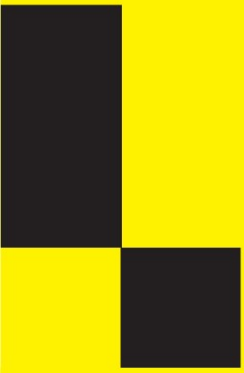
563 LECTURE NOTES IN ECONOMICS AND MATHEMATICAL SYSTEMS



Alberto Seeger

Recent Advances in Optimization

Springer

Lecture Notes in Economics and Mathematical Systems 563

Alberto Seeger (Ed.)

Recent Advances in Optimization

Editor

Prof. Alberto Seeger
University of Avignon
Department of Mathematics
33, rue Louis Pasteur
84000 Avignon, France
E-mail: alberto.seeger@univavignon.fr

ISSN 0075-8442
ISBN-10 3-540-28257-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-28257-0 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera ready by author
Cover design: *Erich Kirchner*, Heidelberg

Printed on acid-free paper 42/3130Jö 5 4 3 2 1 0

Preface

This volume contains the Proceedings of the Twelfth French-German-Spanish Conference on Optimization held at the University of Avignon in 2004. We refer to this conference by using the acronym FGS-2004.

During the period September 20-24, 2004, about 180 scientists from around the world met at Avignon (France) to discuss recent developments in optimization and related fields. The main topics discussed during this meeting were the following:

1. smooth and nonsmooth continuous optimization problems,
2. numerical methods for mathematical programming,
3. optimal control and calculus of variations,
4. differential inclusions and set-valued analysis,
5. stochastic optimization,
6. multicriteria optimization,
7. game theory and equilibrium concepts,
8. optimization models in finance and mathematical economics,
9. optimization techniques for industrial applications.

The Scientific Committee of the conference consisted of F. Bonnans (Rocquencourt, France), J.-B. Hiriart-Urruty (Toulouse, France), F. Jarre (Düsseldorf, Germany), M.A. Lopez (Alicante, Spain), J.E. Martinez-Legaz (Barcelona, Spain), H. Maurer (Münster, Germany), S. Pickenhain (Cottbus, Germany), A. Seeger (Avignon, France), and M. Thera (Limoges, France).

The conference FGS-2004 is the 12th of the series of French-German meetings which started in Oberwolfach in 1980 and was continued in Confolant (1981), Luminy (1984), Irsee (1986), Varetz (1988), Lambrecht (1991), Dijon (1994), Trier (1996), Namur (1998), Montpellier (2000), and Cottbus (2002).

Since 1998, this series of meetings has been organized under the participation of a third European country. In 2004, the guest country was Spain. The conference promoted, in particular, the contacts between researchers of the three

involved countries and provide a forum for sharing recent results in theory and applications of optimization.

The conference FGS-2004 was organized by the "Group of Nonlinear Analysis and Optimization" of the University of Avignon. As chairman of the Organizing Committee, I would like to acknowledge the following institutions for their financial or material support:

- Région Provence-Alpes-Cote d'Azur
- Université d'Avignon et des Pays de Vaucluse
- Agroparc: Technopole Régional d'Avignon
- Mairie d'Avignon
- Institut National de Recherche en Informatique et en Automatique

For the sake of convenience, the contributions appearing in this volume are splitted in four different groups:

Part I. Optimization Theory and Algorithms,
Part II. Optimal Control and Calculus of Variations,
Part III. Game Theory,
Part IV. Modeling and Numerical Testing.

Each contribution has been examined by one or two referees. The evaluation process has been more complete and thorough for the contributions appearing in Parts I, II, and III. The papers in Part IV are less demanding from a purely mathematical point-of-view (no theorems, propositions, etc). Their principal concern is either the modeling or the computer resolution of specific optimization problems arising in industry and applied sciences.

I would like to thank all the contributors for their effort and the anonymous referees for their comments and suggestions. The help provided by Mrs Monique Lefebvre (Secretarial Office of FGS-2004) and the staff of Springer-Verlag is also greatly appreciated.

Avignon, September 2005 *Alberto Seeger*

Contents

Part II Optimal Control and Calculus of Variations

Part III Game Theory

Part IV Industrial Applications and Numerical Testing

List of Contributors

Felipe Alvarez
Universidad de Chile/Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático
Casilla 170/3, Correo 3
Santiago, Chile
falvarez@dim.uchile.cl

António Sanches Antunes
University of Minho
Portugal
asanches@ipg.pt

M. Madalena Araujo
Minho University/Systems and Production Department
Braga, Portugal
mmaraujo@dps.uminho.pt

Rodica Branzei
Alexandru Ioan Cuza University/Faculty of Computer Science
Iasi, Romania
branzeir@info.uaic.ro

Hans Georg Bock
IWR Heidelberg
Heidelberg, Germany

Alexandre Cabot
Université de Limoges/Laboratoire LACO
Limoges, France
alexandre.cabot@unilim.fr

Francesc Carreras
Polytechnic University of Catalonia/Dep. of Applied Mathematics II and Industrial Engineering School
Terrassa, Spain
francesc.carreras@upc.edu

Stephan Dempe
Tech. University Bergakademie Freiberg/Dep. of Mathematics and Computer Sciences
Akademiestr. 6
09596 Freiberg, Germany
dempe@math.tu-freiberg.de

Moritz Diehl
IWR Heidelberg
Heidelberg, Germany

Oltin Dogaru
University Politehnica of Bucharest/Department of Mathematics
Splaiul Independenţei 313
060042 Bucharest, Romania

Isabel A.C.P. Espíritu-Santo
Minho University/Systems and Production Department
Braga, Portugal
iapinho@dps.uminho.pt

Edite M.G.P. Fernandes
Universidade do Minho Campus de Gualtar/Departamento de Producão e Sistemas, Escola de Engenharia
4710-057 Braga, Portugal
emgpf@dps.uminho.pt

José Fernández
University of Murcia/Dep. Statistics and Operations Research
Murcia, Spain

Massimiliano Ferrara
University of Messina/Faculty of Economics
Via dei Verdi 75
98122 Messina, Italy
mferrara@unime.it

Eugénio C. Ferreira
Minho University/Centre of Biological Engineering
Braga, Portugal
ecferreira@deb.uminho.pt

Steven Gabriel
University of Maryland, College Park/Department of Civil and Environmental Engineering
20742 Maryland, U.S.A.
sgabriel@umd.edu

Joseph Gergaud
ENSEEIHT-IRIT/UMR CNRS 5505
2, rue Camichel
31071 Toulouse, France
gergaud@enseeiht.fr

Ivan Ginchev
Technical University of Varna/Dep. of Mathematics
Studentska Str. 1
9010 Varna, Bulgaria
ginchev@ms3.tu-varna.acad.bg

Angelo Guerraggio
University of Insubria/Department of Economics
Via Ravasi 2
21100 Varese, Italy
aguerraggio@eco.uninsubria.it

Martin Gugat
Universität Erlangen-Nürnberg/Lehrstuhl 2 für Angewandte Mathematik
Martensstr. 3
91058 Erlangen, Germany
gugat@am.uni-erlangen.de

Elvira Hernández
UNED/ Depto. de Matemática Aplicada, E.T.S.I. Industriales
c/ Juan del Rosal 12
28040 Madrid, Spain
ehernandez@ind.uned.es

Alexey F. Izmailov
Moscow State University/ Dep. of Operations Research
Leninskiye Gori, GSP-2
119992 Moscow, Russia
izmaf@ccas.ru

Bienvenido Jiménez
Universidad de Salamanca/Depto. de Economía e Historia Económica, Facultad de Economía y Empresa
Campus Miguel de Unamuno, s/n,
37007 Salamanca, Spain
bjimen1@encina.pntic.mec.es

Alexander Kaplan
University of Trier/Department of Mathematics
54286 Trier, Germany
Al.Kaplan@tiscali.de

Igor V. Konnov
Kazan University/Department of Applied Mathematics
Kazan, Russia
ikonnov@ksu.ru

Valeriya Lykina
Brandenburg Univ. of Technology
Cottbus, Germany
lykina@math.tu-cottbus.de

Sebastian Lohse
Tech. University Bergakademie
Freiberg/Dep. of Mathematics and
Computer Sciences
Akademiestr. 6
09596 Freiberg, Germany

Alexander Mänz
ASPECTA Lebensversicherung AG
Germany
AMaenz@aspecta.com

Juan Enrique Martínez-Legaz
Universitat Autònoma de
Barcelona/Departament d'Economia
i d'Història Econòmica
08193 Bellaterra, Spain
juanenrique.martinez@uab.es

Pierre Martinon
ENSEEIHT-IRIT/UMR CNRS 5505
2, rue Camichel
31071 Toulouse, France
martinon@enseeiht.fr

Christian Meyer
Technische Universität
Berlin/Institut für Mathematik
Str. des 17. Juni 136
D-10623 Berlin, Germany
cmeyer@math.tu-berlin.de

M. Teresa T. Monteiro
University of Minho
Portugal
tm@dps.uminho.pt

Stefano Moretti
University of Genoa/Department of
Mathematics
Genoa, Italy
moretti@dima.unige.it

Henk Norde
Tilburg University/CentER and
Department of Econometrics and
Operations Research
Tilburg, The Netherlands
h.norde@uvt.nl

Vicente Novo
UNED/ Depto. de Matemática
Aplicada, E.T.S.I. Industriales
c/ Juan del Rosal 12
28040 Madrid, Spain
vnovo@ind.uned.es

Zsolt Páles
University of Debrecen/Institute of
Mathematics
4029 Debrecen, Pf. 12, Hungary
pales@math.klte.hu

Blas Pelegrín
University of Murcia/Dep. Statistics
and Operations Research
Murcia, Spain

Sabine Pickenhain
Brandenburg Univ. of Technology
Cottbus, Germany
sabine@math.tu-cottbus.de

Frank Plastria
Vrije Universiteit Brussel/MOSI-
Dep. of Mathematics, Operational
Research and Information Systems
for Management
Brussel, Belgium

María Albina Puente
Polytechnic University of Catalonia/
Dep. of Applied Mathematics III and
Polytechnic School
Manresa, Spain
m.albina.puente@upc.edu

Gerhard Reinelt
IWR Heidelberg
Heidelberg, Germany

Matteo Rocca
University of Insubria/Department of Economics
Via Ravasi 2
21100 Varese, Italy
mrocca@eco.uninsubria.it

Helena Sofia Rodrigues
University of Minho
Portugal
helena.rodrigues@ipb.pt

Sebastian Sager
IWR Heidelberg
Heidelberg, Germany
sebastian.sager
@iwr.uni-heidelberg.de

Johannes P. Schlöder
IWR Heidelberg
Heidelberg, Germany

Yves Smeers
Université catholique de Louvain/
Dep. of Mathematical Engineering and Center for Operations Research and Econometrics
Louvain-la-Neuve, Belgium
smeers@core.ucl.ac.be

Mikhail V. Solodov
Instituto de Matemática Pura e Aplicada
Estrada Dona Castorina 110, Jardim Botânico,
RJ 22460-320, Rio de Janeiro, Brazil.
solodov@impa.br

Ionel Ţevy
University Politehnica of Bucharest/
Department of Mathematics
Splaiul Independenţei 313
060042 Bucharest, Romania

Rainer Tichatschke
University of Trier/Department of Mathematics
54286 Trier, Germany
tichat@uni-trier.de

