 1.48032082$
	$r_8^{(0)} = -2$		$\rho_8^* = 1.77862488$
$z_0^{(0)} = -5$,	$r_1^{(0)} = -0.5$	$y_0^* = 0.35275252$	$\rho_1^* = 16.71211859$
$z_1^{(0)} = -5$,	$r_2^{(0)} = -0.5$	$y_1^* = 0.14724748$	$\rho_2^* = 10.13139616$
$z_2^{(0)} = -5$,	$r_3^{(0)} = -0.5$	$y_2^* = 0.25275252$	$\rho_3^* = 12.72947321$
$z_3^{(0)} = -5$,	$r_4^{(0)} = -0.5$	$y_3^* = 0.20000000$	$\rho_4^* = 11.32254620$
	$r_5^{(0)} = -0.5$		$\rho_5^* = 19.99732265$
	$r_6^{(0)} = -0.5$		$\rho_6^* = 16.18876216$
	$r_7^{(0)} = -0.5$		$\rho_7^* = 8.53671165$
	$r_8^{(0)} = -0.5$		$\rho_8^* = 10.13066345$

B. Proof of Theorem 2

In problem (2), the objective function

$$\sum_{(i,j)\in S} \log(y_{ij})$$

is differentiable and strictly concave, and its feasible region is compact, hence a maximizing value of (y^*, ρ^*) exists. Moreover, since the objective function is a strictly concave function in y, it implies that there exists a unique optimal solution for the end-to-end session rates vector y.

Applying Lemma 1, we can conclude that the equivalent convex problem (3) has an optimal solution (z^*, r^*) with unique z^*. Thus, there exists a dual optimal solution $\left(\lambda_{ij}^{(1)^*}, \lambda_{ij}^{(2)^*}, \gamma_l^*\right) \in \mathbb{R}^{2|S|+|L|}$ that, together with

Table 5. The optimal solutions given by the PDIP algorithm with $\alpha = 0.1$, $\beta = 0.8$

Initial wireless link transmission rates		Optimal end-end session rates	Optimal wireless link transmission rates
$z_0^{(0)} = -2,$	$r_1^{(0)} = -1$	$y_0^* = 0.35275252$	$\rho_1^* = 2.36844983$
$z_1^{(0)} = -2,$	$r_2^{(0)} = -1$	$y_1^* = 0.14724748$	$\rho_2^* = 1.29607831$
$z_2^{(0)} = -2,$	$r_3^{(0)} = -1$	$y_2^* = 0.25275252$	$\rho_3^* = 1.80564383$
$z_3^{(0)} = -2,$	$r_4^{(0)} = -1$	$y_3^* = 0.20000000$	$\rho_4^* = 1.52644375$
	$r_5^{(0)} = -1$		$\rho_5^* = 2.76882804$
	$r_6^{(0)} = -1$		$\rho_6^* = 2.14179347$
	$r_7^{(0)} = -1$		$\rho_7^* = 1.16910744$
	$r_8^{(0)} = -1$		$\rho_8^* = 1.41303270$
$z_0^{(0)} = -3,$	$r_1^{(0)} = -2$	$y_0^* = 0.35275252$	$\rho_1^* = 2.63438605$
$z_1^{(0)} = -3,$	$r_2^{(0)} = -2$	$y_1^* = 0.14724748$	$\rho_2^* = 1.45111659$
$z_2^{(0)} = -3,$	$r_3^{(0)} = -2$	$y_2^* = 0.25275252$	$\rho_3^* = 1.95895545$
$z_3^{(0)} = -3,$	$r_4^{(0)} = -2$	$y_3^* = 0.20000000$	$\rho_4^* = 1.65426526$
	$r_5^{(0)} = -2$		$\rho_5^* = 3.13373673$
	$r_6^{(0)} = -2$		$\rho_6^* = 2.43738586$
	$r_7^{(0)} = -2$		$\rho_7^* = 1.22790485$
	$r_8^{(0)} = -2$		$\rho_8^* = 1.47099894$
$z_0^{(0)} = -5,$	$r_1^{(0)} = -0.5$	$y_0^* = 0.35275252$	$\rho_1^* = 10.55114238$
$z_1^{(0)} = -5,$	$r_2^{(0)} = -0.5$	$y_1^* = 0.14724748$	$\rho_2^* = 6.54517622$
$z_2^{(0)} = -5,$	$r_3^{(0)} = -0.5$	$y_2^* = 0.25275252$	$\rho_3^* = 8.91290134$
$z_3^{(0)} = -5,$	$r_4^{(0)} = -0.5$	$y_3^* = 0.20000000$	$\rho_4^* = 8.02515194$
	$r_5^{(0)} = -0.5$		$\rho_5^* = 12.71012289$
	$r_6^{(0)} = -0.5$		$\rho_6^* = 10.63664631$
	$r_7^{(0)} = -0.5$		$\rho_7^* = 5.90488864$
	$r_8^{(0)} = -0.5$		$\rho_8^* = 7.20092715$

(z^*, r^*), satisfies the KKT conditions (4), (5), (6), and (7). According to Theorem (1), the KKT conditions (4), (5), (6), and (7) have a unique solution $\left(\lambda_{ij}^{(1)^*}, \lambda_{ij}^{(2)^*}\right) = 0 \in \mathbb{R}^{2|S|}$ for variable λ. Note that the functions $f(z, r)$ and $h_l(z, r)$, $\forall l \in L$ do not consist of variable r. Therefore, the KKT conditions (4), (5), (6), and (7) are reduced to the system (9), (10), (11), and (12), where the system (10), (11), and (12) is given by

$$\begin{aligned} g_{ij}^{(1)}(z^*, r^*) &\leq 0, \quad \forall (i,j) \in S, \\ g_{ij}^{(2)}(z^*, r^*) &\leq 0, \quad \forall (i,j) \in S, \\ h_l(z^*) &\leq 0, \quad \forall l \in L, \end{aligned} \tag{10}$$

$$\gamma_l^* \geq 0, \quad \forall l \in L, \tag{11}$$

$$\gamma_l^* h_l(z^*) = 0, \quad \forall l \in L. \tag{12}$$

Due to the unique existence of vector $z^* \in \mathbb{R}^{|S|}$ and the inequality constraint functions $g_{ij}^{(1)}(z,r)$ and $g_{ij}^{(2)}(z,r)$, $\forall (i,j) \in S$, we arrive at the conclusion that there exist infinitely many solutions of vectors $r^* \in \mathbb{R}^{|M|}$ which satisfy relation (10). Thus there is a unique value of $z^* \in \mathbb{R}^{|S|}$ and there are many corresponding values of $r^* \in \mathbb{R}^{|M|}$ such that $(z^*, r^*) \in \mathbb{R}^{|S|+|M|}$ satisfies the KKT conditions (9), (10), (11), and (12). Therefore, we proved that the convex optimization problem (3) always has a unique optimal solution $z^* = \left(z_{ij}^* : (i,j) \in S\right)$ and infinitely many optimal solutions of $r^* = \left(r_{s,t}^* : (s,t) \in E_w, w \in W\right)$. Q.E.D.

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization, Cambridge University Press, Cambridge, (2004)
2. Chiang, M.: Balancing transport and physical layer in wireless multihop networks: Jointly optimal congestion control and power control. IEEE J. Sel. Area. Comm. 23(1), 104–116 (2005)
3. Kelly, F.: Charging and rate control for elastic traffic. Eur. Trans. Telecommun. 8, 33–37, (1997)
4. Kelly, F., Maullo, A., Tan, D.: Rate control for communication networks: Shadow prices, proportional fairness and stability. J. Oper. Res. Soc. 49(3), 237–252 (1998)
5. Loi, L.C., Hwang, W.-J.: A new approach to solve the rate control problem in wired-cum-wireless networks. J. Korea Multimed. Soc. 9(12), 28–40 (2006)
6. Loi, L.C., Hwang, W.-J.: Optimization wireless link transmission rates in wired-cum-wireless networks. Proc. Korea Multimed. Soc. 9(2), 195–198 (2006)
7. Low, S.H., Lapsley, D.E.: Optimization flow control, I: Basic algorithm and convergence. IEEE/ACM Trans. Network. 7(6), 861–874 (1999)
8. Wang, X., Kar, K.: Throughput modelling and fairness issues in CSMA/CA based ad-hoc networks. Proc. INFOCOM 1, 23–34 (2005)
9. Wang, X., Kar, K.: Cross-layer rate optimization for proportional fairness in multihop wireless networks with random access. IEEE J. Select. Area. Commun. 24(8), 1548–1558 (2006)
10. Wang, X., Kar, K., Low, S.H.: Cross-layer rate optimization in wired-cum-wireless networks. Proceeding of 19th International Teletraffic Congress (ITC), Beijing, China (2005)
11. Wang, X., Kar, K., Low, S.H.: End-to-end fair rate optimization in wired-cum-wireless networks. Ad Hoc Network. 7, 473–485 (2009)

Integer Programming of Biclustering Based on Graph Models

Neng Fan[1], Altannar Chinchuluun[2], and Panos M. Pardalos[3]

[1] Center for Applied Optimization, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
andynfan@ufl.edu
[2] Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, UK
a.chinchuluun@imperial.ac.uk
[3] Center for Applied Optimization, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
pardalos@ufl.edu

Summary. In this chapter, biclustering is studied in a mathematical prospective, including bipartite graphs and optimization models via integer programming. A correspondence between biclustering and graph partitioning is established. In the optimization models, different cuts are used and the integer programming models are presented. We prove that the spectral biclustering for Ratio cut and Normalized cut are the relaxation forms of these integer programming models, and also the Minmax cut for biclustering is equivalent to Normalized cut for biclustering.

Key words: biclustering, integer programming, spectral clustering, graph partitioning, ratio cut, normalized cut, minmax cut

1 Introduction

With large amounts of data collected from different areas, finding the relevant information among them appears to be very important. Data mining is a process of doing this and one hot research area is data clustering, which deals with techniques to classify data into different groups. Many algorithms were designed to face the challenges in data clustering, and a survey of algorithms can be found in [15] while several applications in biological networks are discussed in [2].

In data clustering, data points are grouped with respect to the relations between each other, but the attributes of these data are not classified. Biclustering (co-clustering, two-mode clustering) model was introduced in [12] and

The research of the second author is supported by MOBILE, ERC Advanced Grant No 226462.

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_23,

recently used in gene expression analysis [4]. Different from clustering, biclustering can simultaneously group both data and their attributes. For example, for the data of gene expression microarray, all gene samples together form the data, while each gene has different functions (called features). Biclustering techniques will group gene samples and features while each group of genes is corresponding to a specific function. Mathematically, this kind of data will be stored in a matrix with numerical entries.

Many algorithms were designed to solve the biclustering problem, and surveys of these methods can be found in [1, 11] and recent algorithms in [7]. To measure the differences between biclusters, two mostly used are the Ratio cut [8] and the Normalized cut [14]. There are also many other different measurements of difference [5, 9, 13, 16], but the authors always used many different kinds of approach to model the problem. In this chapter, a more general approach will be introduced based on bipartite graph.

This chapter is organized as follows: In Section 2, mathematical representations of biclustering are presented. In Section 3, correspondence between graph partitioning and biclustering is established. In Section 4, the integer programming models for Ratio cut, Normalized cut, Minmax cut, and ICA cut are introduced with relaxation forms. Section 5 concludes the chapter.

2 Biclustering Models

As mentioned above, data for biclustering usually is stored in a rectangular matrix. Using the example of data from gene expression microarray with n samples and m features of genes, let $A = (a_{ij})_{n\times m}$ be the data matrix where each row of A corresponds to a sample, each column to a kind of feature, and each entry a_{ij} denotes the expression level (or called weight) of a gene sample i with a specific feature j.

In [1], Busygin, Prokopyev, and Pardalos presented a mathematical definition of biclustering. Before giving the definition of biclustering, the partition of a matrix is defined first.

Definition 1. *Given a data matrix $A = (a_{ij})_{n\times m}$, its* partition *is defined as a collection of subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ of its rows such that*

$$\begin{gathered}\mathcal{S}_i \subseteq \{1, \ldots, n\} (i = 1, \ldots, k),\\ \mathcal{S}_1 \cup \mathcal{S}_2 \cup \ldots \cup \mathcal{S}_k = \{1, \ldots, n\},\\ \mathcal{S}_i \cap \mathcal{S}_j = \emptyset, i, j = 1, \ldots, k, i \neq j,\end{gathered}$$

and a corresponding collection of subsets $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ of its columns such that

$$\begin{gathered}\mathcal{F}_i \subseteq \{1, \ldots, m\} (i = 1, \ldots, k),\\ \mathcal{F}_1 \cup \mathcal{F}_2 \cup \ldots \cup \mathcal{F}_k = \{1, \ldots, m\},\\ \mathcal{F}_i \cap \mathcal{F}_j = \emptyset, i, j = 1, \ldots, k, i \neq j,\end{gathered}$$

where $k (1 \leq k \leq \min\{n, m\})$ is the number of parts it partitions to.

In a mathematical point of view, both the rows and the columns of the matrix are partitioned into k parts. The pairs $(\mathcal{S}_i, \mathcal{F}_i)$ are submatrices in the diagonal of the matrix after properly rearranging the rows and columns of A.

The biclustering is expressed in the form of a partition of the data matrix A and a bicluster is a submatrix of A with a pair of groups $(\mathcal{S}_i, \mathcal{F}_i)$ of both samples and features. The data matrix A used is kind of "sample–feature" one, which is different from the matrix usually used in clustering as "sample–sample" type. For the biclusters $(\mathcal{S}_i, \mathcal{F}_i), i = 1, \ldots, k$, this does not mean that the samples in $\mathcal{S}_i$ cannot have features outside $\mathcal{F}_i$. In some cases, some sample may have high expression level outside its corresponding feature group. Generally, a bicluster reflects the features of samples in groups, not individually.

For biclustering, the objectives are to maximize intra similarity of samples according to features in a bicluster and minimize the inter similarity of samples from different biclusters with respect to features. In order to achieve these objectives, many different objective functions are defined to measure the similarity or dissimilarity as discussed below.

3 General Approach to Biclustering

3.1 Graph Partitioning

Since different objective functions are defined to measure the similarity or dissimilarity among parts, many approaches are used in different papers to transform the biclustering problem into optimization models. Here, based on graph theory, a general approach is discussed. Before discussing transformations, several concepts used in graph theory are defined.

Definition 2. *An (undirected)* graph $G = (V, E)$ *consists of a set of vertices* $V = \{v_1, v_2, \ldots, v_{|V|}\}$ *and a set of edges* $E = \{(i, j) : \text{edge between } v_i \text{ and } v_j, i, j \leq |V|\}$, *where* $|V|$ *is the number of vertices. A* bipartite graph *is defined as a graph* $G = (U, V, E)$, *where* U, V *are two disjoint sets of vertices, and* E *is the set of edges between vertices from* U *and* V *while no edge appears between any vertices from* U *or* V.

For an edge $(i, j) \in E$ of the bipartite graph $G = (U, V, E)$, let $w(i, j)$ be the associated weight of edge (i, j). For the cases we considered in this chapter, the edges (i, j) and (j, i) are the same and $w(i, j) = w(j, i)$. Based on the weights of edges, there are some useful matrices defined in the following.

Definition 3. *Several weighted matrices of the graph* $G = (V, E)$ *are defined as follows:*

(1) The adjacency weighted matrix $W = (w_{ij})_{|V|\times|V|}$ *of the graph is defined as*

$$w_{ij} = \begin{cases} w(i, j), & \text{if the edge } (i, j) \text{ exists,} \\ 0, & \text{otherwise.} \end{cases}$$

(2) The weighted degree d_i *of vertex* v_i *is defined as*

$$d_i = \sum_{j:(i,j)\in E} w(i,j),$$

and the degree matrix $D = (d_{ij})_{|V|\times|V|}$ *of the graph is a diagonal matrix as*

$$d_{ij} = \begin{cases} d_i, & \text{if } i{=}j, \\ 0, & \text{otherwise.} \end{cases}$$

(3) The Laplacian matrix $L = (l_{ij})_{|V|\times|V|}$ *of a graph is a symmetric matrix with one row and column for each vertex such that*

$$l_{ij} = \begin{cases} d_i, & \text{if } i = j, \\ -w(i,j), & \text{if the edge } (i,j) \text{ exists}, \\ 0, & \text{otherwise.} \end{cases}$$

Clearly, from the definitions, Laplacian matrix satisfies $L = D-W$. Besides this property, there are many propositions of this matrix. In [5, 13] the authors gave some ones, the properties of this matrix-related biclustering will be listed in Proposition 1. Before that the definitions of partitions and cut on graph $G = (V, E)$ are defined.

Definition 4. *A* bipartition of graph *for* $G = (V, E)$ *is defined as two subsets* V_1, V_2 *of* V *such that* $V_1 \cup V_2 = V, V_1 \cap V_2 = \emptyset$.

More generally, a k-partition of graph *is the collection of* k *subsets* $V_1, V_2, \ldots, V_k$ *such that* $V_1 \cup \ldots \cup V_k = V, V_i \cap V_j = \emptyset$ *for* $i, j \in \{1, 2, \ldots, k\}$ *and* $i \neq j$.

In addition, a balanced graph partitioning *is defined as a graph partitioning with the size difference between any two parts at most 1 (almost equal size for all parts).*

For a bipartite graph $G = (U, V, E)$, the graph partitioning will perform on both vertex set U, V, i.e., $U = U_1 \cup U_2, V = V_1 \cup V_2$ where $U_1 \cap U_2 = \emptyset, V_1 \cap V_2 = \emptyset$. Similarly, for k-partition of a bipartite graph, both U and V are partitioned into k disjoint parts.

The balanced graph partitioning is to divide the vertex set into the same size or at most 1 difference in size. So for a k-partition of a graph with n vertices, each part has the size $\llcorner n/k \lrcorner$ or $\llcorner n/k \lrcorner + 1$, where $\llcorner n/k \lrcorner$ is the biggest integer less than or equal to n/k. For a weighted graph, both vertices and edges can be weighted. The balanced graph partitioning is a partition of V into k disjoint parts such that the parts have approximately equal weight (total weight of all vertices within one part).

Definition 5. *Suppose the vertex set* V *of a graph is partitioned into two disjoint subsets* V_1, V_2*, the corresponding graph* cut *is defined as*

$$\mathrm{cut}(V_1, V_2) = \sum_{(i,j)\in E, i\in V_1, j\in V_2} w_{ij}.$$

For the case of k-partition, the k-cut is

$$\mathrm{cut}(V_1, V_2, \ldots, V_k) = \sum_{1\le i<j\le k} \mathrm{cut}(V_i, V_j).$$

An edge with two ends belonging to two different parts is called a cut edge. The fact that $\mathrm{cut}(V_1, V_2) = \mathrm{cut}(V_2, V_1)$ can be easily drawn from that weighted matrix is symmetric.

In the following, the notation $\mathrm{cut}(V_a, V_b)$ is used as the total weight of edges with one end in V_a and another in V_b, whether V_a and V_b are disjoint or not (the two vertex set can be the same, or one is a subset of another). For example, the notation $\mathrm{cut}(V_1, V_1)$ is the sum of weights of edges with two ends in vertex set V_1 and $\mathrm{cut}(V_1, V) = \mathrm{cut}(V_1, V_1 \cup V_2) = 2\mathrm{cut}(V_1, V_1) + \mathrm{cut}(V_1, V_2)$.

3.2 Bipartite Partitioning and Biclustering

Now, bipartite graph is used to model biclustering. Given the "sample–feature" type matrix $A = (a_{ij})_{n\times m}$ with n samples and m features, where a_{ij} is the expression level of feature j in sample i, we construct the corresponding bipartite graph $G = (U, V, E)$ in the following steps:

- The vertex set U represents n samples, and each vertex u_i in U corresponds to a sample $i, 1 \le i \le n$;
- The vertex set V represents m features, and each vertex v_j in V corresponds to a feature $j, 1 \le j \le m$;
- An edge $(i, j) \in E$ with weight $w_{ij} = a_{ij}$ is between a vertex $u_i \in U$ and a vertex $v_j \in V, 1 \le i \le n, 1 \le j \le m$ if $a_{ij} \ne 0$.

The corresponding adjacency weighted matrix of the bipartite graph $G = (U, V, E)$ is expressed in the form of data matrix A as

$$W = (w_{ij})_{(n+m)\times(n+m)} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \tag{1}$$

and the degree matrix of the bipartite graph $G = (U, V, E)$ is

$$D = \begin{pmatrix} D_U & 0 \\ 0 & D_V \end{pmatrix}, \tag{2}$$

where the diagonal elements of D_U and D_V are weighted degree of vertices belonging to U and V, and all other elements are 0.

Thus, the Laplacian matrix of the bipartite graph $G = (U, V, E)$ for data set A is

$$L = D - W = \begin{pmatrix} D_U & -A \\ -A^T & D_V \end{pmatrix}. \tag{3}$$

In this section, we always consider the case of bipartition $\{(U_1 \cup V_1), (U_2 \cup V_2)\}$ of the bipartite graph $G = (U, V, E)$, that is, a biclustering of gene expression data divides the samples and features into two pairs $(\mathcal{S}_1, \mathcal{F}_1)$ and $(\mathcal{S}_2, \mathcal{F}_2)$. Here, the vertices in U_i correspond to rows of $\mathcal{S}_i (i = 1, 2)$ and vertices in V_i correspond to columns of $\mathcal{F}_i (i = 1, 2)$. In this way, a partition of the matrix data A for biclustering is transformed into a bipartition of bipartite graph $G = (U, V, E)$ with weighted matrix W.

4 Integer Programming of Partitioning

In order to classify which part the vertex belongs to, the decision indicator variable is defined as follows.

Definition 6. *The* indicator variable *on $G = (U, V, E)$ to classify the vertices' part belongings is defined as a vector*

$$p = \begin{pmatrix} p_U \\ p_V \end{pmatrix}, \tag{4}$$

where $p_U = (p_1, p_2, \ldots, p_n)^T$ and $p_V = (p_{n+1}, \ldots, p_{n+m})^T$ are used to classify vertex set U and V, respectively, and

$$p_i = \begin{cases} 1, & i \in U_1 \\ -1, & i \in U_2 \end{cases} \text{and } p_j = \begin{cases} 1, & j \in V_1 \\ -1, & j \in V_2 \end{cases}.$$

The indicator variable is a decision variable, that is, we want to obtain such variable after our computation to decide the partitioning of graph. Here we define another useful vector $e = (1, 1, \ldots, 1)^T$ with all elements being 1 (the dimension of this vector we use below will conform with other vectors or matrices).