Stef Tijs
Tilburg University/CentER and Department of Econometrics and Operations Research
Tilburg, The Netherlands
S.h.tijs@uvt.nl

Boglárka Tóth
University of Murcia/Dep. Statistics and Operations Research
Murcia, Spain

Fredi Tröltzsch
Technische Universität
Berlin/Institut für Mathematik
Str. des 17. Juni 136
D-10623 Berlin, Germany
troeltz@math.tu-berlin.de

Constantin Udrişte
University Politehnica of Bucharest/
Department of Mathematics
Splaiul Independenţei 313
060042 Bucharest, Romania
udriste@mathem.pub.ro

A. Ismael F. Vaz
Universidade do Minho Campus de Gualtar/Departamento de Producão e Sistemas, Escola de Engenharia,
4710-057 Braga, Portugal
aivaz@dps.uminho.pt

Silvia Vogel
Technische Universitäet Ilmenau
Ilmenau, Germany
Silvia.Vogel@tu-ilmenau.de

Marcus Wagner
Cottbus University of Technology/Dep. of Mathematics
Karl-Marx-Str. 17, P.O. Box 101344
D-03013 Cottbus, Germany
wagner@math.tu-cottbus.de

Alexander J. Zaslavski
Technion, Dep. of Mathematics
Haifa, Israel
ajzasl@techunix.technion.ac.il

Part I

Optimization Theory and Algorithms

On the Asymptotic Behavior of a System of Steepest Descent Equations Coupled by a Vanishing Mutual Repulsion*

F. Alvarez[1]** and A. Cabot[2]

[1] Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático, Universidad de Chile, Casilla 170/3, Correo 3, Santiago, Chile.
falvarez@dim.uchile.cl

[2] Laboratoire LACO, Université de Limoges, Limoges, France.
alexandre.cabot@unilim.fr

Summary. We investigate the behavior at infinity of a special dissipative system, which consists of two steepest descent equations coupled by a non-autonomous conservative repulsion. The repulsion term is parametrized in time by an asymptotically vanishing factor. We show that under a simple slow parametrization assumption the limit points, if any, must satisfy an optimality condition involving the repulsion potential. Under some additional restrictive conditions, requiring in particular the equilibrium set to be one-dimensional, we obtain an asymptotic convergence result. Finally, some open problems are listed.

1 Introduction

Throughout this paper, H is a real Hilbert space with scalar product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. Let $\phi : H \to \mathbb{R}$ be a C^1 function and suppose that the set of critical points of ϕ is nonempty, that is,

$$S := \{x \in H \mid \nabla\phi(x) = 0\} \neq \emptyset.$$

A standard first-order method for finding a point in S consists in following the "Steepest Descent" trajectories:

$$\text{(SD)} \quad \dot{x} + \nabla\phi(x) = 0, \quad t \geq 0.$$

The evolution equation SD defines a dissipative dynamical system in the sense that every solution $x(t)$ satisfies $\frac{d}{dt}\phi(x(t)) = -\|\nabla\phi(x(t))\|^2$ so that $\phi(x(t))$

* This work was partially supported by the French-Chilean research cooperation program ECOS/CONICYT C04E03. The research was partly realized while the second author was visiting the first one at the CMM, Chile.

** The first author was supported by Fondecyt 1020610, Fondap en Matemáticas Aplicadas and Programa Iniciativa Científica Milenio.

decreases as long as $\nabla\phi(x(t)) \neq 0$. Since the stationary solutions of SD are described by S, it is natural to expect the corresponding solution $x(t)$ to approach the set S as $t \to \infty$. Indeed, under additional hypotheses, it is possible to ensure convergence at infinity to a local minimizer of ϕ (we refer the reader to [7, 8] for more details). However, we may be interested in additional information about S when ϕ has multiple critical points. For instance, we would like to compare some of them to select the best ones according to some additional criteria. We could also be interested in some properties of S such as unboundedness directions, symmetries, diameter estimates, etc. A possible strategy may be to "explore" the state space by solving a system of simultaneous SD equations. In order to reinforce the exploration aspect, and motivated by the second-order in time system treated in [11], we propose to introduce a perturbation term which models an asymptotically vanishing repulsion. More precisely, in this paper we study the following non-autonomous coupled system:

$$\text{(SDVR)} \qquad \begin{cases} \dot{x} + \nabla\phi(x) + \varepsilon(t)\nabla V(x-y) = 0, \\ \dot{y} + \nabla\phi(y) - \varepsilon(t)\nabla V(x-y) = 0. \end{cases}$$

Here the function $V : H \to \mathbb{R}$ is of class $\mathcal{C}^1$ and satisfies the repulsion condition

$$\forall x \in H \setminus \{0\}, \quad \langle \nabla V(x), x \rangle < 0, \tag{1}$$

while the parametrization map $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ tends to zero as $t \to \infty$:

$$\lim_{t\to\infty} \varepsilon(t) = 0. \tag{2}$$

This evolution problem will be referred to as the "Steepest Descent and Vanishing Repulsion" (SDVR) system.

As a simple illustration of the type of behavior that SDVR may exhibit, suppose $H = \mathbb{R}^n$ and consider the case of a quadratic objective function $\phi(x) = \frac{1}{2}\langle Ax, x\rangle$ with $A \in \mathbb{R}^{n\times n}$ being symmetric and positive semi-definite, together with the quadratic repulsion potential $V(x) = -\frac{1}{2}\|x\|^2$. The corresponding SDVR system is

$$\begin{cases} \dot{x} + Ax - \varepsilon(t)(x-y) = 0, \\ \dot{y} + Ay + \varepsilon(t)(x-y) = 0, \end{cases}$$

whose solution is explicitly given by

$$\begin{cases} x(t) = e^{-tA}\left[\frac{x_0+y_0}{2} + \frac{x_0-y_0}{2} e^{2\int_0^t \varepsilon(\tau)d\tau}\right], \\ y(t) = e^{-tA}\left[\frac{x_0+y_0}{2} - \frac{x_0-y_0}{2} e^{2\int_0^t \varepsilon(\tau)d\tau}\right]. \end{cases}$$

As $\varepsilon(t)$ vanishes when $t \to \infty$, if A is positive definite then $\lim_{t\to\infty} x(t) = \lim_{t\to\infty} y(t) = 0$, independently of the improper integral $\int_0^\infty \varepsilon(t)dt$. Suppose now that $\ker A \neq \{0\}$ and take $v \in \ker A \setminus \{0\}$. Remark that v

is a direction of unboundedness for $S = \ker A$. Since $e^{-tA}v = v$, we get $\langle x(t) - y(t), v\rangle = e^{2\int_0^t \varepsilon(\tau)d\tau}\langle x_0 - y_0, v\rangle$. When $\langle x_0 - y_0, v\rangle \neq 0$, the asymptotic behavior along the direction given by v depends strongly on the improper integral $\int_0^\infty \varepsilon(t)dt$. In that case, if $\int_0^\infty \varepsilon(\tau)d\tau = \infty$ then the repulsion forces $x(t)$ and $y(t)$ to diverge towards infinity following opposite directions. Notice that in this example $\inf V = -\infty$ and $\|\nabla V(x)\| \to \infty$ as $\|x\| \to \infty$. However, an analogous divergent behavior can occur for a repulsion potential V that is bounded from below and satisfies $\|\nabla V(x)\| \to 0$ as $\|x\| \to \infty$. For instance, take $H = \mathbb{R}$ and $\phi \equiv 0$ (so that $S = \mathbb{R}$), and suppose that $V \in \mathcal{C}^1(\mathbb{R})$ is such that $V(x) = |x|^{-1}$ for all $|x| \geq 1$. If $y_0 \leq -1$ and $x_0 \geq 1$ then the system we have to solve is given by

$$\begin{cases} \dot{x} - \varepsilon(t)/(x-y)^2 = 0, \\ \dot{y} + \varepsilon(t)/(x-y)^2 = 0. \end{cases}$$

Since $\frac{d}{dt}(x-y) = 2\varepsilon(t)/(x-y)^2$, we have that $x(t) - y(t) = ((x_0 - y_0)^3 + 6\int_0^t \varepsilon(\tau)d\tau)^{1/3}$, which diverges if and only if $\int_0^\infty \varepsilon(t)dt = \infty$.

From these examples we infer that the repulsion term $\pm\varepsilon(t)\nabla V(x-y)$ is asymptotically effective as soon as $\varepsilon(t)$ vanishes sufficiently slow as $t \to \infty$, and moreover, it is apparent that the adequate condition is

$$\int_0^\infty \varepsilon(t)\, dt = \infty. \tag{3}$$

Such a "slow parametrization" condition has already been pointed out by many authors in various contexts (cf. [3, 4, 9, 11]). Since $\varepsilon(t)$ vanishes when $t \to \infty$, it is quite easy to prove the convergence of the gradients $\nabla\phi(x)$ and $\nabla\phi(y)$ toward 0. The examples above show that under unboundedness of S we may observe divergence to infinity. Divergence can be prevented under coercivity of ϕ and the natural question that arises is the convergence of the trajectory $(x(t), y(t))$ as $t \to \infty$. This is a difficult problem due to the non convexity of the repulsive potential V (see [10] for positive results in a convex framework). In this direction, a one-dimensional convergence result has been obtained in [11] for a second-order in time version of SDVR.

The paper is organized as follows. In section 2, we state some general convergence properties for the SDVR system and we show that the slow parametrization assumption (3) forces the limit points to satisfy an optimality condition involving ∇V and the normal cone of S. This normal condition[3] is new and allows to reformulate some results of [11] in a more elegant way. In section 3 we derive a sharp convergence result when the equilibrium set S is one-dimensional. In the last section, we precise our results when ϕ is the square of a distance function. Due to the first-order (in time) structure of SDVR, our asymptotic selection results are sharper than in [11].

[3] This optimality condition has been found independently by M.-O. Czarnecki (University Montpellier II).

Notations. We use the standard notations of convex analysis. In particular, given a convex set $C \subset H$, we denote by $d_C(x)$ (resp. $P_C(x)$) the distance of the point $x \in H$ to the set C (resp. the best approximation to x from C). For every $x \in C$, the set $N_C(x)$ stands for the normal cone of C at x. Given any set $D \subset H$, the closed convex hull of D is denoted by $\overline{\text{co}}(D)$. Given $a, b \in H$, we define $[a,b] = \{a+\lambda(b-a) \mid \lambda \in [0,1]\}$ and $]a,b[= \{a+\lambda(b-a) \mid \lambda \in]0,1[\}$.

2 General Asymptotic Results

From now on, suppose that the functions $\phi : H \to \mathbb{R}$, $V : H \to \mathbb{R}$ and $\epsilon : \mathbb{R}_+ \to \mathbb{R}_+$, which are assumed to be of class $\mathcal{C}^1$, satisfy the following set of hypotheses $(\mathcal{H})$:

$$(\mathcal{H}_1) \begin{cases} i - \phi \text{ and } V \text{ are bounded from below on } H, \text{ with } \inf V = 0. \\ ii - \nabla\phi \text{ and } \nabla V \text{ are Lipschitz continuous on bounded sets of } H. \end{cases}$$

$$(\mathcal{H}_2) \begin{cases} i - \text{ The map } \varepsilon \text{ is non-increasing, i.e. } \dot{\varepsilon}(t) \leq 0 \quad \forall t \in \mathbb{R}_+. \\ ii - \text{ The map } \varepsilon \text{ is Lipschitz continuous on } \mathbb{R}_+. \\ iii - \lim_{t\to\infty} \varepsilon(t) = 0. \end{cases}$$

Let us begin our study of SDVR by noticing that it can be rewritten as a single vectorial equation in $H^2 = H \times H$. Indeed, let us set $X = (x,y) \in H^2$, $\Phi(X) = \phi(x) + \phi(y)$ and $U(X) = V(x-y)$. With such notations, SDVR is equivalent to

$$\dot{X} + \nabla\Phi(X) + \varepsilon(t)\nabla U(X) = 0, \tag{4}$$

where Φ and U are differentiable functions on H^2 satisfying the analogue to $(\mathcal{H}_1)$, that is

$$(\mathcal{H}_1^{vec}) \begin{cases} i - \Phi \text{ and } U \text{ are bounded from below on } H^2, \text{ with } \inf U = 0. \\ ii - \nabla\Phi \text{ and } \nabla U \text{ are Lipschitz continuous on bounded sets of } H^2. \end{cases}$$

Set

$$E(t) = \Phi(X(t)) + \varepsilon(t)U(X(t)) = \phi(x(t)) + \phi(y(t)) + \varepsilon(t)V(x(t)-y(t)).$$

By differentiating E with respect to time, we obtain

$$\dot{E} = -\|\dot{X}\|^2 + \dot{\varepsilon}\, U(X) = -\|\dot{x}\|^2 - \|\dot{y}\|^2 + \dot{\varepsilon}\, V(x-y) \leq 0.$$

Thus E is non-increasing, defining a Lyapounov-like function for (4). This is a useful tool in the study of the asymptotic stability of equilibria. Lyapounov methods and other powerful tools (like the Lasalle invariance principle) have been developed to study such a question. We refer the reader to the abundant literature on this subject; see, for instance, [2, 13, 14]. In this specific case, some standard arguments relying on the non-increasing and bounded from below function $E(t)$ permit to prove the next result, which we state without proof.

Proposition 1. *Assume that* $(\mathcal{H}_1^{vec})$ *and* $(\mathcal{H}_2)$ *hold. Then,*

(i) $\forall\, X_0 \in H^2$, *there exists a unique solution* $X : \mathbb{R}_+ \to H^2$ *of* (4), *which is of class* C^1 *and satisfies* $X(0) = X_0$. *Moreover,* $\dot{X} \in L^2([0,\infty); H^2)$.

(ii) Assuming additionally that $\{X(t)\}_{t\geq 0}$ *is bounded in* H^2 *(which is the case for example if* Φ *is coercive, i.e.* $\lim_{\|X\|\to\infty} \Phi(X) = \infty$*), then* $\lim_{t\to\infty} \dot{X}(t) = 0$ *and* $\lim_{t\to\infty} \nabla\Phi(X(t)) = 0$.

(iii) If Φ *is convex and* $\{X(t)\}_{t\geq 0}$ *is bounded then* $\lim_{t\to\infty} \Phi(X(t)) = \inf\Phi$.

The natural question that arises is the convergence of the trajectory $X(t)$ as $t \to \infty$. When $\varepsilon \equiv 0$, (4) reduces to the steepest descent dynamical system associated with Φ. In that case, there are different conditions ensuring the asymptotic convergence towards an equilibrium. For instance, it is well-known that under convexity of Φ, the trajectories weakly converge to a minimum of Φ (cf. Bruck [8]). This last result can be generalized when ε tends to zero fast enough; indeed, we have

Proposition 2. *In addition to* $(\mathcal{H}_1^{vec})$ *and* $(\mathcal{H}_2)$, *assume that* Φ *is convex with* $\operatorname{argmin}\Phi \neq \emptyset$. *If* $\int_0^\infty \varepsilon(t)\,dt < \infty$ *then every solution* $X(t)$ *of* (4) *weakly converges to a minimum of* Φ *as* $t \to \infty$.

We omit the proof of this result because it is similar to that given in [1] for second-order in time systems, which has been revisited with slight variants in [4, 5, 9, 11]. Notice that under the conditions of Proposition 2, any minimizer of Φ is asymptotically attainable. As the following result shows, that is not the case when the parametrization $\varepsilon(t)$ satisfies (3).

Lemma 1. *Assume that* $(\mathcal{H}_1^{vec})$, $(\mathcal{H}_2)$ *and* (3) *hold. Let* $X(t)$ *be a solution to* (4) *and suppose that* $X(t) \to X_\infty$ *strongly as* $t \to \infty$. *Then,*
(i) (Convex case)[4] *If* Φ *is convex, then* $X_\infty \in \operatorname{argmin}\Phi$ *and*

$$-\nabla U(X_\infty) \in N_{\operatorname{argmin}\Phi}(X_\infty). \tag{5}$$

(ii) (General case) We have $X_\infty \in C := \{X \in H^2 \mid \nabla\Phi(X) = 0\}$ *and*

$$-\nabla U(X_\infty) \in \bigcap_{W\in\mathcal{N}(X_\infty)} \overline{\operatorname{co}}\left(\mathbb{R}_+\nabla\Phi(W)\right), \tag{6}$$

where $\mathcal{N}(X_\infty)$ *denotes the set of neighborhoods of* X_∞ *and the set* $\mathbb{R}_+\nabla\Phi(W)$ *is defined by* $\mathbb{R}_+\nabla\Phi(W) := \{\lambda\,\nabla\Phi(x) \mid \lambda \in \mathbb{R}_+,\, x \in W\}$.

Proof. (i) From Proposition 1(*ii*), $\lim_{t\to\infty} \nabla\Phi(X(t)) = 0$ and hence $X_\infty \in \operatorname{argmin}\Phi$. Let $w \in \operatorname{argmin}\Phi$ so that $\nabla\Phi(w) = 0$. By convexity, $\nabla\Phi$ is monotone and we have

$$\forall v \in H^2, \quad \langle \nabla\Phi(v), v - w\rangle \geq 0. \tag{7}$$

[4] This result has been obtained simultaneously by M.-O. Czarnecki (University Montpellier II).

Taking the scalar product of (4) by $X(\cdot) - w$ and integrating on $[0,t]$, we obtain

$$\tfrac{1}{2}\|X(t)-w\|^2-\tfrac{1}{2}\|X(0)-w\|^2+\int_0^t \langle \nabla\Phi(X(s))+\varepsilon(s)\nabla U(X(s)), X(s)-w\rangle\, ds = 0.$$

Using (7), we get

$$\int_0^t \varepsilon(s)\langle \nabla U(X(s)), X(s)-w\rangle\, ds \leq \tfrac{1}{2}\|X(0)-w\|^2 - \tfrac{1}{2}\|X(t)-w\|^2.$$

Recalling that $\int_0^\infty \varepsilon(t)\, dt = \infty$, we deduce that

$$\langle \nabla U(X_\infty), X_\infty - w\rangle = \lim_{t\to\infty} \langle \nabla U(X(t)), X(t)-w\rangle \leq 0,$$

otherwise, we would have $\lim_{t\to\infty} \int_0^t \varepsilon(s)\langle \nabla U(X(s)), X(s)-w\rangle\, ds = \infty$, which is impossible. This being true for any $w \in \operatorname{argmin} \Phi$, we conclude that (5) holds.

(ii) Again, $X_\infty \in C$ due to Proposition 1 (*ii*). Next, let $W \in \mathcal{N}(X_\infty)$ and $v \in H^2$. Suppose that for every $w \in W$, $\langle \nabla\Phi(w), v\rangle \leq 0$. Since $X(t) \to X_\infty$, there exists $t_0 \geq 0$ such that for all $t \geq t_0$, $X(t) \in W$, and consequently

$$\forall t \geq t_0,\ \langle \nabla\Phi(X(t)), v\rangle \leq 0. \tag{8}$$

Integrating (4) on $[t_0,t]$ we obtain

$$\int_{t_0}^t \varepsilon(s)\langle \nabla U(X(s)), v\rangle\, ds = \langle X(t_0) - X(t), v\rangle - \int_{t_0}^t \langle \nabla\Phi(X(s)), v\rangle\, ds.$$

From (8), we get $\int_{t_0}^t \varepsilon(s)\langle \nabla U(X(s)), v\rangle ds \geq \langle X(t_0) - X(t), v\rangle$, $\forall t \geq t_0$. By (3), we deduce that

$$\langle \nabla U(X_\infty), v\rangle = \lim_{t\to\infty} \langle \nabla U(X(t)), v\rangle \geq 0.$$

This proves that, for every $v \in H^2$ and $w \in W$, if $\langle \nabla\Phi(w), v\rangle \leq 0$ then $\langle \nabla U(X_\infty), v\rangle \geq 0$, which amounts to

$$\forall v \in (\mathbb{R}_+\nabla\Phi(W))^o,\ \langle \nabla U(X_\infty), v\rangle \geq 0, \tag{9}$$

where $(\mathbb{R}_+\nabla\Phi(W))^o$ stands for the polar cone of the conic hull of $\nabla\Phi(W)$. By (9), the vector $-\nabla U(X_\infty)$ belongs to $(\mathbb{R}_+\nabla\Phi(W))^{oo}$, the polar cone of $(\mathbb{R}_+\nabla\Phi(W))^o$. Finally the bipolar theorem (cf., for example, [6]) ensures that $-\nabla U(X_\infty) \in \overline{\text{co}}\,(\mathbb{R}_+\nabla\Phi(W))$, which completes the proof. □

Remark 1. Condition (5) for the convex case expresses a necessary condition for X_∞ to be a local minimum of the function U on the set $\operatorname{argmin} \Phi$. In the general case, the set arising in (6) is closely related to the normal cone to C at X_∞. However, Lemma 1(i) cannot be viewed as a special case of Lemma 1(ii).

3 Convergence for a One-Dimensional Equilibrium Set

When ϕ has non-isolated critical points, the general results of the previous section for the vectorial form (4) of SDVR do not ensure the asymptotic convergence of the solution $(x(t), y(t))$ under the slow parametrization condition (3). If ϕ and V are both convex then it is possible to prove the asymptotic convergence to a pair (x_∞, y_∞) that minimizes $(x, y) \mapsto V(x - y)$ on $\operatorname{argmin} \phi \times \operatorname{argmin} \phi$ (see [10]). Although the repulsion condition (1) is not compatible with the convexity of V, the asymptotic selection principle given by Lemma 1 establishes that the "candidates" to be limit points must satisfy an analogous extremality condition depending on $U(x, y) = V(x - y)$. In a one-dimensional scalar setting, a convergence theorem for a second-order in time system involving a repulsion term has been proved in [11]. Next, we show that this type of result is valid for SDVR. To our best knowledge, convergence in the general higher dimensional case is an open problem.

From now on, we assume the following hypotheses on the function ϕ:

$$\text{for every bounded sequence } (x_n) \subset H,$$
$$\lim_{n\to\infty} \|\nabla\phi(x_n)\| = 0 \Rightarrow \lim_{n\to\infty} d_S(x_n) = 0, \tag{10}$$

$$\text{the map } \phi \text{ is coercive and } S = [a, b] \text{ for some } a, b \in H. \tag{11}$$

If $a \neq b$ then we suppose that for every $x \in H$,

$$\text{if } P_\Delta(x) \in S \text{ then } \nabla\phi(x) \text{ is orthogonal to } \Delta, \tag{12}$$

where Δ is the straight line $\Delta := \{a + \lambda(b - a) \mid \lambda \in \mathbb{R}\}$.