Proposition 1. *The Laplacian Matrix L in (3) has these propositions:*

(1) L is symmetric positive semidefinite.
(2) 0 is an eigenvalue of L and e is the corresponding eigenvector.
(3) If the graph G has c connected components, then L has c eigenvalues that equal 0.
(4) For any vector x, $x^T L x = \sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2$, and additionally, for any scalars α, β,

$$(\alpha x + \beta e)^T L(\alpha x + \beta e) = \alpha^2 x^T L x.$$

(5) Rayleigh Quotient of L is

$$\frac{p^T L p}{p^T p} = \frac{1}{n+m} \cdot 4\mathrm{cut}(V_1, V_2).$$

(6) Let λ_2 be the second smallest eigenvalue of L,

$$\lambda_2 = \min \frac{x^T L x}{x^T x}, \quad x^T e = 0, x \neq 0.$$

For the proof of the first five properties, we refer to [5, 13]. For the last property, there is a proof with many generalized forms in spectral graph theory [3].

As shown above, the objectives of biclustering are to maximize intra similarity and minimize the inter similarity. Additionally, "to avoid unnatural bias for partitioning out small sets of points" [14], balanced biclusters are considered with respect to different objective functions. Therefore, correspondingly on a graph, the purpose is to find a balanced partitioning such that the corresponding data items in each part are highly related and the weight of cut edges is minimized. Here first the intra and inter bicluster similarities are expressed in the forms of matrices in (1), (2), and (3) and indicator variable (4).

Theorem 1. *For a bipartition $\{(U_1 \cup V_1), (U_2 \cup V_2)\}$ of the bipartite graph $G = (U, V, E)$, we have the following results:*

(1) The intra similarity is

$$\begin{aligned} \mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_2, V_2) &= \sum_{i \in U_1, j \in V_1, (i,j) \in E} w_{ij} + \sum_{i \in U_2, j \in V_2, (i,j) \in E} w_{ij} \quad (5) \\ &= \frac{1}{4} p^T (D + W) p. \end{aligned}$$

(2) The inter similarity is

$$\begin{aligned} \mathrm{cut}(U_1 \cup V_1, U_2 \cup V_2) &= \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1) \quad (6) \\ &= \sum_{\substack{i \in U_1, j \in V_2, (i,j) \in E \\ \text{and } i \in U_2, j \in V_1, (i,j) \in E}} w_{ij} \\ &= \frac{1}{4} p^T (D - W) p = \frac{1}{4} p^T L p. \end{aligned}$$

Proof. The identity (5) refers to the similarities in two groups $U_1 \cup V_1$ and $U_2 \cup V_2$, and it is the total weight of edges within each group. For decision indicator (4), it has four sub-vectors $p_{U_1}, p_{V_1}, p_{U_2}, p_{V_2}$ Where first two have all elements 1 and the other all -1. Similarly, for matrices D_U, D_V, A, we decompose them into corresponding submatrices with respect to p by properly rearranging the matrices, i.e.,

$$D + W = \begin{pmatrix} D_U & A \\ A^T & D_V \end{pmatrix} = \begin{pmatrix} D_{U_1} & 0 & A_{U_1,V_1} & A_{U_1,V_2} \\ 0 & D_{U_2} & A_{U_2,V_1} & A_{U_2,V_2} \\ A_{V_1,U_1} & A_{V_1,U_2} & D_{V_1} & 0 \\ A_{V_2,U_1} & A_{V_2,U_2} & 0 & D_{V_2} \end{pmatrix}.$$

Therefore, using these forms,

$$
\begin{aligned}
&\frac{1}{4}p^T(D+W)p \\
&= \frac{1}{4}(p_{U_1}^T, p_{U_2}^T, p_{V_1}^T, p_{V_2}^T)\begin{pmatrix} D_{U_1} & 0 & A_{U_1,V_1} & A_{U_1,V_2} \\ 0 & D_{U_2} & A_{U_2,V_1} & A_{U_2,V_2} \\ A_{V_1,U_1} & A_{V_1,U_2} & D_{V_1} & 0 \\ A_{V_2,U_1} & A_{V_2,U_2} & 0 & D_{V_2} \end{pmatrix}\begin{pmatrix} p_{U_1} \\ p_{U_2} \\ p_{V_1} \\ p_{V_2} \end{pmatrix} \\
&= \frac{1}{4}(e^T, -e^T, e^T, -e^T)\begin{pmatrix} D_{U_1} & 0 & A_{U_1,V_1} & A_{U_1,V_2} \\ 0 & D_{U_2} & A_{U_2,V_1} & A_{U_2,V_2} \\ A_{V_1,U_1} & A_{V_1,U_2} & D_{V_1} & 0 \\ A_{V_2,U_1} & A_{V_2,U_2} & 0 & D_{V_2} \end{pmatrix}\begin{pmatrix} e \\ -e \\ e \\ -e \end{pmatrix} \\
&= \frac{1}{4}\Big(e^T D_{U_1} e + e^T D_{V_1} e + e^T D_{U_2} e + e^T D_{V_2} e \\
&\quad + e^T A_{U_1,V_1} e - e^T A_{U_1,V_2} e - e^T A_{U_2,V_1} e + e^T A_{U_2,V_2} e \\
&\quad + e^T A_{V_1,U_1} e - e^T A_{V_1,U_2} e - e^T A_{V_2,U_1} e + e^T A_{V_2,U_2} e\Big) \\
&= \frac{1}{4}\left(\sum_{i\in U_1} d_i + \sum_{j\in V_1} d_j + \sum_{i\in U_2} d_i + \sum_{j\in V_2} d_j \right. \\
&\quad + \sum_{i\in U_1, j\in V_1} a_{ij} - \sum_{i\in U_1, j\in V_2} a_{ij} - \sum_{i\in U_2, j\in V_1} a_{ij} + \sum_{i\in U_2, j\in V_2} a_{ij} \\
&\quad \left. + \sum_{j\in V_1, i\in U_1} a_{ij} - \sum_{j\in V_1, i\in U_2} a_{ij} - \sum_{j\in V_2, i\in U_1} a_{ij} + \sum_{j\in V_2, i\in U_2} a_{ij}\right) \\
&= \frac{1}{4}\left(\sum_{i\in U_1, j\in V_1\cup V_2} a_{ij} + \sum_{j\in V_1, i\in U_1\cup U_2} a_{ij} + \sum_{i\in U_2, j\in V_1\cup V_2} a_{ij} + \sum_{j\in V_2, i\in U_1\cup U_2} a_{ij} \right. \\
&\quad \left. +2\left(\sum_{i\in U_1, j\in V_1} a_{ij} - \sum_{i\in U_1, j\in V_2} a_{ij} - \sum_{i\in U_2, j\in V_1} a_{ij} + \sum_{i\in U_2, j\in V_2} a_{ij}\right)\right) \\
&= \frac{1}{4}\cdot 4\left(\sum_{i\in U_1, j\in V_1} a_{ij} + \sum_{i\in U_2, j\in V_2} a_{ij}\right) \\
&= \sum_{i\in U_1, j\in V_1} a_{ij} + \sum_{i\in U_2, j\in V_2} a_{ij},
\end{aligned}
$$

which finishes the proof by the relation of $w_{ij} = a_{ij}$ if $(i,j) \in E, a_{ij} \neq 0$.

Analogically, the identity of (6) refers to inter similarities between two groups $U_1 \cup V_1$ and $U_2 \cup V_2$, and it is the total weight of edges between two groups. The notation of inter similarity is $\mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1) = \mathrm{cut}(U_1 \cup V_1, U_2 \cup V_2)$ from the fact $\mathrm{cut}(U_1, U_2) = \mathrm{cut}(V_1, V_2) = 0$. By Proposition 1 (5) where $p^T p = n + m$ is a constant, this inter similarity is obtained in the form of p and L. And also we can also use the proof of intra similarity to decompose matrices into submatrices. □

Obviously, the biclustering requires that

$$\max \frac{1}{4}p^T(D+W)p \quad \text{and} \quad \min \frac{1}{4}p^T Lp$$

from Theorem 1. However, the objective $\max \frac{1}{4}p^T(D+W)p$ produces "tight" biclusters while $\min \frac{1}{4}p^T Lp$ may produce quite unequal biclusters and isolated vertices. Both are not satisfying the balanced partitioning requirement. Thus, the cut $\frac{1}{4}p^T Lp$ between two parts is called general cut in order to distinguish it from other cuts used in biclustering.

In addition, for a given data matrix A, minimizing the inter similarity $\frac{1}{4}p^T Lp$ is equivalent to maximizing the intra similarity $\frac{1}{4}p^T(D+W)p$ from the fact

$$\frac{1}{4}p^T Lp+\frac{1}{4}p^T(D+W)p=\frac{1}{4}p^T(D-W)p+\frac{1}{4}p^T(D+W)p=\frac{1}{2}p^T Dp=\frac{1}{2}\sum_{i\in U\cup V} d_i,$$

a constant related to A as shown in Theorem 1. Thus, the classic minimizing cut($s-t$ cut) problem can be written by L as the integer programming formulation

$$\begin{aligned} \min \quad & \frac{1}{4}p^T Lp \\ \text{s.t.} \quad & p=(p_1,\ldots,p_{n+m})^T, p_i \in \{-1,1\}. \end{aligned}$$

Many previous research used quite a lot of different objective functions to obtain the balance between "tight" and "quite unequal" biclusters. Among them, two famous cuts have been introduced: Ratio cut and Normalized cut, which are both based on the inter similarity. Besides, some other functions are also reviewed in the following.

Definition 7. *For a bipartition* $\{(U_1 \cup V_1),(U_2 \cup V_2)\}$ *of the bipartite graph* $G=(U,V,E)$*, the* Ratio cut *is defined as*

$$R((U_1 \cup V_1),(U_2 \cup V_2)) = \frac{\text{cut}(U_1 \cup V_1, U_2 \cup V_2)}{|U_1 \cup V_1|} + \frac{\text{cut}(U_2 \cup V_2, U_1 \cup V_1)}{|U_2 \cup V_2|}, \quad (7)$$

and the Normalized cut *is defined as*

$$N((U_1 \cup V_1),(U_2 \cup V_2)) = \frac{\text{cut}(U_1 \cup V_1, U_2 \cup V_2)}{d_{P_1}} + \frac{\text{cut}(U_2 \cup V_2, U_1 \cup V_1)}{d_{P_2}}, \quad (8)$$

where $d_{P_1}=\sum_{i\in(U_1\cup V_1)} d_i, d_{P_2}=\sum_{j\in(U_2\cup V_2)} d_j$.

In the above definitions of Ratio cut and Normalized cut, cut $(U_1 \cup V_1, U_2 \cup V_2)$ is the inter similarity between the bipartitions $\{(U_1 \cup V_1),(U_2 \cup V_2)\}$, and either $|U_i \cup V_i|$ or d_{P_i} can be viewed as the total degree of vertices of group $U_i \cup V_i$. The first one assumes every vertex has weight as 1 and second one chooses weight as the vertex's weighted degree. In addition,

$$\begin{aligned} d_{P_1} &= \sum_{i\in(U_1\cup V_1)} d_i \\ &= \mathrm{cut}(U_1, V) + \mathrm{cut}(U, V_1) \\ &= \mathrm{cut}(U_1, V_1 \cup V_2) + \mathrm{cut}(U_1 \cup U_2, V_1) \\ &= \mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_2, V_1) \\ &= 2\mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1), \end{aligned}$$

and similarly $d_{P_2} = 2\mathrm{cut}(U_2, V_2) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)$.