Remark 2. Condition (10) holds automatically when $\dim H < \infty$, but (11) and (12) are stringent. Take $\phi := f \circ d_{[a,b]}$ where $f \in \mathcal{C}^1(\mathbb{R}_+; \mathbb{R})$ and $d_{[a,b]}$ refers to the distance function to the segment $[a, b]$. If the function f is such that $f'(0) = 0$ and $f'(x) > 0$ for every $x > 0$, then the function ϕ satisfies (10), (11) and (12). Note that the function ϕ is a C^1 function due to the assumption $f'(0) = 0$.

On the repulsion potential V, we assume that there exists a scalar function $\gamma : H \to \mathbb{R}_{++}$ such that

$$\forall x \in H,\ \nabla V(x) = -\gamma(x)x. \tag{13}$$

Theorem 1. *Under hypotheses $(\mathcal{H})$, let $(x(t), y(t))$ be a solution to SDVR. If (10)-(13) hold, then:*

(i) *There exists $(x_\infty, y_\infty) \in [a, b]^2$ such that $\lim_{t\to\infty}(x(t), y(t)) = (x_\infty, y_\infty)$.*

(ii) *Suppose that $a \neq b$ and let us denote by Γ_a (resp. Γ_b) the connected component of $cl(\Delta \backslash S)$ such that $a \in \Gamma_a$ (resp. $b \in \Gamma_b$). Assume that $x_\infty = y_\infty = \ell$ and $P_\Delta(x(0)) \neq P_\Delta(y(0))$. Then ℓ equals a or b and*

- *$\ell = a$ implies $\Big(P_\Delta(x(t)), P_\Delta(y(t))\Big) \in \Gamma_a^2$ for every $t \geq 0$.*
- *$\ell = b$ implies $\Big(P_\Delta(x(t)), P_\Delta(y(t))\Big) \in \Gamma_b^2$ for every $t \geq 0$.*

(iii) *Suppose that the slow parametrization condition (3) holds. If* $P_\Delta(x(0)) \neq P_\Delta(y(0))$ *then* $(x_\infty, y_\infty) \in \{a,b\}^2$. *When in addition* $a \neq b$, *if* $x_\infty = y_\infty = a$ *(resp.* $x_\infty = y_\infty = b$*), then we have* $(P_\Delta(x(t)), P_\Delta(y(t))) \in \Gamma_a^2$ *(resp.* $(P_\Delta(x(t)), P_\Delta(y(t))) \in \Gamma_b^2$*) for every* $t \geq 0$.

Proof. (i) From the coercivity of ϕ, we deduce the boundedness of the map $t \mapsto (x(t), y(t))$ and hence in view of Proposition 1 (ii), we have $\lim_{t\to\infty} \nabla\phi(x(t)) = \lim_{t\to\infty} \nabla\phi(y(t)) = 0$. From assumption (10), it ensues that

$$\lim_{t\to\infty} d_S(x(t)) = \lim_{t\to\infty} d_S(y(t)) = 0. \tag{14}$$

If $a = b$ the set S is reduced to the singleton $\{a\}$ and the convergence of $x(t)$ and $y(t)$ toward a is immediate. Now assume that the segment line S is not trivial. Since $S \subset \Delta$, we have for every $x \in H$, $\|x - P_\Delta(x)\| = d_\Delta(x) \leq d_S(x)$. Hence, in view of (14), we obtain

$$\lim_{t\to\infty} \|x(t) - P_\Delta(x(t))\| = \lim_{t\to\infty} \|y(t) - P_\Delta(y(t))\| = 0.$$

As a consequence, the convergence of $x(t)$ (resp. $y(t)$) as $t \to \infty$ is equivalent to the convergence of $P_\Delta(x(t))$ (resp. $P_\Delta(y(t))$, which amounts to the convergence of $\langle x(t), b-a\rangle$ (resp. $\langle y(t), b-a\rangle$) as $t \to \infty$. For every $t \geq 0$, set $\alpha(t) := \langle x(t), b-a\rangle$ and $\beta(t) := \langle y(t), b-a\rangle$. From SDVR, we obtain

$$\dot\alpha(t) + \langle \nabla\phi(x(t)), b-a\rangle - \varepsilon(t)\,\gamma(x(t)-y(t))\,(\alpha(t)-\beta(t)) = 0. \tag{15}$$

$$\dot\beta(t) + \langle \nabla\phi(y(t)), b-a\rangle + \varepsilon(t)\,\gamma(x(t)-y(t))\,(\alpha(t)-\beta(t)) = 0. \tag{16}$$

We have that $\{\langle x, b-a\rangle \mid x \in S\} = [\lambda, \mu]$ for some $\lambda < \mu$. It is immediate to check that, for every $x \in H$, $\langle x, b-a\rangle \in [\lambda,\mu]$ is equivalent to $P_\Delta(x) \in S$, so that we can reformulate assumption (12) as

$$\langle x, b-a\rangle \in [\lambda,\mu] \Rightarrow \langle \nabla\phi(x), b-a\rangle = 0. \tag{17}$$

In particular, for every $t \geq 0$, we have that $\alpha(t) \in [\lambda,\mu]$ (resp. $\beta(t) \in [\lambda,\mu]$) implies $\langle \nabla\phi(x(t)), b-a\rangle = 0$ (resp. $\langle \nabla\phi(y(t)), b-a\rangle = 0$). Since the ω-limit sets of $\{x(t)\}_{t\geq 0}$ and $\{y(t)\}_{t\geq 0}$ are included in S, it is clear that:

$$\left[\liminf_{t\to\infty} \alpha(t), \limsup_{t\to\infty} \alpha(t)\right] \subset [\lambda,\mu] \quad \text{and} \quad \left[\liminf_{t\to\infty} \beta(t), \limsup_{t\to\infty} \beta(t)\right] \subset [\lambda,\mu].$$

We are now going to prove the convergence of $\alpha(t)$ and $\beta(t)$ as $t \to \infty$ by distinguishing three cases:

Case 1: For all $t \geq 0$, we have $\min\{\alpha(t), \beta(t)\} \geq \mu$ or $\max\{\alpha(t), \beta(t)\} \leq \lambda$. Without loss of generality, we can assume that for every $t \geq 0$, $\alpha(t) \geq \mu$ and $\beta(t) \geq \mu$. We deduce that $\liminf_{t\to\infty} \alpha(t) \geq \mu$ and $\liminf_{t\to\infty} \beta(t) \geq \mu$. Since $\limsup_{t\to\infty} \alpha(t) \leq \mu$ and $\limsup_{t\to\infty} \beta(t) \leq \mu$, we conclude that $\lim_{t\to\infty} \alpha(t) = \lim_{t\to\infty} \beta(t) = \mu$.

Case 2: There exist $c \in]\lambda, \mu[$ and $t_0 \geq 0$ such that either $\alpha(t_0) < c < \beta(t_0)$ or $\beta(t_0) < c < \alpha(t_0)$. Suppose $\alpha(t_0) < c < \beta(t_0)$. Let us first prove that

$$\forall t \geq t_0,\ \alpha(t) < c < \beta(t). \tag{18}$$

Let us set $t_\infty := \sup\{t \geq t_0,\ \forall u \in [t_0, t],\ \alpha(u) < c < \beta(u)\}$. Let us argue by contradiction and assume that $t_\infty < \infty$. We then have:

$$\forall t \in [t_0, t_\infty[\ \alpha(t) < c < \beta(t). \tag{19}$$

From the continuity of the maps $t \mapsto \alpha(t)$ and $t \mapsto \beta(t)$, we have $\alpha(t_\infty) = c$ or $\beta(t_\infty) = c$. Without any loss of generality, let us assume that $\alpha(t_\infty) = c$. Using again the continuity of the map α, there exists $t_1 \in [t_0, t_\infty]$ such that $\forall t \in [t_1, t_\infty]$, $\alpha(t) \geq \lambda$. Let us now use the differential equation (15) satisfied by α. Since $\alpha(t) \in [\lambda, c]$ for every $t \in [t_1, t_\infty]$, we deduce from (17) that $\langle \nabla\phi(x(t)), b - a\rangle = 0$. On the other hand, the sign of $\alpha - \beta$ is negative on $[t_1, t_\infty]$, so that equation (15) yields $\forall t \in [t_1, t_\infty]$, $\dot{\alpha}(t) \leq 0$. As a consequence, we have $c = \alpha(t_\infty) \leq \alpha(t_1)$, which contradicts (19). Therefore, we conclude that $t_\infty = +\infty$, which ends the proof of (18).

Case 2.a: First assume that $\alpha(t) \geq \lambda$ for every $t \geq t_0$. From (17) and the fact that $\alpha(t) \in [\lambda, c]$, we deduce that $\langle \nabla\phi(x(t)), b - a\rangle = 0$. This combined with (15) and the negative sign of $\alpha(t) - \beta(t)$ implies that $\dot{\alpha}(t) \leq 0$ for every $t \geq t_0$. As a consequence, $\lim_{t\to\infty} \alpha(t)$ exists.

Case 2.b: Now assume that there exists $t_1 \geq t_0$ such that $\alpha(t_1) < \lambda$. Let us first prove that

$$\forall t \geq t_1, \quad \alpha(t) \leq \lambda. \tag{20}$$

Let us argue by contradiction and assume that there exists $t_2 \geq t_1$ such that $\alpha(t_2) > \lambda$. Let

$$t_3 := \inf\{t \in [t_1, t_2],\ \forall u \in [t, t_2],\ \alpha(u) \geq \lambda\}.$$

From the continuity of α, we have $\alpha(t_3) = \lambda$. The definition of t_3 shows that $\alpha(t) \geq \lambda$ for every $t \in [t_3, t_2]$. In particular, we have $\alpha(t) \in [\lambda, c]$, which in view of (17) implies that $\langle \nabla\phi(x(t)), b - a\rangle = 0$. This combined with (15) and the negative sign of $\alpha(t) - \beta(t)$ yields $\dot{\alpha}(t) \leq 0$ for every $t \in [t_3, t_2]$. Hence, we infer that $\lambda < \alpha(t_2) \leq \alpha(t_3) = \lambda$, a contradiction which ends the proof of (20). From (20), we deduce that $\limsup_{t\to\infty} \alpha(t) \leq \lambda$. Since on the other hand, $\liminf_{t\to\infty} \alpha(t) \geq \lambda$, we conclude that $\lim_{t\to\infty} \alpha(t) = \lambda$.

The proof of the convergence of $\beta(t)$ follows the same lines and is left to the reader.

Case 3: There exist $c \in]\lambda, \mu[$ and $t_0 \geq 0$ such that $\alpha(t_0) = \beta(t_0) = c$. It is clear that the constant map $t \in [t_0, \infty[\mapsto (c, c)$ satisfies the differential equations (15) and (16). From the uniqueness of the Cauchy problem at t_0, we deduce that $\alpha(t) = \beta(t) = c$ for every $t \geq t_0$.

We let the reader check that all cases are recovered by the previous three ones.

(ii) If case 2 holds, it is immediate that $\lim_{t\to\infty}\alpha(t) \neq \lim_{t\to\infty}\beta(t)$, thus implying that $\lim_{t\to\infty}x(t) \neq \lim_{t\to\infty}y(t)$. If case 3 occurs, we obtain by reversing the time that $\alpha(0)=\beta(0)$ and hence $P_\Delta(x(0))=P_\Delta(y(0))$. Therefore, if $\lim_{t\to\infty}x(t)=\lim_{t\to\infty}y(t)=\ell$ and $P_\Delta(x(0)) \neq P_\Delta(y(0))$, case 1 necessary holds which means that

$$\forall t\geq 0,\ (P_\Delta(x(t)),P_\Delta(y(t)))\in \Gamma_a^2 \ \text{ or } \ \forall t\geq 0,\ (P_\Delta(x(t)),P_\Delta(y(t)))\in \Gamma_b^2.$$

In the first eventuality, we have $\ell = a$, while in the second one we obtain $\ell = b$.

(iii) First assume that $x_\infty = y_\infty$. From (ii), we deduce that x_∞ and y_∞ are extremal points of $S=[a,b]$. Now assume that $x_\infty \neq y_\infty$. Let us apply Lemma 1 (ii) by taking into account the fact that $\nabla V(x) = -\gamma(x)\,x$ and $\gamma(x)>0$ for every $x\in H$. Condition (6) yields

$$x_\infty - y_\infty \in \bigcap_{W_1\in\mathcal{N}(x_\infty)} \overline{\text{co}}(\mathbb{R}_+\nabla\phi(W_1)) \ \text{ and } \ y_\infty - x_\infty \in \bigcap_{W_2\in\mathcal{N}(y_\infty)} \overline{\text{co}}(\mathbb{R}_+\nabla\phi(W_2)).$$

Let us argue by contradiction and assume that $x_\infty \in]a,b[$ (resp. $y_\infty\in]a,b[$). It is then clear that

$$\bigcap_{W_1\in\mathcal{N}(x_\infty)} \overline{\text{co}}\,(\mathbb{R}_+\nabla\phi(W_1)) \subset \Delta_0^\perp \quad \text{and} \quad \bigcap_{W_2\in\mathcal{N}(y_\infty)} \overline{\text{co}}\,(\mathbb{R}_+\nabla\phi(W_2)) \subset \Delta_0^\perp,$$

where $\Delta_0 := \Delta-\Delta = \mathbb{R}\,(b-a)$. Therefore $x_\infty - y_\infty \in \Delta_0^\perp$. Since $x_\infty - y_\infty \in \Delta_0$, we conclude that $x_\infty = y_\infty$, a contradiction. The rest of the statement is an immediate consequence of (ii). □

4 Further Convergence Results

Under the assumption of slow parametrization, Theorem 1 shows that, either the solutions x and y of SDVR converge to the opposite extremities of S, or they have the same limit. Since our aim is a global exploration of S, the second case clearly appears as the pathological one. Our purpose in this section is to find sufficient conditions on ϕ and V ensuring the convergence toward the opposite extremities of ϕ. We will restrict the analysis to the functions of the form $\phi := d_S^2$.

Lemma 2. *Under the hypotheses of Theorem 1, take* $\phi(x)=\frac{\delta}{2}\|x-p\|^2$ *for some* $\delta\in\mathbb{R}_+$ *and* $p\in H$. *Suppose moreover that the map* γ *in* (13) *satifies* $\liminf_{x\to 0}\gamma(x)>0$. *If* (3) *holds then for every straight line* L *going through the point* p *and satisfying* $P_L(x(0))\neq P_L(y(0))$, *there exists* $T\geq 0$ *such that* $p\in]P_L(x(t)),P_L(y(t))[$ *for all* $t\geq T$.

Proof. Set $x_0=x(0)$ and $y_0=y(0)$. Let us denote by u a director vector of L. The assumption $P_L(x_0)\neq P_L(y_0)$ amounts to saying that $\langle x_0,u\rangle \neq \langle y_0,u\rangle$.

Without any loss of generality, one can assume that $\langle x_0, u\rangle > \langle y_0, u\rangle$. Taking into account the particular form of ϕ and V and adding (resp. subtracting) the first and second equation of SDVR, we find respectively

$$\dot{x}(t) + \dot{y}(t) + \delta(x(t) + y(t) - 2p) = 0$$

$$\dot{x}(t) - \dot{y}(t) + \delta(x(t) - y(t)) - 2\varepsilon(t)\,\gamma\,(x(t) - y(t))\,(x(t) - y(t)) = 0.$$

Taking the scalar product of these equations by the vector u and setting $\alpha(t) := \langle x(t), u\rangle$ (resp. $\beta(t) := \langle y(t), u\rangle$), we obtain:

$$\dot{\alpha}(t) + \dot{\beta}(t) + \delta(\alpha(t) + \beta(t) - 2\langle p, u\rangle) = 0 \tag{21}$$

$$\dot{\alpha}(t) - \dot{\beta}(t) + \delta(\alpha(t) - \beta(t)) - 2\varepsilon(t)\,\gamma\,(x(t) - y(t))\,(\alpha(t) - \beta(t)) = 0. \tag{22}$$

It is clear in view of equation (22) that if the quantity $\alpha(t) - \beta(t)$ takes the value 0 for some $t \geq 0$ then $\alpha(t) - \beta(t) = 0$ for every $t \geq 0$. Since by assumption $\alpha(0) - \beta(0) > 0$, we deduce that $\alpha(t) - \beta(t) > 0$ for every $t \geq 0$. From the assumption $\liminf_{x\to 0} \gamma(x) > 0$, there exist $\eta > 0$ and $m > 0$ such that, for every $\|x\| \leq \eta$, we have $\gamma(x) \geq m$. Since ϕ admits p as a unique strong minimum, we clearly have $\lim_{t\to\infty} x(t) = \lim_{t\to\infty} y(t) = p$ and hence $\lim_{t\to\infty} x(t) - y(t) = 0$. We deduce the existence of $t_0 \geq 0$ such that, for every $t \geq t_0$, we have $\gamma(x(t) - y(t)) \geq m$. This last inequality combined with (22) gives

$$\dot{\alpha}(t) - \dot{\beta}(t) + \delta(\alpha(t) - \beta(t)) \geq 2m\,\varepsilon(t)(\alpha(t) - \beta(t)).$$

Multiplying by $e^{\delta t}$, we obtain

$$\frac{d}{dt}\left[e^{\delta t}(\alpha(t) - \beta(t))\right] \geq 2m\,\varepsilon(t)\,e^{\delta t}(\alpha(t) - \beta(t)).$$

By integrating this differential equation between t_0 and t, we find:

$$\alpha(t) - \beta(t) \geq (\alpha(t_0) - \beta(t_0))\,e^{-\delta(t-t_0)} \exp \int_{t_0}^{t} 2m\,\varepsilon(s)\,ds. \tag{23}$$

On the other hand, a simple integration of (21) on $[t_0, t]$ yields

$$\alpha(t) + \beta(t) - 2\langle p, u\rangle = (\alpha(t_0) + \beta(t_0) - 2\langle p, u\rangle)\,e^{-\delta(t-t_0)}. \tag{24}$$

Relations (23) and (24) imply that

$$\alpha(t) - \langle p, u\rangle \geq \frac{e^{-\delta(t-t_0)}}{2}\left(\alpha(t_0) + \beta(t_0) - 2\langle p, u\rangle + (\alpha(t_0) - \beta(t_0))e^{\int_{t_0}^{t} 2m\varepsilon(s)ds}\right)$$

$$\beta(t) - \langle p, u\rangle \leq \frac{e^{-\delta(t-t_0)}}{2}\left(\alpha(t_0) + \beta(t_0) - 2\langle p, u\rangle - (\alpha(t_0) - \beta(t_0))e^{\int_{t_0}^{t} 2m\varepsilon(s)ds}\right)$$

Since $\int_{t_0}^{\infty} \varepsilon(s)\,ds = \infty$, we obtain the existence of $T \geq t_0$ such that $\beta(t) < \langle p, u\rangle < \alpha(t)$ for every $t \geq T$. This means that $p \in]P_L(x(t)), P_L(y(t))[$ for every $t \geq T$. □

Remark 3. The assumption $\liminf_{x\to 0}\gamma(x) > 0$ means that the repulsion term $\nabla V(x)$ is not negligible with respect to x when $x \to 0$. Suppose that the function V is defined by $V(x) := \theta(\|x\|^2)$, where $\theta : \mathbb{R}_+ \to \mathbb{R}$ is a decreasing function of class $\mathcal{C}^1$. In this case, the condition $\liminf_{x\to 0}\gamma(x) > 0$ is equivalent to $\theta'(0) < 0$.

In the next theorem, we assume that the function ϕ equals $\phi := d^2_{[a,b]}$, for some $a, b \in H$. We show that the assumption $\liminf_{x\to 0}\gamma(x) > 0$ implies that the trajectories x and y converge to opposite extremities of the segment line $[a, b]$. In this case, the repulsion term is strong enough to push the trajectories x and y away from one another.

Theorem 2. *Consider a segment line $[a, b] \subset H$, included in some straight line Δ and let us define the function ϕ by $\phi := \frac{\delta}{2} d^2_{[a,b]}$ for some $\delta > 0$. Under the hypotheses of Theorem 1, we suppose moreover that the map γ in* (13) *satifies $\liminf_{x\to 0}\gamma(x) > 0$, and that the slow parametrization condition* (3) *holds. Let $(x, y) : \mathbb{R}_+ \to H^2$ be the unique trajectory of SDVR with initial conditions $(x_0, y_0) \in H^2$ satisfying $P_\Delta(x_0) \neq P_\Delta(y_0)$. Then we have*

$$\lim_{t\to\infty}(x(t), y(t)) = (a, b) \qquad or \qquad \lim_{t\to\infty}(x(t), y(t)) = (b, a).$$

Proof. When $a = b$, the function ϕ admits the real a as a unique strong minimum and we obviously have $\lim_{t\to\infty} x(t) = \lim_{t\to\infty} y(t) = a$. From now on, let us assume that $a \neq b$. In view of Remark 2, the function $\phi := \frac{\delta}{2} d^2_{[a,b]}$ satisfies hypotheses (10)-(12). Hence Theorem 1 applies and one of the following cases holds

(*i*) $\lim_{t\to\infty}(x(t), y(t)) = (a, b)$ or $\lim_{t\to\infty}(x(t), y(t)) = (b, a)$.
(*ii*) $\lim_{t\to\infty}(x(t), y(t)) = (a, a)$ and $\forall t \geq 0$, $(P_\Delta(x(t)), P_\Delta(y(t))) \in \Gamma_a^2$.
(*iii*) $\lim_{t\to\infty}(x(t), y(t)) = (b, b)$ and $\forall t \geq 0$, $(P_\Delta(x(t)), P_\Delta(y(t))) \in \Gamma_b^2$.

Let us argue by contradiction and assume that case (i) does not hold. Without any loss of generality, we can assume that case (*ii*) holds. On the half-space E_a defined by $E_a := \{x \in H, \quad P_\Delta(x) \in \Gamma_a\}$, the function ϕ coincides with the function $x \mapsto \frac{\delta}{2}\|x - a\|^2$. From Lemma 2 applied with $p = a$ and the straight line Δ, we obtain the existence of $t_0 \geq 0$ such that $a \in]P_\Delta(x(t)), P_\Delta(y(t))[$ for $t \geq t_0$. This shows that either $P_\Delta(x(t)) \notin \Gamma_a$ or $P_\Delta(y(t)) \notin \Gamma_a$, which gives a contradiction. □

When the assumption $\liminf_{x\to 0}\gamma(x) > 0$ does not hold, it is possible to choose initial conditions so as to force the corresponding trajectories to converge toward the same limit. The next proposition provides us with a counterexample in the case $H = \mathbb{R}$.