In biclustering, both (3) and (8) cuts have to be minimized since they are both generalized forms of cut in Definition 5. Furthermore, for Normalized cut, it can be transformed into an equivalent maximum problem.

Theorem 2. *To minimize the Normalized cut (8) is equivalent to maximize*

$$\frac{\mathrm{cut}(U_1, V_1)}{d_{P_1}} + \frac{\mathrm{cut}(U_2, V_2)}{d_{P_2}}. \tag{9}$$

Proof. From the facts that $d_{P_1} = 2\mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)$ and $d_{P_2} = 2\mathrm{cut}(U_2, V_2) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)$,

$$\begin{aligned} & 2\left(\frac{\mathrm{cut}(U_1, V_1)}{d_{P_1}} + \frac{\mathrm{cut}(U_2, V_2)}{d_{P_2}}\right) \\ &= \frac{2\mathrm{cut}(U_1, V_1)}{2\mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)} + \\ &\quad + \frac{2\mathrm{cut}(U_2, V_2)}{2\mathrm{cut}(U_2, V_2) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)} \\ &= 1 - \frac{\mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)}{2\mathrm{cut}(U_1, V_1) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)} + \\ &\quad + 1 - \frac{\mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)}{2\mathrm{cut}(U_2, V_2) + \mathrm{cut}(U_1, V_2) + \mathrm{cut}(U_2, V_1)} \\ &= 2 - \frac{\mathrm{cut}(U_1 \cup V_1, U_2 \cup V_2)}{d_{P_1}} - \frac{\mathrm{cut}(U_2 \cup V_2, U_1 \cup V_1)}{d_{P_2}} \\ &= 2 - N((U_1 \cup V_1), (U_2 \cup V_2)). \end{aligned}$$

Thus to minimize $N(V_1, V_2)$ is equivalent to maximize $\frac{\mathrm{cut}(U_1,V_1)}{d_{P_1}} + \frac{\mathrm{cut}(U_2,V_2)}{d_{P_2}}$. □

The cuts $\mathrm{cut}(U_1, V_1)$, $\mathrm{cut}(U_2, V_2)$ in Theorem 2 are intra similarity of two parts $U_1 \cup V_1, U_2 \cup V_2$ of the bipartition, respectively, and from this theorem, a cut based on inter similarity has been transformed into one based on intra similarity. This theorem also indicates that Normalized cut is similar to general cut that both have the property of minimizing inter similarity being equivalent to maximize intra similarity.

4.1 Ratio Cut

Now we are beginning to use the defined cuts for bipartition on bipartite graph $G = (U, V, E)$ and obtain its corresponding biclusters of data matrix A. Assume that $|U_1| = n_1, |U_2| = n_2$ with $n_1 + n_2 = n$ and $|V_1| = m_1, |V_2| = m_2$ with $m_1 + m_2 = m$.

Theorem 3. *Defined the indicator vector $y = (y_u, y_v)$ as*

$$y_i = \begin{cases} \sqrt{(n_2 + m_2)/((n_1 + m_1)(n + m))}, & i \in U_1 \cup V_1 \\ -\sqrt{(n_1 + m_1)/((n_2 + m_2)(n + m))}, & i \in U_2 \cup V_2 \end{cases}$$

where y_u is the vector of elements of y_i with $i \in U_1 \cup U_2$ and y_v is the vector of elements of y_i with $i \in V_1 \cup V_2$, the Ratio cut of $\{(U_1 \cup V_1), (U_2 \cup V_2)\}$ of the bipartite graph $G = (U, V, E)$ can be expressed as $y^T Ly$.

Proof. Since the Ratio cut $R((U_1 \cup V_1), (U_2 \cup V_2))$ can be expressed as

$$\begin{aligned} R((U_1 \cup V_1), (U_2 \cup V_2)) &= \frac{\text{cut}(U_1 \cup V_1, U_2 \cup V_2)}{n_1 + m_1} + \frac{\text{cut}(U_2 \cup V_2, U_1 \cup V_1)}{n_2 + m_2} \\ &= \frac{\text{cut}(U_1 \cup V_1, U_2 \cup V_2) \times ((n_1 + m_1) + (n_2 + m_2))}{(n_1 + m_1)(n_2 + m_2)} \\ &= \text{cut}(U_1 \cup V_1, U_2 \cup V_2) \times \frac{(n + m)}{(n_1 + m_1)(n_2 + m_2)}, \end{aligned}$$

and by Theorem 1(2), the Ratio cut is

$$R((U_1 \cup V_1), (U_2 \cup V_2)) = \frac{1}{4} p^T Lp \times \frac{(n + m)}{(n_1 + m_1)(n_2 + m_2)},$$

where p is the indicator variable from Definition 4 and L is Laplacian matrix.

The vector y can be written as

$$y = \frac{n + m}{2\sqrt{(n_1 + m_1)(n_2 + m_2)(n + m)}} p + \frac{n_2 + m_2 - n_1 - m_1}{2\sqrt{(n_1 + m_1)(n_2 + m_2)(n + m)}} e,$$

and by Proposition 1 (4) of Laplacian matrix L,

$$\begin{aligned} y^T Ly &= \left(\frac{n + m}{2\sqrt{(n_1 + m_1)(n_2 + m_2)(n + m)}} \right)^2 p^T Lp \\ &= \frac{(n + m)}{4(n_1 + m_1)(n_2 + m_2)} p^T Lp. \end{aligned}$$

Therefore, the Ratio cut can be expressed by y and L as

$$R((U_1 \cup V_1), (U_2 \cup V_2)) = y^T Ly.$$

□

Formally, to minimize Ratio cut for biclustering can be modeled as the following mixed binary integer program:

$$\begin{aligned}
\min \quad & y^T L y \qquad (10)\\
\text{s.t.} \quad & y = \frac{n+m}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}}p \\
& \quad + \frac{n_2+m_2-n_1-m_1}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}}e, \\
& p = (p_1, \ldots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\
& n_1 + m_1 = \sum_i (p_i+1)/2, \\
& n_2 + m_2 = \sum_i (1-p_i)/2, \\
& n_1 + n_2 = n, m_1 + m_2 = m \\
& p_i \in \{-1, 1\}, i = 1, \ldots, n+m.
\end{aligned}$$

As in Theorem 3, the elements of vector y can be either positive or negative which indicates the part belongings of each vertex, and y has the property that $y^T y = 1, y^T e = 0$ implying from the constraints. These properties of y are summarized in the following theorem.

Theorem 4. *The nonzero vector y defined in Theorem 3 satisfies the identities $y^T y = 1$ and $y^T e = 0$.*

Proof. The property of y is formulated by the following steps:

$$\begin{aligned}
y^T y &= \sum_{i=1}^{n+m} \left(\frac{n+m}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} p_i + \frac{n_2+m_2-n_1-m_1}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} \right)^2 \\
&= \sum_{i=1}^{n+m} \left(\frac{(n+m)^2}{4(n_1+m_1)(n_2+m_2)(n+m)} p_i^2 + \frac{(n_2+m_2-n_1-m_1)^2}{4(n_1+m_2)(n_2+m_2)(n+m)} + \right. \\
&\quad \left. + \frac{(n+m)(n_2+m_2-n_1-m_1)}{2(n_1+m_1)(n_2+m_2)(n+m)} p_i \right) \\
&= \frac{(n+m)^2}{4(n_1+m_1)(n_2+m_2)} + \frac{(n_2+m_2-n_1-m_1)^2}{4(n_1+m_2)(n_2+m_2)} + \\
&\quad + \frac{(n_1+m_1-n_2-m_2)(n_2+m_2-n_1-m_1)}{2(n_1+m_1)(n_2+m_2)} \\
&= \frac{(n+m)^2 - (n_2+m_2-n_1-m_1)^2}{4(n_1+m_1)(n_2+m_2)} \\
&= \frac{(n+m+n_2+m_2-n_1-m_1)(n+m-n_2-m_2+n_1+m_1)}{4(n_1+m_1)(n_2+m_2)} \\
&= \frac{(2n_2+2m_2)(2n_1+2m_1)}{4(n_1+m_1)(n_2+m_2)} \\
&= 1,
\end{aligned}$$

and for the identity $y^T e = 0$,

$$\begin{aligned} y^T e &= \sum_{i=1}^{n+m} \left(\frac{n+m}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} p_i + \frac{n_2+m_2-n_1-m_1}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} \right) \\ &= \frac{n+m}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} \sum_{i=1}^{n+m} p_i + \frac{n_2+m_2-n_1-m_1}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} (n+m) \\ &= \frac{n+m}{2\sqrt{(n_1+m_1)(n_2+m_2)(n+m)}} [(n_1+m_1-n_2-m_2)+(n_2+m_2-n_1-m_1)] \\ &= 0. \end{aligned}$$

□

Therefore, a relaxation of this formulation can be solved for bipartition based on this property of y, and the relaxation program is

$$\begin{aligned} \min \quad & y^T L y \\ \text{s.t.} \quad & y^T y = 1,\ y^T e = 0,\ y \neq 0, \end{aligned}$$

where y is any real number. In addition, from Proposition 1 (6) and the fact $y^T y = 1$, a constant, the above formulation can be written as

$$\begin{aligned} \min \quad & \frac{y^T L y}{y^T y} \\ \text{s.t.} \quad & y^T e = 0, y \neq 0, \end{aligned} \tag{11}$$

with solution of objective value as second smallest eigenvalue and y as corresponding eigenvector. This gives the reason in [5, 8, 16] why they can use formulation (11) to do biclustering. The sign of elements of y, either positive or negative, is used to classify the two groups as $p_i = 1$ and $p_i = -1$.

Another easier but equivalent form of formulation (10) is given by the following binary integer program

$$\begin{aligned} \min \quad & \frac{n+m}{4(n_1+m_1)(n_2+m_2)} p^T L p \\ \text{s.t.} \quad & p = (p_1, \ldots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\ & n_1 + m_1 = \sum_i (p_i + 1)/2, \\ & n_2 + m_2 = \sum_i (1 - p_i)/2, \\ & n_1 + n_2 = n, m_1 + m_2 = m \\ & p_i \in \{-1, 1\}, i = 1, \ldots, n+m. \end{aligned} \tag{12}$$

4.2 Normalized Cut and Minmax Cut

For Normalized cut, define the indicator variable

$$\begin{aligned} y_i &= \begin{cases} \sqrt{d_{P_2}/(d_{P_1}d)}, & i \in U_1 \cup V_1, \\ -\sqrt{d_{P_1}/(d_{P_2}d)}, & i \in U_2 \cup V_2, \end{cases} \\ &= \frac{d_{P_1}+d_{P_2}}{2\sqrt{d_{P_1}d_{P_2}d}}p_i + \frac{d_{P_2}-d_{P_1}}{2\sqrt{d_{P_1}d_{P_2}d}}e_i, \end{aligned} \tag{13}$$

where $d_{P_1} = \sum_{i\in U_1\cup V_1} d_i, d_{P_2} = \sum_{j\in U_2\cup V_2} d_j$, $d = d_{P_1} + d_{P_2}$, and $(p_1, \cdots, p_{n+m}) = p$ is decision indicator variable defined in Definition 4.

The Normalized cut of bipartition $\{(U_1 \cup V_1), (U_2 \cup V_2)\}$ of the bipartite graph $G = (U, V, E)$ can also be expressed as $y^T Ly$, which can be proved with the similar methods in Theorem 3. We present it as a theorem without proof in the following.

Theorem 5. *With the y defined in* (13)*, the Normalized cut* (8) *can be expressed as*

$$N((U_1 \cup V_1), (U_2 \cup V_2)) = y^T Ly.$$

Thus, by this relation of Normalized cut, the problem of minimizing Normalized cut is the following mixed binary integer program

$$\begin{aligned} \min \quad & y^T Ly \\ \text{s.t.} \quad & y = \frac{d_{P_1}+d_{P_2}}{2\sqrt{d_{P_1}d_{P_2}d}}p + \frac{d_{P_2}-d_{P_1}}{2\sqrt{d_{P_1}d_{P_2}d}}e, \\ & p = (p_1, \ldots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\ & p_i \in \{-1, 1\}, i = 1, \ldots, n+m, \\ & d_{P_1} = \sum_{i:p_i=1} d_i, \\ & d_{P_2} = \sum_{j:p_j=-1} d_j, \\ & d = d_{P_1} + d_{P_2}. \end{aligned} \tag{14}$$

Now, the constraints within the above formulation have the properties $y^T Dy = 1, y^T De = 0$, which is different from those in Ratio cut.

Theorem 6. *The nonzero vector y defined in formulation* (14) *satisfies the identities* $y^T Dy = 1$ *and* $y^T De = 0$.

Proof. Since D is a diagonal matrix with nonzero elements $d_1, \ldots, d_{n+m}$ on its diagonal, and from $d_{P_1} = \sum_{i:p_i=1} d_i, d_{P_2} = \sum_{j:p_j=-1} d_j, d = d_{P_1} + d_{P_2}$,

$$
\begin{aligned}
y^T Dy = & \sum_{i=1}^{n+m} d_i y_i^2 \\
= & \sum_{i=1}^{n+m} d_i \left(\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} p_i + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} e_i \right)^2 \\
= & \sum_{p_i=1} d_i \left(\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right)^2 + \\
& + \sum_{p_i=-1} d_i \left(-\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right)^2 \\
= & \, d_{P_1} \left(\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right)^2 + \\
& + d_{P_2} \left(-\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right)^2 \\
= & \, d_{P_1} \frac{d^2 + 2(d_{P_2}^2 - d_{P_1}^2) + (d_{P_2} - d_{P_1})^2}{4 d_{P_1} d_{P_2} d} + \\
& + d_{P_2} \frac{d^2 - 2(d_{P_2}^2 - d_{P_1}^2) + (d_{P_2} - d_{P_1})^2}{4 d_{P_1} d_{P_2} d} \\
= & \, \frac{d^3 + d(d_{P_2} - d_{P_1})^2 + 2(d_{P_1} - d_{P_2})(d_{P_2}^2 - d_{P_1}^2)}{4 d_{P_1} d_{P_2} d} \\
= & \, \frac{d^3 + d(d_{P_2} - d_{P_1})^2 - 2d(d_{P_2} - d_{P_1})^2}{4 d_{P_1} d_{P_2} d} \\
= & \, \frac{d^2 - (d_{P_2} - d_{P_1})^2}{4 d_{P_1} d_{P_2}} \\
= & \, \frac{(d + d_{P_2} - d_{P_1})(d - d_{P_2} + d_{P_1})}{4 d_{P_1} d_{P_2}} \\
= & \, 1,
\end{aligned}
$$

and for another identity,

$$
\begin{aligned}
y^T De = & \sum_{i=1}^{n+m} d_i y_i \\
= & \sum_{i=1}^{n+m} d_i \left(\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} p_i + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} e_i \right) \\
= & \, d_{P_1} \left(\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right) + d_{P_2} \left(-\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + \frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= (d_{P_1} - d_{P_2})\frac{d_{P_1} + d_{P_2}}{2\sqrt{d_{P_1} d_{P_2} d}} + (d_{P_1} + d_{P_2})\frac{d_{P_2} - d_{P_1}}{2\sqrt{d_{P_1} d_{P_2} d}} \\
&= \frac{(d_{P_1} + d_{P_2})(d_{P_1} - d_{P_2}) - (d_{P_1} + d_{P_2})(d_{P_1} - d_{P_2})}{2\sqrt{d_{P_1} d_{P_2} d}} \\
&= 0.
\end{aligned}
$$

□

Thus by the properties of y, the above formulation is relaxing to

$$
\begin{aligned}
\min \quad & \frac{y^T L y}{y^T D y} \\
\text{s.t.} \quad & y^T D e = 0, y \neq 0,
\end{aligned}
\tag{15}
$$

and by Proposition 1, the solution of this problem is to find the second smallest eigenvalue and corresponding eigenvector of generalized eigenvalue problem $Ly = \lambda Dy$. This also gives a proof that the Normalized cut can be solved by generalized eigenvalue problem in [5, 14].

Similarly as Ratio cut, the equivalent binary integer program for Normalized cut in formulation (14) is

$$
\begin{aligned}
\min \quad & \frac{d}{4 d_{P_1} d_{P_2}} p^T L p \\
s.t. \quad & p = (p_1, \ldots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\
& p_i \in \{-1, 1\}, i = 1, \ldots, n+m, \\
& d_{P_1} = \sum_{i:p_i=1} d_i, \\
& d_{P_2} = \sum_{j:p_j=-1} d_j, \\
& d = d_{P_1} + d_{P_2}.
\end{aligned}
\tag{16}
$$

For the two constraints $d_{P_1} = \sum_{i:p_i=1} d_i$ and $d_{P_2} = \sum_{j:p_j=-1} d_j$, they can be written as $d_{P_1} = \sum_{i=1}^{n} \frac{p_i+1}{2} d_i$ and $d_{P_2} = \sum_{i=1}^{n} \frac{1-p_i}{2} d_i$.

For the equivalent form of maximizing as shown in Theorem 2, we first show that the form of intra similarity of Normalized cut can be expressed as $\frac{1}{2} y^T (D + W) y$, i.e.,

$$
\frac{\mathrm{cut}(U_1, V_1)}{d_{P_1}} + \frac{\mathrm{cut}(U_2, V_2)}{d_{P_2}} = \frac{1}{2} y^T (D + W) y.
$$

In fact, from the proof of Theorem 2 and $y^T D y = 1$ from Theorem 6, we have the identities

$$
\begin{aligned}
2\left(\frac{\mathrm{cut}(U_1,V_1)}{d_{P_1}}+\frac{\mathrm{cut}(U_2,V_2)}{d_{P_2}}\right) &= 2-N((U_1\cup V_1),(U_2\cup V_2))\\
&= 2y^TDy-y^TLy\\
&= 2y^TDy-y^T(D-W)y\\
&= y^T(D+W)y.
\end{aligned}
$$

From Theorem 1(1), this is the intra similarity, which should be maximized in biclustering. The relaxation program for this is

$$
\begin{aligned}
\max \quad & \frac{y^T(D+W)y}{2}\\
\text{s.t} \quad & y^TDy=1, y^TDe=0, y\neq 0,
\end{aligned}
$$

or in the form of

$$
\begin{aligned}
\max \quad & \frac{y^T(D+W)y}{2y^TDy}=\frac{1}{2}+\frac{y^TWy}{2y^TDy}\\
\text{s.t.} \quad & y^TDe=0, y\neq 0,
\end{aligned}
\tag{17}
$$

In [6], Ding et al. defined a Minmax cut of bipartition $\{(U_1\cup V_1),(U_2\cup V_2)\}$ of $G=(U,V,E)$ as

$$
\text{Minmax Cut}=\frac{\mathrm{cut}(U_1\cup V_1,U_2\cup V_2)}{\mathrm{cut}(U_1\cup V_1,U_1\cup V_1)}+\frac{\mathrm{cut}(U_2\cup V_2,U_1\cup V_1)}{\mathrm{cut}(U_2\cup V_2,U_2\cup V_2)}
$$

and used the indicator vector as in (13). Then they proved that minimizing Minmax cut is equivalent to max $\frac{y^TWy}{y^TDy}$ with constraints $y^TDe=0, y\neq 0$. As we have shown above, this kind of Minmax cut is equivalent to Normalized cut for biclustering.

To solve the formulation (17), the method is same as formulation (15) with the solution y as the eigenvector corresponding to second smallest eigenvalue of generalized eigenvalue problem $Ly=\lambda Dy$.

4.3 ICA Cut

In [13], they defined a cut called ICA (Isoperimetric co-clustering) cut of bipartition $\{(U_1\cup V_1),(U_2\cup V_2)\}$ of $G=(U,V,E)$ as

$$
\frac{\mathrm{cut}(U_1\cup V_1,U_2\cup V_2)}{d_{P_1}}
$$

and let the indicator vector be

$$
y_i=\begin{cases}1/\sqrt{d_{P_1}}, & i\in U_1\cup V_1,\\ 0, & i\in U_2\cup V_2.\end{cases}
$$

This kind of ICA cut can be expressed as $y^T Ly = \frac{1}{4d_{P_1}} p^T Lp$, and the mixed integer program with ICA cut as objective function is

$$\begin{aligned} \min \quad & y^T Ly \\ \text{s.t.} \quad & y = \frac{1}{2\sqrt{d_{P_1}}}(p+e), \\ & p = (p_1, \ldots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\ & p_i \in \{-1, 1\}, i = 1, \ldots, n+m, \\ & d_{P_1} = \sum_{i:p_i=1} d_i. \end{aligned} \tag{18}$$

The decision variable y has the property $y^T De = \sum_{i=1}^{n+m} d_i y_i = d_{P_1} \frac{1}{\sqrt{d_{P_1}}} = \sqrt{d_{P_1}}$. If the volume d_{P_1} of $U_1 \cup V_1$ is fixed as a constant c^2, where $0 < c^2 < \sum_i d_i$, $y^T De = c$. A relaxation form of the above formulation can be stated as

$$\begin{aligned} \min \quad & y^T Ly \\ \text{s.t.} \quad & y^T De = c, y \neq 0. \end{aligned} \tag{19}$$

In [13], the Lagrange multiplier is used to solve it as

$$\frac{d(y^T Ly - \lambda(y^T De - c))}{dy} = 2Ly - \lambda De,$$

and assuming it to be zero and ignoring the $\lambda, 2$, they solve $Ly = De$ to obtain y. The integer programming form for this problem is

$$\begin{aligned} \min \quad & \frac{1}{4d_{P_1}} p^T Lp \\ \text{s.t.} \quad & p = (p_1, \cdots, p_n, p_{n+1}, \ldots, p_{n+m})^T, \\ & p_i \in \{-1, 1\}, i = 1, \ldots, n+m, \\ & d_{P_1} = \sum_{i:p_i=1} d_i. \end{aligned} \tag{20}$$

Whether what kind of objective functions or cuts are used, the above models all divide both samples and features into two groups. In [5], Dhillon used other models to obtain k groups based on above optimization models instead of hierarchical method. The idea is to use k-means algorithm on the obtained eigenvector to obtain the desired k parts of biclustering.

In the above of using Ratio cut, Normalized cut or others, the general steps of doing biclustering are

- Choosing or defining objective function with respect to cut and weights of vertices or edges;

- Defining indicator vector y and finding the relation with p, e;
- Using the Propositions of L to rewrite the objective function and find the constraints; and
- Designing algorithms to solve the optimization models.