Proposition 3. *Take any function $\phi : \mathbb{R} \to \mathbb{R}$ satisfying $\phi(x) = x^2/2$ for every $x \in \mathbb{R}_+$. Assume that the functions $V : H \to \mathbb{R}$ and $\varepsilon : \mathbb{R}_+ \to \mathbb{R}_+$ satisfy $(\mathcal{H})$. Suppose that there exist $M > 0$ and $\delta > 1$ such that,*

$$\forall x \in \mathbb{R}, \quad |V'(x)| \leq M\,|x|^\delta.$$

Let $(x, y) : \mathbb{R}_+ \to \mathbb{R}^2$ be the unique trajectory of SDVR with initial conditions (x_0, y_0). Then there exist $r > 0$ and a function $\theta : [0, r[\to \mathbb{R}_+$ such that, for every $x_0 > 0$ and $y_0 > 0$ with $|y_0 - x_0| < r$,

$$\theta(|y_0 - x_0|) \leq x_0 + y_0 \quad \Longrightarrow \quad \forall t \geq 0, \quad x(t) \geq 0 \quad \text{and} \quad y(t) \geq 0. \tag{25}$$

For such initial conditions, we have $\lim_{t\to\infty} x(t) = \lim_{t\to\infty} y(t) = 0$.

Proof. Let us consider the function $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ defined by $\tilde{\phi}(x) = x^2/2$ for every $x \in \mathbb{R}$ and let $(\tilde{x}, \tilde{y})$ be the unique trajectory of SDVR associated with $\tilde{\phi}$. If $(\tilde{x}, \tilde{y})$ is proved to satisfy the property (25), then $(\tilde{x}, \tilde{y})$ is also the solution of SDVR associated with any function ϕ coinciding with $\tilde{\phi}$ on $\mathbb{R}_+$. As a consequence, without loss of generality, we can assume that $\phi = \tilde{\phi}$. The SDVR system then reduces to:

$$\text{(SDVR)} \qquad \begin{cases} \dot{x}(t) + x(t) + \varepsilon(t)V'(x-y)(t) = 0 \\ \dot{y}(t) + y(t) - \varepsilon(t)V'(x-y)(t) = 0. \end{cases}$$

By adding the first and the second equation of SDVR, we obtain $\dot{x}(t) + \dot{y}(t) + x(t) + y(t) = 0$, which immediately yields

$$x(t) + y(t) = (x_0 + y_0)\, e^{-t}. \tag{26}$$

Without any loss of generality, one may assume that $y_0 < x_0$. We then have $y(t) < x(t)$ for every $t \geq 0$. From the assumption on V, we have for every $t \geq 0$, $V'(x-y)(t) \geq -M\,(x(t) - y(t))^\delta$. Let us subtract the first and the second equation of SDVR by taking into account the previous inequality

$$\dot{x}(t) - \dot{y}(t) + x(t) - y(t) - 2M\varepsilon(t)(x(t) - y(t))^\delta \leq 0.$$

We now multiply by e^t and set $u(t) = e^t(x(t) - y(t))$ to obtain

$$\dot{u}(t) \leq 2M\varepsilon(t)e^t(x(t) - y(t))^\delta = 2M\varepsilon(t)e^{-(\delta-1)t}\, u^\delta(t).$$

Let us integrate the previous inequality on $[0, t]$ to find

$$-\frac{1}{\delta - 1}\left(\frac{1}{u^{\delta-1}(t)} - \frac{1}{u^{\delta-1}(0)}\right) \leq 2M \int_0^t \varepsilon(s)e^{-(\delta-1)s} ds.$$

Setting

$$C = 2M(\delta - 1) \int_0^\infty \varepsilon(s)e^{-(\delta-1)s}\, ds,$$

we deduce

$$\frac{1}{u^{\delta-1}(t)} \geq \frac{1}{u^{\delta-1}(0)} - C. \tag{27}$$

Setting $r = C^{-\frac{1}{\delta-1}}$, we observe that if $u(0) = x_0 - y_0 < r$ then the second member of (27) is positive. Inequality (27) is then equivalent to

$$u(t) \leq \left(\frac{1}{u^{\delta-1}(0)} - C\right)^{-\frac{1}{\delta-1}} = (x_0 - y_0)\,(1 - C\,(x_0 - y_0)^{\delta-1})^{-\frac{1}{\delta-1}}.$$

Defining the function $\theta : [0, r[\to \mathbb{R}_+$ by $\forall z \in [0, r[,\ \theta(z) = z\,(1 - C\,z^{\delta-1})^{-\frac{1}{\delta-1}}$, the previous inequality can be rewritten as

$$x(t) - y(t) \leq \theta(x_0 - y_0)\,e^{-t}. \tag{28}$$

Note that the previous inequality remains true when $x_0 = y_0$, in which case $x(t) = y(t)$ for every $t \geq 0$. By combining (26) and (28), we finally obtain $y(t) \geq \frac{e^{-t}}{2}(x_0 + y_0 - \theta(x_0 - y_0))$. It is then clear that $\theta(x_0 - y_0) \leq x_0 + y_0$ implies $y(t) \geq 0$ for every $t \geq 0$. Since $x(t) \geq y(t)$, we also have $x(t) \geq 0$ for every $t \geq 0$. □

5 Open Questions and Further Remarks

Below are listed some open questions and possible directions for future investigation. Assumptions of Theorem 1 are very stringent: the set S of equilibria of ϕ is one-dimensional and the level curves of ϕ are colinear to the direction of S. We conjecture that the result of Theorem 1 remains true without assumption (12). More generally, the extension of Theorem 1 to the case of multidimensional equilibrium sets is open. The proof technique that we use in the paper cannot be immediately extended to these situations.

From Theorem 1, the trajectories x and y of SDVR may possibly coincide at the limit when $t \to +\infty$, even if the function V modelizes a repulsive potential. To avoid this eventuality, a natural idea consists in introducing a "singular" potential V defined on $H \setminus \{0\}$ such that $\lim_{x \to 0} V(x) = +\infty$. This type of potential plays a central role in gravitational or electromagnetic theories. For example, when $V(x) = 1/\|x\|$ it corresponds to the electric potential between two particles having the same sign. For further details, we refer the reader to [12], where the author studies the dynamics of a pair of oscillators coupled by a singular potential.

Another extension consists in studying the system of $N \geq 3$ steepest descent equations coupled by a mutual repulsion. For large values of N, such a coupled system could help in finding a global description of the set of minima of ϕ and also estimates of its size.

For numerical purposes, it would be interesting to study a discretized version of SDVR by using a finite differencing scheme. These developments are out of the scope of this paper but certainly indicate a matter for future research.

References

1. F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM J. Control Optim., 38:1102-1119, 2000.
2. V. Arnold. Equations Différentielles Ordinaires. Editions de Moscou, 1974.
3. H. Attouch and R. Cominetti. A dynamical approach to convex minimization coupling approximation with the steepest descent method. J. Differential Equations 128(2):519-540, 1996.
4. H. Attouch and M.-O. Czarnecki. Asymptotic control and stabilization of nonlinear oscillators with non isolated equilibria. J. Differential Equations 179:278-310, 2002.
5. H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method. I The continuous dynamical system. Commun. Contemp. Math. 2 (1):1-34, 2000.
6. V. Barbu and T. Precupanu. Convexity and Optimization in Banach Spaces. 2nd ed., D. Reidel, Dordrecht, Boston, 1986.
7. H. Brézis. Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert. North-Holland Mathematics Studies, No. 5., North-Holland Publishing Co., Amsterdam-London, 1973.
8. R.E. Bruck. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. J. Funct. Anal. 18:15-26, 1975.
9. A. Cabot. Inertial gradient-like dynamical system controlled by a stabilizing term, J. Optim. Theory Appl. 120:275-303, 2004.
10. A. Cabot. The steepest descent dynamical system with control. Applications to constrained minimization. ESAIM Control Optim. Calc. Var., 10:243-258, 2004.
11. A. Cabot and M.-O. Czarnecki. Asymptotic control of pairs of oscillators coupled by a repulsion, with non isolated equilibria. SIAM J. Control Optim. 41(4):1254-1280, 2002.
12. M.-O. Czarnecki. Asymptotic control of pairs of oscillators coupled by a repulsion, with non isolated equilibria II: the singular case. SIAM J. Control Optim. 42(6):2145-2171, 2004
13. W. Hirsch and S. Smale. Differential Equations, Dynamical Systems and Linear Algebra, Academic Press, New York, 1974.
14. J.P. Lasalle and S. Lefschetz. Stability by Lyapounov's Direct Method with Applications. Academic Press, New York, 1961.

Inverse Linear Programming

S. Dempe[1] and S. Lohse[2]

[1] Technical University Bergakademie Freiberg, Department of Mathematics and Computer Sciences, Akademiestr. 6, 09596 Freiberg, Germany.
dempe@math.tu-freiberg.de

[2] Technical University Bergakademie Freiberg, Department of Mathematics and Computer Sciences, Akademiestr. 6, 09596 Freiberg, Germany

Summary. Let $\Psi(b,c)$ be the solution set mapping of a linear parametric optimization problem with parameters b in the right hand side and c in the objective function. Then, given a point x^0 we search for parameter values $\overline{b}$ and $\overline{c}$ as well as for an optimal solution $\overline{x} \in \Psi(\overline{b},\overline{c})$ such that $\|\overline{x} - x^0\|$ is minimal. This problem is formulated as a bilevel programming problem. Focus in the paper is on optimality conditions for this problem. We show that, under mild assumptions, these conditions can be checked in polynomial time.

1 Introduction

Let $\Psi(b,c) = \operatorname{argmax}\{c^\top x : Ax = b, x \geq 0\}$ denote the set of optimal solutions of a linear parametric optimization problem

$$\max\left\{c^\top x : Ax = b, x \geq 0\right\}\ , \tag{1}$$

where the parameters of the right hand side and in the objective function are elements of given sets

$$\mathcal{B} = \{b : Bb = \tilde{b}\}\ , \quad \mathcal{C} = \{c : Cc = \tilde{c}\}\ ,$$

respectively. Throughout this note, $A \in \mathbb{R}^{m\times n}$ is a matrix of full row rank m, $B \in \mathbb{R}^{p\times m}, C \in \mathbb{R}^{q\times n}, \tilde{b} \in \mathbb{R}^p$ and $\tilde{c} \in \mathbb{R}^q$. This data is fixed once and for all.

Let $x^0 \in \mathbb{R}^n$ also be fixed. Our task is to find values $\overline{b}$ and $\overline{c}$ for the parameters, such that $x^0 \in \Psi(\overline{b},\overline{c})$ or, if this is not possible, x^0 is at least close to $\Psi(\overline{b},\overline{c})$. Thus we consider the following bilevel programming problem

$$\min\left\{\|x - x^0\| :\ x \in \Psi(b,c),\ b \in \mathcal{B},\ c \in \mathcal{C}\right\}\ , \tag{2}$$

which has a convex objective function $x \in \mathbb{R}^n \mapsto f(x) := \|x - x^0\|$, but not necessarily a convex feasible region. We consider in this note an arbitrary

(semi)norm $\|\cdot\|$, not necessarily the Euclidean norm. In fact, we are specially thinking in a polyhedral norm like, for instance, the l_1-norm.