4.4 Spectral and Integer Programming Biclustering

Since the Laplacian matrix of a graph is widely studied in graph theory, called spectral graph theory [3], this method of biclustering for solving problems (11), (15), (17) is used as the term of spectral biclustering [1]. In the book [3], Chung has demonstrated the spectral graph theory and its application for the isoperimetric problem. In spectral biclustering, the problem is to concentrate on computing the eigenvalues and eigenvectors. However, for large-scale data matrix, this is still computationally difficult.

For integer programming for biclustering in (12), (16), (20), these are all nonlinear integer programming models. The methods for solving nonlinear programming can be found in [10] and also some methods from mixed (nonlinear) integer programming such as Outer Approximation methods, Branch-and-Bound, Extended Cutting Plane methods, and Generalized Benders Decomposition.

5 Conclusion

In this chapter, different measurements of cut are transformed into optimization models after properly choosing the indicator variables. This gave a general approach to use optimization models based on Laplacian matrix from data matrix for biclustering. How to solve these optimization models more efficiently is still under future considerations.

The Ratio cut, Normalized cut, and Minmax cut all have the relaxation forms that can be solved by computing the eigenvalues and eigenvectors of matrices. In addition, we show that Minmax cut is equivalent to Normalized cut for biclustering.

References

1. Busygin, S., Prokopyev, O., Pardalos, P.M.: Biclustering in data mining. Comp. Oper. Res. 35, 2964–2987 (2008)
2. Chaovalitwongse, W.A., Butenko, S., Pardalos, P.M.: Clustering Challenges in Biological Networks, World Scientific, Singapore (2009)
3. Chung, F.R.K.: Spectral graph theory, Conference Board of the Mathematical Sciences (Vol. 92), American Mathematical Society, Washington, DC (1997)
4. Cheng, Y., Church, G.M.: Biclustering of expression data. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology 93–103 (2000)

5. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. Proceedings of the 7th ACM SIGKDD International Conference on Knowledging Discovery and Data Mining (KDD), 26–29 (2001)
6. Ding, C.H.Q., He, X., Zha, H., Gu, M., Simon, H.D.: A min-max cut algorithm for graph partitioning and data clustering. Proc. ICDM, 107–114 (2001)
7. Fan, N., Boyko, N., Pardalos, P.M.: Recent Advances of Data Biclustering with Application in Computational Neuroscience. In: W.A. Chaovalitwongse, P.M. Pardalos, P. Xanthopoulos (Eds.), Computational Neuroscience, Springer Optimization and Its Applications (Vol. 38), Springer, Berlin (2010)
8. Hagen, L., Kahng, A.B.: New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Comput-Aided Des. 11(9), 1074–1085 (1992).
9. Lee, C.-H., Zaine, O.R., Park, H.-H., Huang, J., Greiner, R.: Clustering high dimensional data: A graph-based relaxed optimization approach. Inf. Sci. 178(23), 4501–4511 (2008)
10. Li, D., Sun, X.: Nonlinear integer programming, series. Int. Ser. Operat. Res. Manage Sci. 84 (2006)
11. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE Trans. Comput. Biol. Bioinform. 1(1), 24–45 (2004)
12. Mirkin, B.: Mathematical Classification and Clustering, Kluwer, Dordrecht (1996)
13. Rege, M., Dong, M., Fotouhi, F.: Bipartite isoperimetric graph partitioning for data co-clustering. Data Min. Know. Disc, 16, 276–312 (2008)
14. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intelli 22(8), 888–905 (2000)
15. Xu, R., Wunsch II, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw., 16(3), 645–678 (2005)
16. Zha, H., He, X., Ding, C., Simon, H., Gu, M.: Bipartite graph partitioning and data clustering. Proceedings of the 10th International Conference on Information and Knowledge Mmanagement, 25–32 (2001)

A Random Arrival Time Best-Choice Problem with Uniform Prior on the Number of Arrivals

Mitsushi Tamaki[1] and Qi Wang[2]

[1] Department of Business Administration, Aichi University, Aichi, Japan
tamaki@vega.aichi-u.ac.jp
[2] Department of Business Administration, Aichi University, Aichi, Japan
07dm1401@moon.aichi-u.ac.jp

Summary. Suppose that a random number N of rankable applicants appear and their arrival times are i.i.d. random variables having a known distribution function. A method of choosing the best applicant is investigated when a prior on N is uniform on $\{1, 2, \ldots, n\}$. An exact form of the optimal selection rule is derived. Stewart first studied this problem, but examined only the case of the non-informative prior, i.e., the limiting case of $n \to \infty$, so our result can be considered as a generalization of Stewart's result.

Key words: secretary problem, optimal stopping, bayesian updating, OLA rule, e^{-1}-rule, relative rank

1 Introduction

We first review some well-known results for the *classical best-choice problem* and its variation. A known number n of applicants appear one by one in random order with all $n!$ permutations equally likely. As each applicant appears, we rank the applicant relative to those preceding him and decide to either select or reject the current applicant with the objective of maximizing the probability of success, i.e., choosing the very best. An applicant once rejected cannot be recalled later. For convenience we call an applicant a *candidate* if he is better than all his predecessors. Clearly we never stop with an applicant who is not a candidate. It is well known that, in this classical best-choice problem, the optimal rule is of the threshold type with value $s_1(n)$ described as follows : Let $s_1(n) - 1$ applicants go by, and select the first candidate, if any, from time $s_1(n)$ onward, where $s_1(n)$ is computed as

$$s_1(n) = \min\left\{ k \geq 1 : \sum_{j=k+1}^{n} \frac{1}{j-1} \leq 1 \right\}. \tag{1}$$

A. Chinchuluun et al. (eds.), *Optimization and Optimal Control*,
Springer Optimization and Its Applications 39, DOI 10.1007/978-0-387-89496-6_24,

Evidently $s_1(n)/n \to e^{-1}$ as $n \to \infty$. The values of $s_1(n)$ are given in [10].

Reference [11] is the first to introduce the uncertainty about the number N of available applicants. A prior distribution of N, i.e., $p_m = P\{N = m\}, m \geq 1$, is given and, conditional on $N = m$, $m!$ arrival orders are assumed to be equally likely. They studied the best-choice problem with such priors as uniform, Poisson, and geometric. In the uniform case having a prior

$$p_m = \frac{1}{n}, \quad 1 \leq m \leq n, \tag{2}$$

for a known n, they found that the optimal rule is also of the threshold type with value $s_2(n)$, defined as

$$s_2(n) = \min\left\{k \geq 1 : \sum_{i=k}^{n} \frac{1}{i}\left(1 - \sum_{j=k+1}^{i} \frac{1}{j-1}\right) \geq 0\right\}, \tag{3}$$

where the vacuous sum is assumed to be zero. It can be shown that $s_2(n) \leq s_1(n)$ and $s_2(n)/n \to e^{-2}$ as $n \to \infty$.

Instead of having the applicants appear in discrete time, we may have them appear in continuous time. As such one, the *random arrival time best-choice problem* can be described as follows : Let $X_1, X_2, \ldots, X_N$ be continuous i.i.d. random variables with values in $[0, T]$ possibly infinite and common c.d.f. F, where N is an integer-valued random variable independent of X_k's. X_k is thought of as the arrival time of the kth best applicant and N represents the total number of applicants. The objective is to maximize the probability of success. F is assumed to be uniform on $[0, 1]$ without loss of generality, because a change of time $Z_k = F(X_k), 1 \leq k \leq N$ makes Z_k uniform on $[0, 1]$.

In this chapter we consider a random arrival time best-choice problem having the uniform prior given by (2). Reference [15] first studied this problem, but examined only the case of the non-informative prior, i.e., the limiting case of $n \to \infty$ in the prior. He showed that the optimal rule has the following simple form, called e^{-1}-*rule* by [2] later: Reject all the applicants that appear before time e^{-1} and then select the first candidate if any (Stewart took F to be exponential, so the statement is here adjusted to the uniform). We will review this in the Remark of Section 2. However, when n is finite, the optimal rule becomes complicated because it then depends not only on the arrival time of the candidate but also on the number of arrivals observed up to that time. Our main result can be summarized as follows. This proof is given in Section 2.

Theorem 1. *Define, for* $n \geq 2$,

$$s_3(n) = \min\left\{k \geq 1 : \sum_{j=k+1}^{n} \frac{1}{j} \leq 1\right\}. \tag{4}$$

Then there exists a non-decreasing sequence $\{t_k^*(n)\colon 1 \leq k < s_3(n)\}$ *such that the optimal rule chooses the kth applicant (i.e., kth arrival) if and only if*

he is a candidate and appears at time later than $1 - t_k^*(n)$*, where* $t_k^*(n)$ *is a unique solution* $t \in (0,1)$ *to the equation*

$$\sum_{i=0}^{n-k} \frac{(i+k-1)!}{i!} t^i = \sum_{i=1}^{n-k} \frac{(i+k-1)!}{i!} \left(\sum_{j=1}^{i} \frac{1}{j+k-1} \right) t^i, \tag{5}$$

while, if no stoppage has occurred on the first $s_3(n) - 1$ *arrivals, the optimal rule chooses any candidate if any, irrespective of his arrival time. Moreover we have* $s_3(n) = s_1(n)$ *or* $s_1(n) - 1$ *in addition to the obvious relation* $s_3(n) = s_1(n+1) - 1$.

Table 1 presents the numerical values of $s_3(n)$ and $\{t_k^*(n)\}$ for specified values of n. We observe that, for a given k, $t_k^*(n)$ is decreasing in n and approaches the value $0.6321 = 1 - e^{-1}$ very quickly. From this table and the fact that the optimality of the e^{-1}-rule for the non-informative case, it is easily conjectured that, as n tends to infinity, $t_k^*(n)$ converges to $1 - e^{-1}$ for each k. This can be confirmed by Lemma 6.

Table 1. $s_3(n)$ and $\{t_k^*(n)\}$ for several values of n

n	$s_3(n)$	$t_1^*(n)$	$t_2^*(n)$	$t_3^*(n)$	$t_4^*(n)$	$t_5^*(n)$	$t_6^*(n)$	$t_7^*(n)$	$t_8^*(n)$	$t_9^*(n)$
2	1									
3	1									
4	2	0.8957								
5	2	0.7561								
6	2	0.6987								
7	3	0.6704	0.8572							
8	3	0.6551	0.7554							
9	3	0.6462	0.7057							
10	4	0.6409	0.6782	0.8300						
11	4	0.6377	0.6619	0.7499						
12	5	0.6357	0.6517	0.7069	0.9432					
13	5	0.6344	0.6451	0.6815	0.8088					
14	5	0.6336	0.6408	0.6655	0.7432					
15	6	0.6331	0.6380	0.6550	0.7058	0.8961				
16	6	0.6327	0.6361	0.6480	0.6826	0.7914				
17	6	0.6325	0.6348	0.6432	0.6674	0.7365				
18	7	0.6324	0.6339	0.6399	0.6571	0.7037	0.8612			
19	7	0.6323	0.6333	0.6376	0.6500	0.6825	0.7770			
20	8	0.6322	0.6329	0.6359	0.6449	0.6683	0.7302	0.9770		
30	11	0.6321	0.6321	0.6322	0.6326	0.6338	0.6369	0.6441	0.6594	0.6931
40	15	0.6321	0.6321	0.6321	0.6321	0.6322	0.6324	0.6329	0.6341	0.6369
50	19	0.6321	0.6321	0.6321	0.6321	0.6321	0.6321	0.6322	0.6322	0.6325

Let $n = 100$. Then $s_3(100) = 37$ and the values of $t_k^* = t_k^*(100)$ for some selected values of k are given in Table 2.