Bilevel programming problems have been intensively investigated, see the monographs [2, 3] and the annotated bibliography [4]. Inverse linear programming problems have been investigated in the paper [1], where it is shown that the inverse problem to e.g. a shortest path problem can again be formulated as a shortest path problem and there is no need to solve a bilevel programming problem. However, the main assumption in [1] that there exist parameter values $\overline{b} \in \mathcal{B}$ and $\overline{c} \in \mathcal{C}$ such that $x^0 \in \Psi(\overline{b}, \overline{c})$ seems to be rather restrictive. Hence, we will not use this assumption.

Throughout the paper the following system is supposed to be infeasible:

$$A^\top y = c, \quad Cc = \tilde{c}. \tag{3}$$

Otherwise every solution of

$$Ax = b, \quad x \geq 0, \quad Bb = \tilde{b},$$

would be feasible for (2), which means that (2) reduces to

$$\min \left\{ \|x - x^0\| : \; Ax = b, \; x \geq 0, \; Bb = \tilde{b} \right\},$$

which is a convex optimization problem.

2 Reformulation as an MPEC

First we transform (2) via the Karush-Kuhn-Tucker conditions into a mathematical program with equilibrium constraints (MPEC) [5] and we get

$$\begin{aligned} \|x - x^0\| &\longrightarrow \min_{x,b,c,y} \\ Ax &= b \\ x &\geq 0 \\ A^\top y &\geq c \\ x^\top (A^\top y - c) &= 0 \\ Bb &= \tilde{b} \\ Cc &= \tilde{c}\,. \end{aligned} \tag{4}$$

The next thing which should be clarified is the notion of a local optimal solution.

Definition 1. *A point $\overline{x}$ is a local optimal solution of problem (2) if there exists a neighborhood U of $\overline{x}$ such that $\|x - x^0\| \geq \|\overline{x} - x^0\|$ for all x, b, c with $b \in \mathcal{B}$, $c \in \mathcal{C}$ and $x \in U \cap \Psi(b, c)$.*

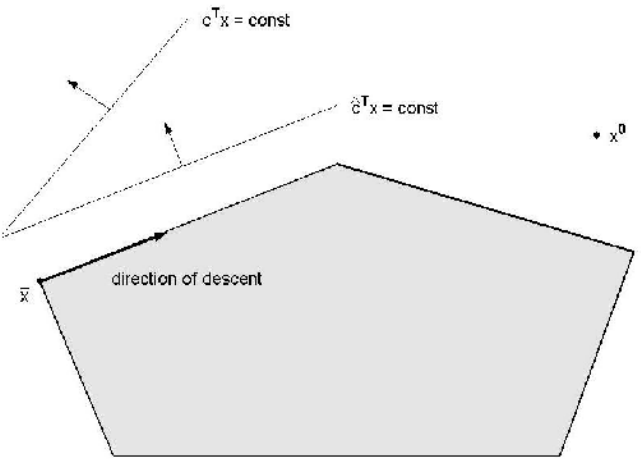


Fig. 1. Definition of a local optimal solution

Using the usual definition of a local optimal solution of problem (4) it can be easily seen that for each local optimal solution $\overline{x}$ of problem (2) there are $\overline{b}, \overline{c}, \overline{y}$ such that $(\overline{x}, \overline{b}, \overline{c}, \overline{y})$ is a local optimal solution of problem (4), cf. [3]. The opposite implication is in general not true.

Theorem 1. *Let* $\mathcal{B} = \{\overline{b}\}$, $\{\overline{x}\} = \Psi(\overline{b}, c)$ *for all* $c \in U \cap \mathcal{C}$, *where* U *is some neighborhood of* $\overline{c}$. *Then,* $(\overline{x}, \overline{b}, \overline{c}, \overline{y})$ *is a local optimal solution of* (4) *for some dual variables* $\overline{y}$.

The proof of Theorem 1 is fairly easy and therefore it is omitted. Figure 1 can be used to illustrate the fact of the last theorem. The points $\overline{x}$ satisfying the assumptions of Theorem 1 are the vertices of the feasible set of the lower level problem given by the dashed area in this figure.

3 Optimality via Tangent Cones

Now we consider a feasible point $\overline{x}$ of problem (2) and we want to decide whether $\overline{x}$ is local optimal or not. To formulate suitable optimality conditions certain subsets of the index set of active inequalities in the lower level problem need to be determined. Let

$$I(\overline{x}) = \{i : \overline{x}_i = 0\}$$

be the index set of active indices. Then every feasible solution x of (2) close enough to $\overline{x}$ satisfies $x_i > 0$ for all $i \notin I(\overline{x})$. Complementarity slackness motivates us to define the following index sets, too:

- $I(c,y) = \{i : (A^\top y - c)_i > 0\}$
- $\mathcal{I}(\overline{x}) = \{I(c,y) : A^\top y \geq c,\ (A^\top y - c)_i = 0\ \forall i \notin I(\overline{x}),\ Cc = \tilde{c}\}$
- $I^0(\overline{x}) = \bigcap_{I \in \mathcal{I}(\overline{x})} I.$

Remark 1. If an index set I belongs to the family $\mathcal{I}(\overline{x})$ then $I^0(\overline{x}) \subseteq I \subseteq I(\overline{x})$.

An efficient calculation of the index set $I^0(\overline{x})$ is necessary for the evaluation of the optimality conditions below. By contrast, the knowledge of the family $\mathcal{I}(\overline{x})$ itself is not necessary.

Remark 2. We have $j \in I(\overline{x}) \setminus I^0(\overline{x})$ if and only if the system

$$\begin{aligned} (A^\top y - c)_i &= 0 \quad \forall i \notin I(\overline{x}) \\ (A^\top y - c)_j &= 0 \\ (A^\top y - c)_i &\geq 0 \quad \forall i \in I(\overline{x}) \setminus \{j\} \\ Cc &= \tilde{c} \end{aligned}$$

is feasible. Furthermore $I^0(\overline{x})$ is an element of $\mathcal{I}(\overline{x})$ if and only if the system

$$\begin{aligned} (A^\top y - c)_i &= 0 \quad \forall i \notin I^0(\overline{x}) \\ (A^\top y - c)_i &\geq 0 \quad \forall i \in I^0(\overline{x}) \\ Cc &= \tilde{c} \end{aligned}$$

is feasible.

Now we are able to transform (4) into a locally equivalent problem, which does not explicitly depend on c and y.

Lemma 1. *$\overline{x}$ is a local optimal solution of (2) if and only if $\overline{x}$ is a (global) optimal solution of all problems (A_I)*

$$\begin{aligned} \|x - x^0\| &\longrightarrow \min_{x,b} \\ Ax &= b \\ x &\geq 0 \\ x_i &= 0 \quad \forall i \in I \\ Bb &= \tilde{b} \end{aligned} \tag{A_I}$$

with $I \in \mathcal{I}(\overline{x})$.

Proof. Let $\overline{x}$ be a local optimal solution of (2) and assume that there is a set $I \in \mathcal{I}(\overline{x})$ with $\overline{x}$ being not optimal for (A_I). Then there exists a sequence $\{x^k\}_{k \in \mathbb{N}}$ of feasible solutions of (A_I) with $\lim_{k \to \infty} x^k = \overline{x}$ and $\|x^k - x^0\| < \|\overline{x} - x^0\|$ for all k. Consequently $\overline{x}$ can not be a local optimal solution to (2)

since $I \in \mathcal{I}(\overline{x})$ implies that all x^k are also feasible for (2). Conversely, let $\overline{x}$ be an optimal solution of all problems (A_I) and assume that there is a sequence $\{x^k\}_{k\in\mathbb{N}}$ of feasible points of (2) with $\lim_{k\to\infty} x^k = \overline{x}$ and $\|x^k - x^0\| < \|\overline{x} - x^0\|$ for all k. For k sufficiently large the elements of this sequence satisfy the condition $x_i^k > 0$ for all $i \notin I(\overline{x})$ and due to the feasibility of x^k for (2) there are sets $I \in \mathcal{I}(\overline{x})$ such that x^k is feasible for problem (A_I). Because $\mathcal{I}(\overline{x})$ consists only of a finite number of sets, there is a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ where x^{k_j} are all feasible for a fixed problem (A_I). So we contradict the optimality of $\overline{x}$ for this problem (A_I). □

Corollary 1. *We can also consider*

$$\begin{aligned} \|x - x^0\| &\longrightarrow \min_{x,b,I} \\ Ax &= b \\ x &\geq 0 \\ x_i &= 0 \quad \forall i \in I \\ Bb &= \tilde{b} \\ I &\in \mathcal{I}(\overline{x}) \end{aligned} \tag{5}$$

to check if $\overline{x}$ is a local optimal solution of (2). Here the index set I is a minimization variable. Problem (5) combines all the problems (A_I) into one problem and means that we have to find a best one between all the optimal solutions of the problems (A_I) for $I \in \mathcal{I}(\overline{x})$.

In what follow we use the notation

$$T_I(\overline{x}) = \{d |\ \exists r:\ Ad = r,\ Br = 0,\ d_i \geq 0\ \ \forall i \in I(\overline{x}) \setminus I,\ d_i = 0\ \ \forall i \in I\}\ .$$

This set corresponds to the tangent cone (relative to x only) to the feasible set of problem (A_I) at the point $\overline{x}$. The last lemma obviously implies the following necessary and sufficient optimality condition.

Lemma 2. *$\overline{x}$ is a local optimal solution of (5) if and only if $f'(\overline{x}, d) \geq 0$ for all*

$$d \in T(\overline{x}) := \bigcup_{I\in\mathcal{I}(\overline{x})} T_I(\overline{x})\ .$$

Remark 3. $T(\overline{x})$ is the (not necessarily convex) tangent cone (relative x) of problem (5) at the point $\overline{x}$.

Corollary 2. *The condition $I^0(\overline{x}) \in \mathcal{I}(\overline{x})$ implies $T_{I^0(\overline{x})}(\overline{x}) = T(\overline{x})$.*

Remark 4. If f is differentiable at $\overline{x}$, then saying that $f'(\overline{x}, \cdot)$ is nonnegative over $T(\overline{x})$ is obviously equivalent to saying that

$$f'(\overline{x}, d) \geq 0 \quad \forall d \in \operatorname{conv} T(\overline{x})\ , \tag{6}$$

where the "conv" indicates the convex hull operator.

As shown in the next example, without differentiability assumption, (6) is sufficient for optimality but not necessary.