Table 2. $t_k^* = t_k^*(100)$ for several values of k

k	1–20	25	30	31	32	33	34	35	36
t_k^*	0.6321	0.6329	0.6417	0.6469	0.6548	0.6673	0.6883	0.7287	0.8370

A random arrival time best-choice problem with Poisson prior, which is equivalent to the best-choice problem with a Poisson arrival process, was studied by [8] and [3]. As for the same problem with geometric prior, see [5] and [6]. We are so far concerned with the form of the optimal rule and not with the success probability. Reference [2] recognized the importance of the e^{-1}-rule in the sense that the e^{-1}-rule achieves the success probability greater than e^{-1}, the asymptotic success probability of the classical best-choice problem, whatever the prior distribution of N might be, and [4] generalized this result to the problem with general loss function by embedding the process in the so-called infinite secretary problem defined by [9].

2 Proof of Theorem 1

Let $\{N(t), 0 \le t \le 1\}$ be a counting process defined as

$$N(t) = \sharp\{Z_k : Z_k \le t\}$$

with $N(1) = N$ and focus our attention on the posterior distribution $P\{N = m \mid \mathcal{F}_t\}$ where $\mathcal{F}_t$ denotes the σ-algebra generated by $\{N(s) : s \le t\}$. The posterior distribution depends on the prior $\{p_m\}_{m=1}^{\infty}$, parameter t, and the observation $N(t)$ because of the i.i.d. assumption of the arrival times. Thus the straightforward application of Bayes formulae yields (see, e.g., [6])

$$\begin{aligned} P\{N = m \mid \mathcal{F}_t\} &= \frac{\binom{m}{N(t)} t^{N(t)}(1-t)^{m-N(t)} p_m}{\sum_{k=N(t)}^{\infty} \binom{k}{N(t)} t^{N(t)}(1-t)^{k-N(t)} p_k}, \\ &= \frac{\binom{m}{N(t)}(1-t)^m p_m}{\sum_{k=N(t)}^{\infty} \binom{k}{N(t)}(1-t)^k p_k}. \end{aligned} \tag{6}$$

Let (k,t) denote the state in which we are facing the kth applicant at time $1-t$ who is a candidate (note that t is not an elapsed time but represents the remaining time). Now that the prior distribution is given by (2), the posterior distribution just after observing state (k,t) is given by

$$p(m \mid k,t) = \frac{\binom{m}{k} t^m}{C(k,t)} \tag{7}$$

from (6), where $C(k,t) = \sum_{j=k}^{n} \binom{j}{k} t^j$.

Suppose that we are in state (k,t). Then we have to decide to either select or reject the current candidate. Let $P(k,t)$ be the probability of success by selecting the current candidate. We can compute $P(k,t)$ by conditioning on N. It is easy to see that, conditional on $N=m$, $m!$ arrival orders of these applicants are equally likely, and hence the conditional success probability is given by k/m. Thus unconditioning on N yields

$$\begin{aligned} P(k,t) &= \sum_{m=k}^{n} \frac{k}{m} p(m \mid k,t), \\ &= \frac{1}{C(k,t)} \sum_{m=k}^{n} \binom{m-1}{k-1} t^m. \end{aligned} \tag{8}$$

On the other hand, let $Q(k,t)$ be the success probability when we reject the current candidate and then select, if any, the first candidate that appears. $Q(k,t)$ can also be computed by conditioning on N. The success probability conditional on $N=m$ is known to be $(k/m)\sum_{j=k+1}^{m}(j-1)^{-1}$ for $m>k$. Thus we have

$$\begin{aligned} Q(k,t) &= \sum_{m=k+1}^{n} \left(\frac{k}{m} \sum_{j=k+1}^{m} \frac{1}{j-1} \right) p(m \mid k,t), \\ &= \frac{1}{C(k,t)} \sum_{m=k+1}^{n} \left(\sum_{j=k+1}^{m} \frac{1}{j-1} \right) \binom{m-1}{k-1} t^m. \end{aligned} \tag{9}$$

Now, for a given n (though implicit), let

$$G = \{(k,t) : P(k,t) \geq Q(k,t)\}. \tag{10}$$

G represents the set of states for which stopping immediately is at least as good as continuing for exactly one more transition and then stopping. The rule that stops the first time the process enters a state in G is called the OLA (one-stage look-ahead) stopping rule. It is well known that if G is *closed* in a sense that once $(k,t) \in G$, then $(k+j,s) \in G$ for $j \geq 1, s \leq t$, then the OLA stopping rule is optimal (see [12]). Reference [7] called this case *monotone case*. From (8) and (9), $P(k,t) \geq Q(k,t)$ is equivalent to

$$\sum_{m=k}^{n} \binom{m-1}{k-1} t^m \geq \sum_{m=k+1}^{n} \left(\sum_{j=k+1}^{m} \frac{1}{j-1} \right) \binom{m-1}{k-1} t^m. \tag{11}$$

Remark 1. Let n tend to infinity in (11) and then apply a well-known identity

$$\sum_{m=k}^{\infty} \binom{m-1}{k-1} t^m = \left(\frac{t}{1-t} \right)^k, \quad k \geq 1$$

and another identity shown by [15] and then [3]

$$\sum_{m=k+1}^{\infty}\left(\sum_{j=k+1}^{m}\frac{1}{j-1}\right)\binom{m-1}{k-1}(1-t)^k t^{m-k} = -\log(1-t), \quad k \geq 1.$$

Then (11) can be greatly simplified to

$$1 \geq -\log(1-t)$$

or equivalently

$$t \leq 1 - e^{-1}$$

implying that

$$G = \left\{(k,t) : t \leq 1 - e^{-1}, \text{ irrespective of } k\right\}.$$

Since G is closed, G gives an optimal stopping region. This is just the result Stewart obtained.

We now return to finite n. Let $b_{k,m} = \sum_{j=k+1}^{m}(j-1)^{-1}, 1 \leq k < m$ with $b_{k,k} = 0$ corresponding to the vacuous sum and define, for $k \leq m \leq n$ and $0 \leq t \leq 1$,

$$\phi_{k,m}(t) = (1 - b_{k,m})\binom{m-1}{k-1}t^{m-k}, \tag{12}$$

and also, for $1 \leq k \leq n$,

$$\Phi_{\mathrm{k}}(t) = \sum_{m=k}^{n}\phi_{k,m}(t). \tag{13}$$

Then the inequality (11) is written as

$$\Phi_{\mathrm{k}}(t) \geq 0. \tag{14}$$

The following result is concerned with the form of $G_k = \{t : \Phi_{\mathrm{k}}(t) \geq 0,\ 0 \leq t \leq 1\}$, i.e., the set of t which satisfies the inequality (14).

Lemma 1. *Let n be fixed. Then, for a given k, there exists a value $t_k^*(n) \in [0,1]$ such that*

$$G_k = \{t : 0 \leq t \leq t_k^*(n)\}, \quad 1 \leq k \leq n. \tag{15}$$

Proof. This proof is similar to that of Lemma 3 in [8]. Two cases are distinguished according to the value of k.

Case $1 : s_1(n) \leq k \leq n$. Since $1 - b_{k,m} \geq 0$ for $k \leq m \leq n$ from the definition of $s_1(n)$, $\Phi_k(t)$ is non-decreasing in t with $\Phi_k(0) = 1$, and so $\Phi_k(t) \geq 1,\ 0 \leq t \leq 1$. If we define $t_k^*(n) = 1$, G_k is written as (15).

Case $2 : 1 \leq k < s_1(n)$. It is noted that $b_{k,m}$ is increasing in m, so there exists an integer $c = c(k)$ such that $1 - b_{k,m} \geq 0,\ k \leq m < c$ and $1 - b_{k,m}$

$< 0,\ c \le m \le n$. Denote by $\Phi_k^{(r)}(t)$ be the rth derivative of $\Phi_k(t)$ $(\Phi_k^{(0)}(t) = \Phi_k(t))$. Then we have from (13)

$$\Phi_k^{(r)}(t) = \frac{1}{(k-1)!} \sum_{m=k+r}^{n} (1 - b_{k,m}) \frac{(m-1)!}{(m-k-r)!} t^{m-(k+r)}. \tag{16}$$

Observe that, from the definition of c, $\Phi_k^{(c-k)}(t) < 0,\ 0 \le t \le 1$, and $\Phi_k^{(r)}(0) \ge 0,\ r \le c-k-1$. We show that if we define $G_{k,r} = \{t : \Phi_k^{(r)}(t) \ge 0,\ 0 \le t \le 1\},\ 0 \le r < c-k$, then $G_{k,r}$ can be written as

$$G_{k,r} = \{t : 0 \le t \le v_r^*\} \tag{17}$$

for some value $v_r^* \in [0,1]$. Since $\Phi_k^{(c-k-1)}(0) \ge 0$ and $\Phi_k^{(c-k-1)}(t)$ is decreasing in t, we obviously have expression (17) for $r = c-k-1$ by defining $v_{c-k-1}^* (< 1)$ as a unique solution $t \in [0,1)$ to the equation $\Phi_k^{(c-k-1)}(t) = 0$ if $\Phi_k^{(c-k-1)}(1) < 0$, otherwise by defining $v_{c-k-1}^* = 1$. For $r = c-k-2$, two cases are considered depending on $v_{c-k-1}^* = 1$ or $v_{c-k-1}^* < 1$. If $v_{c-k-1}^* = 1$, then since $\Phi_k^{(c-k-2)}(0) \ge 0$ and $\Phi_k^{(c-k-2)}(t)$ is increasing in t, we immediately have the form (17) by defining $v_{c-k-2}^* = 1$. If $v_{c-k-1}^* < 1$, then $\Phi_k^{(c-k-2)}(t)$ achieves its maximum at $t = v_{c-k-1}^*$, because we have known $\Phi_k^{(c-k-1)}(v_{c-k-1}^*) = 0$ and $\Phi_k^{(c-k)}(v_{c-k-1}^*) < 0$. Since $\Phi_k^{(c-k-2)}(0) \ge 0$, then also $\Phi_k^{(c-k-2)}(v_{c-k-1}^*) \ge 0$. Thus we have expression (17) for $r = c-k-2$ by defining $v_{c-k-2}^* (< 1)$ as a unique solution $t \in [v_{c-k-1}^*, 1)$ to the equation $\Phi_k^{(c-k-2)}(t) = 0$ if $\Phi_k^{(c-k-2)}(1) < 0$, otherwise by defining $v_{c-k-2}^* = 1$. This argument is repeated to yield $G_{k,0} = \{t : 0 \le t \le v_0^*\}$ for some value $v_0^* \in [0,1]$. Since $G_{k,0} = G_k$, we establish (15) if we define $t_k^*(n) = v_0^*$.

Now that the set G_k is shown to have the form (15) to prove that the set G is closed, it suffices to show that $G_1 \subseteq G_2 \subseteq \cdots \subseteq G_n$, or equivalently, $t_1^*(n) \le t_2^*(n) \le \cdots \le t_n^*(n)$. We need the following lemma.

Lemma 2. *We have, for* $2 \le k \le n$,

$$\Phi_k(t) - \Phi_{k-1}(t) = t\Phi_k(t) - \{\phi_{k-1,n}(t) + t\phi_{k,n}(t)\}. \tag{18}$$

Proof. We have, from (13),

$$\Phi_k(t) - \Phi_{k-1}(t) = \sum_{m=k}^{n} \{\phi_{k,m}(t) - \phi_{k-1,m-1}(t)\} - \phi_{k-1,n}(t). \tag{19}$$

However, from (12),

$$\begin{aligned} &\phi_{k,m}(t) - \phi_{k-1,m-1}(t) \\ &\qquad = \left[\left\{\binom{m-1}{k-1} - \binom{m-2}{k-2}\right\} - \left\{\binom{m-1}{k-1} b_{k,m} \right. \right. \\ &\qquad\quad \left. \left. - \binom{m-2}{k-2} b_{k-1,m-1}\right\}\right] t^{m-k}. \end{aligned}$$

Applying to this the following easily verifiable identities (see p. 139 of [15] for the second identity)

$$\binom{m-1}{k-1} - \binom{m-2}{k-2} = \binom{m-2}{k-1}$$
$$\binom{m-1}{k-1} b_{k,m} - \binom{m-2}{k-2} b_{k-1,m-1} = \binom{m-2}{k-1} b_{k,m-1},$$

we have, for $k \leq m \leq n$,

$$\begin{aligned} \phi_{k,m}(t) - \phi_{k-1,m-1}(t) &= (1 - b_{k,m-1}) \binom{m-2}{k-1} t^{m-k}, \\ &= t\phi_{k,m-1}(t), \end{aligned} \tag{20}$$

where $\phi_{k,k-1}(t)$ is interpreted as 0. Substituting (20) into (19) yields

$$\begin{aligned} \Phi_k(t) - \Phi_{k-1}(t) &= t \sum_{m=k}^{n} \phi_{k,m-1}(t) - \phi_{k-1,n}(t), \\ &= t\{ \sum_{m=k}^{n} \phi_{k,m}(t) - \phi_{k,n}(t)\} - \phi_{k-1,n}(t), \\ &= t\Phi_k(t) - \{t\phi_{k,n}(t) + \phi_{k-1,n}(t)\}, \end{aligned}$$

which is the desired result.

Lemma 3. *The sequence* $\{t_k^*(n) : 1 \leq k \leq n\}$ *is non-decreasing in* k*, that is,*

$$t_1^*(n) \leq t_2^*(n) \leq \cdots \leq t_n^*(n).$$

Proof. We have already found $t_k^*(n) = 1$ for $s_1(n) \leq k$ (see Case 1 in the proof of Lemma 1), so we have only to show that $t_{k-1}^*(n) \leq t_k^*(n)$ for $k \leq s_1(n) - 1$. Since $\phi_{k-1,n}(t) \leq 0$ and $\phi_{k,n}(t) \leq 0$, we obtain, from (18),

$$\Phi_k(t) - \Phi_{k-1}(t) \geq t\Phi_k(t),$$

implying that $\Phi_k(t) \geq \Phi_{k-1}(t)$ for t such that $\Phi_k(t) \geq 0$. Thus, considering that $\Phi_k(0) = \Phi_{k-1}(0) = 1$ and G_k and G_{k-1} are given in the form of (15), we can immediately conclude that $t_{k-1}^*(n) \leq t_k^*(n)$.

From Lemma 1 and the continuity of $\Phi_k(t)$, $t_k^*(n)$ is computed, if $t_k^*(n) < 1$, as a unique root t of the equation $\Phi_k(t) = 0$, or equivalently a unique root t of (5) via (13). The followig lemma gives the minimum number k for which $t_k^*(n) = 1$.

Lemma 4.

$$t_k^*(n) = 1 \Longleftrightarrow s_3(n) \leq k \leq n.$$

Proof. From the preceding argument, $t_k^*(n) = 1$ corresponds to $\Phi_k(1) \geq 0$. Thus we can define

$$\tilde{s}_3(n) = \min\{k \geq 1 : \Phi_k(1) \geq 0\} \tag{21}$$

as the minimum number k for which $t_k^*(n) = 1$. We show $\tilde{s}_3(n) = s_3(n)$ by proving

$$\Phi_k(1) = \binom{n}{k}\left[1 - \sum_{j=k+1}^{n} \frac{1}{j}\right]. \tag{22}$$

We have, from (12) and (13),

$$\begin{aligned}
\Phi_k(1) &= \sum_{m=k}^{n} \phi_{k,m}(1), \\
&= \sum_{m=k}^{n} \left(1 - \sum_{j=k+1}^{m} \frac{1}{j-1}\right)\binom{m-1}{k-1}, \\
&= \sum_{m=k}^{n} \binom{m-1}{k-1} - \sum_{j=k}^{n-1} \frac{1}{j} \sum_{m=j+1}^{n} \binom{m-1}{k-1}.
\end{aligned}$$

Using the well-known identity

$$\sum_{m=j+1}^{n} \binom{m-1}{k-1} = \binom{n}{k} - \binom{j}{k}, \quad k \leq j+1 \leq n, \tag{23}$$

where $\binom{j}{k}$ is interpreted as 0 when $j = k-1$, the preceding gives

$$\begin{aligned}
\Phi_k(1) &= \binom{n}{k} - \sum_{j=k}^{n-1} \frac{1}{j}\left[\binom{n}{k} - \binom{j}{k}\right], \\
&= \binom{n}{k}\left[1 - \sum_{j=k}^{n-1} \frac{1}{j}\right] + \sum_{j=k}^{n-1} \frac{1}{j}\binom{j}{k}.
\end{aligned} \tag{24}$$

However,

$$\begin{aligned}
\sum_{j=k}^{n-1} \frac{1}{j}\binom{j}{k} &= \frac{1}{k}\sum_{j=k}^{n-1} \binom{j-1}{k-1}, \\
&= \frac{1}{k}\binom{n-1}{k}, \qquad \text{(again from (23))} \\
&= \binom{n}{k}\left(\frac{1}{k} - \frac{1}{n}\right).
\end{aligned} \tag{25}$$

Applying (25) to (24) yields

$$\Phi_k(1) = \binom{n}{k}\left[1 - \sum_{j=k}^{n-1}\frac{1}{j} + \frac{1}{k} - \frac{1}{n}\right],$$

and hence proves (22).

The followng result shows that the difference between $s_1(n)$ and $s_3(n)$ is at most one.

Lemma 5. *For $n \geq 2$,*

$$s_3(n) \leq s_1(n) \leq s_3(n) + 1.$$

Proof. To show the first inequality $s_3(n) \leq s_1(n)$, it is sufficient to show that

$$b_{s_1(n)+1,n+1} \leq 1, \tag{26}$$

because $b_{s+1,n+1} \leq 1$ for $s \geq s_3(n)$ from the definition of $s_3(n)$. The inequality (26) is verified because we have $b_{s_1(n)+1,n+1} < b_{s_1(n),n}$ and also $b_{s_1(n),n} \leq 1$ from the definition of $s_1(n)$. To show the second inequality $s_1(n) \leq s_3(n)+1$, it is sufficient to show that

$$b_{s_3(n)+1,n} \leq 1, \tag{27}$$

because $b_{s,n} \leq 1$ for $s \geq s_1(n)$ from the definition of $s_1(n)$. The inequality (27) can be verified because we have $b_{s_3(n)+1,n} < b_{s_3(n)+1,n+1}$ and also $b_{s_3(n)+1,n+1} \leq 1$ from the definition of $s_3(n)$.

The above lemmas can now be combined to yield Theorem 1. The following result shows that for each k, $t_k^*(n)$ converges to the same value $1-e^{-1}$, which is consistent with the e^{-1}-rule for the non-informative case.

Lemma 6. *For fixed $k \geq 1$,*

$$\lim_{n\to\infty} t_k^*(n) = 1 - e^{-1}.$$

Proof. Remember that (5) is equivalent to

$$\sum_{m=k}^{n}\binom{m-1}{k-1}t^m = \sum_{m=k+1}^{n}\left(\sum_{j=k+1}^{m}\frac{1}{j-1}\right)\binom{m-1}{k-1}t^m,$$

so $t_k^*(n)$ is also a solution to this equation. Thus the limiting value of $t_k^*(n)$ must satisfy the equation

$$\sum_{m=k}^{\infty}\binom{m-1}{k-1}t^m = \sum_{m=k+1}^{\infty}\left(\sum_{j=k+1}^{m}\frac{1}{j-1}\right)\binom{m-1}{k-1}t^m.$$

We have already seen in the Remark that this equation has a unique root $1-e^{-1}$ for any k. Thus the proof is complete.

3 Concluding Remark

In this chapter, we have derived the explicit expression of the optimal selection rule for the best-choice problem with N uniform on $\{1, 2, \ldots, n\}$ for a given n and have shown that our rule coincides with that of [15] asymptotically. In contrast to the above no-information random arrival time best-choice problem where the observations are the relative ranks of the applicants, the full-information analogue is the problem where the observations are the true values of N applicants $X_1, X_2, \ldots, X_N$, assumed to be i.i.d. random variables from a known continuous distribution taken without loss of generality to be the uniform distribution on the interval $[0, 1]$. To the best of our knowledge, this full-information version has been studied only when N is Poisson (see [13] and [1]. See also [14]). The case in which N is uniform or geometric remains unsolved.

Acknowledgment

We are grateful to the anonymous referees for their careful reading.

References

1. Bojdecki, T.: On optimal stopping of a sequence of independent random variables - probability maximizing approach. Stoch. Proc. Their Appl. 6, 153–163 (1978)
2. Bruss, F.T.: A unified approach to a class of best choice problems with an unknown number of objects. Ann. Probab. 12, 882–889 (1984)
3. Bruss, F.T.: On an optimal selection problem of Cowan and Zabczyk. J. Appl. Probab. 24, 918–928 (1987)
4. Bruss, F.T., Samuels, S.M.: A unified approach to a class of optimal selection problems with an unknown number of options. Ann. Probab. 15, 824–830 (1987)
5. Bruss, F.T., Samuels, S.M.: Conditions for quasi-stationarity of the Bayes rule in selection problems with an unknown number of rankable options. Ann. Probab. 18, 877–886 (1990)
6. Bruss, F.T., Rogers, L.C.G.: Pascal processes and their characterization. Stoch. Proc. Their Appl. 37, 331–338 (1991)
7. Chow, Y., Robbins, H., Siegmund, D.: The Theory of Optimal Stopping, Houghton Mifflin, Boston, MA (1971)
8. Cowan, R., Zabczyk, J.: An optimal selection problem associated with the poisson process. Theory Probab. Appl. 23, 584–592 (1978)
9. Gianini, J., Samuels, S.M.: The infinite secretary problem. Ann. Probab. 4, 418–432 (1976)
10. Gilbert, J.P., Mosteller, F.: Recognizing the maximum of a sequence. J. Am. Stat. Assoc. 61, 35–73 (1966)
11. Presman, E.L., Sonin, I.M.: The best choice problem for a random number of objects. Theory Probab. Appl. 17, 657–668 (1972)

12. Ross, S.M.: Introduction to Stochastic Dynamic Programming, Academic, Orland, CA (1983)
13. Sakaguchi, M.: Optimal stopping problems for randomly arriving offers. Math. Japon. 21, 201–217 (1976)
14. Sakaguchi, M., Tamaki, M.: Optimal stopping problems associated with a non-homogeneous Markov process. Math. Japon. 25, 681–696 (1980)
15. Stewart, T.J.: The secretary problem with an unknown number of options. Opns. Res. 29, 130–145 (1981)