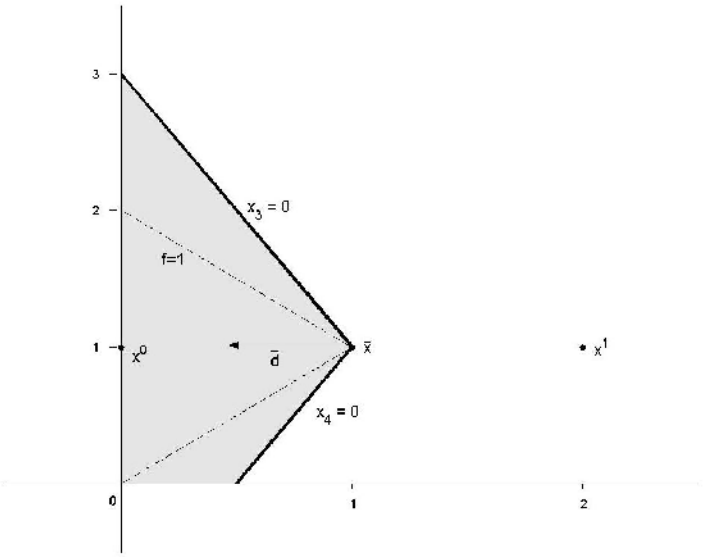


Fig. 2. Illustration of Example 1

Example 1. Let us consider a problem with the l_1-norm restricted to the first two components of x as objective function and

$$A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 2 & -1 & 0 & 1 \end{pmatrix}, \quad \mathcal{B} = \left\{ \begin{pmatrix} 3 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{C} = \left\{ 2e_1^{(4)} + te_2^{(4)} : \ t \in \mathbb{R} \right\},$$

$$x^0 = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 2 \end{pmatrix}, \quad x^1 = \begin{pmatrix} 2 \\ 1 \\ -2 \\ -2 \end{pmatrix} \quad \text{and} \quad \overline{x} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

We consider the point $\overline{x}$. The bold marked lines in Fig. 2 are the feasible set of our problem and the dashed lines are iso-distance-lines with the value 1. So we get the convexified tangent cone as

$$\operatorname{conv} T(\overline{x}) = \{d : \ 2d_1 + d_2 + d_3 = 0; \ 2d_1 - d_2 + d_4 = 0; \ d_3, d_4 \geq 0\} .$$

Finally $\overline{d} = (-1 \ 0 \ 2 \ 2)^\top \in \operatorname{conv} T(\overline{x})$ is a direction of descent with $f'(\overline{x}, \overline{d}) = -1$ although $\overline{x}$ is obviously the global optimal solution. If we choose x^1 (instead of x^0) and the objective function $|x_1 - x_1^1| + |x_2 - x_2^1|$, condition (6) implies the optimality of $\overline{x}$.

Remark 5. Because it is a matter of illustration, we considered the problem with inequality constraints in the lower level. For that reason we used the l_1-norm restricted to the first two components of x as objective function and not the l_1-norm over the whole space $\mathbb{R}^4$. By the way, in this case $\overline{x}$ would not be a local optimal solution.

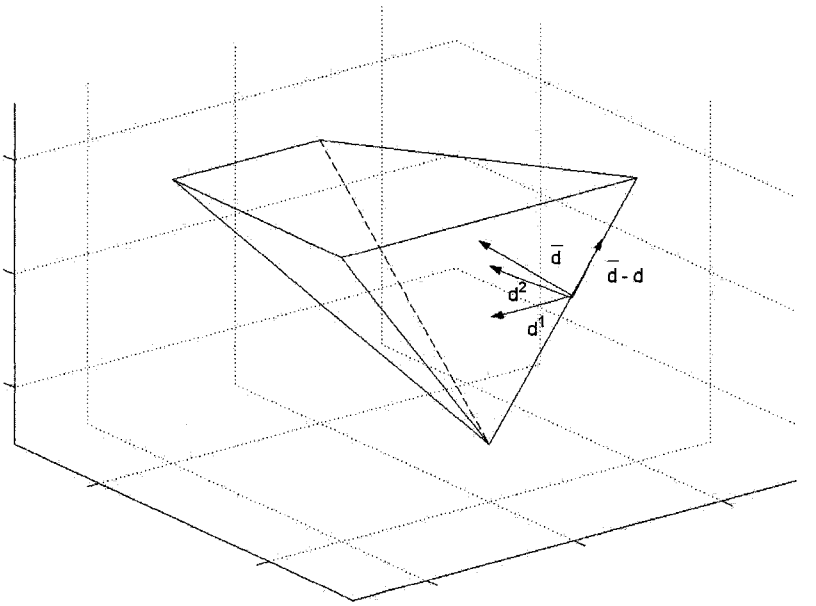


Fig. 3. Illustration of the proof of Theorem 2

4 A Formula for the Tangent Cone

For the verification of the optimality condition (6) an explicit formula for the tangent cone $\operatorname{conv} T(\overline{x})$ is essential. For notational simplicity we suppose $I(\overline{x}) = \{1, \ldots, k\}$ and $I^0(\overline{x}) = \{l+1, \ldots, k\}$ with $l \leq k \leq n$. Consequently all feasible points of (2) sufficiently close to $\overline{x}$ satisfy $x_i = 0$ for all $i \in I^0(\overline{x})$. We pay attention to this fact and consider the following relaxed problem:

$$\begin{aligned} \|x - x^0\| &\longrightarrow \min_{x,b} \\ Ax &= b \\ x_i &\geq 0 \quad i = 1, \ldots, l \\ x_i &= 0 \quad i = l+1, \ldots, k \\ Bb &= \tilde{b}\,. \end{aligned} \tag{7}$$

In what follow we use the notation

$$T_R(\overline{x}) = \{d |\ \exists r : Ad = r,\ Br = 0,\ d_i \geq 0\ i = 1, \ldots, l,\ d_i = 0\ \ i = l+1, \ldots, k\}.$$

This set corresponds to the tangent cone (relative x) of (7) at the point $\overline{x}$. Since $I^0 \subseteq I$ for all $I \in \mathcal{I}(\overline{x})$, it follows immediately that

$$\operatorname{conv} T(\overline{x}) = \operatorname{cone} T(\overline{x}) \subseteq T_R(\overline{x})\,. \tag{8}$$

The point $\overline{x}$ is said to satisfy the full rank condition, if

$$\operatorname{span}(\{A_i :\ i \notin I(\overline{x})\}) = \mathbb{R}^m, \tag{FRC}$$

where A_i denotes the ith column of the matrix A.

Example 2. All non-degenerate vertices of $Ax = b$, $x \geq 0$ satisfy (FRC).

This condition allows us now to establish equality between the cones above.

Theorem 2. *Let (FRC) be satisfied at the point $\overline{x}$. Then equality holds in* (8).

Proof. Let $\overline{d}$ be an arbitrary element of $T_R(\overline{x})$, that means there is a $\overline{r}$ with $A\overline{d} = \overline{r}$, $B\overline{r} = 0$, $\overline{d}_i \geq 0$ $i = 1, \ldots, l$, $\overline{d}_i = 0$ $i = l+1, \ldots, k$. We consider the following linear systems

$$\begin{aligned} Ad &= \delta_{1,j}\overline{r} \\ d_j &= \overline{d}_j \\ d_i &= 0 \quad i = 1, \ldots, k,\ i \neq j \end{aligned} \tag{S_j}$$

for $j = 1, \ldots, l$, where $\delta_{1,j} = 1$ if $j = 1$ and $\delta_{1,j} = 0$ if $j \neq 1$. These systems are all feasible because of (FRC). Furthermore let $d^1, \ldots, d^l$ be (arbitrary) solutions of the systems $(S_1), \ldots, (S_l)$ respectively. We define now the direction $d = \sum_{j=1}^{l} d^j$ and get $d_i = \overline{d}_i$ for $i = 1, \ldots, k$ as well as $Ad = A\overline{d} = \overline{r}$. Because we chose arbitrary vectors $d^1, \ldots, d^l$ it is possible that $d \neq \overline{d}$. But we can achieve equality with a translation of the solution d^1 by a specific vector of $\mathcal{N}(A) = \{z :\ Az = 0\}$. Therefore we define $\hat{d}^1 := d^1 + \overline{d} - d$, and because d^1 is feasible for (S_1) and $d_i = \overline{d}_i$ for $i = 1, \ldots, k$ as well as $Ad = A\overline{d} = \overline{r}$ we get $\hat{d}^1_i = 0$ for all $i = 2, \ldots, k$ and $A\hat{d}^1 = A(d^1 + \overline{d} - d) = \overline{r} + \overline{r} - \overline{r} = \overline{r}$. Hence

$\hat{d}^1$ is also a solution of (S_1). Thus we have $\hat{d}^1 + \sum_{j=2}^{l} d^j = \overline{d} - d + \sum_{j=1}^{l} d^j = \overline{d}$. As a result of the definition of the set $I^0(\overline{x})$ there are index sets $I_j \in \mathcal{I}(\overline{x})$ with $j \notin I_j$ for all $j \in \{1, \ldots, l\} = I(\overline{x}) \setminus I^0(\overline{x})$. So $\hat{d}^1$ is an element of the tangent cone of problem (A_{I_1}) and d^j are elements of the tangent cones of the problems (A_{I_j}) for $j = 2, \ldots, l$, see the definition of these cones. Finally $\overline{d}$ is the sum of a finite number of elements of $T(\overline{x})$ and therefore $T_R(\overline{x}) \subseteq \operatorname{cone} T(\overline{x})$. □

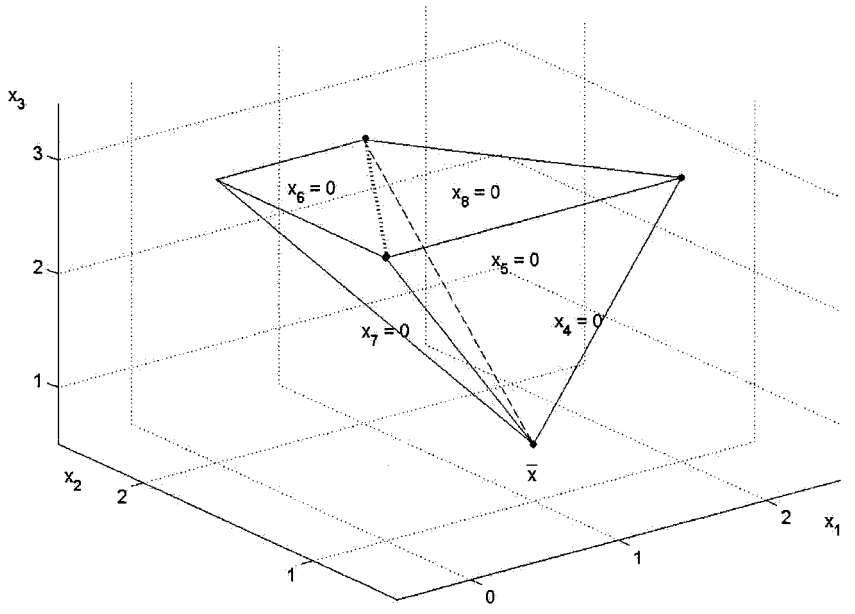


Fig. 4. Illustration of Example 3

By combining Lemma 2 and Remarks 2 and 4, one obtains:

Corollary 3. *Let $\overline{x}$ be a point of differentiability of f. Then, at most n systems of linear equalities\inequalities are needed to be investigated in order to compute the index set $I^0(\overline{x})$. Furthermore, verification of local optimality of a feasible point of problem (2) is possible in polynomial time.*

Example 3. This example will show that (FRC) is not necessary for equality in (8).

$$\begin{array}{rrrrrrrrr} & x_2 & & - x_4 & & & & = 1 \\ 2x_1 + & 2x_2 & - x_3 & & + x_5 & & & = 3 \\ & 2x_2 & - x_3 & & & + x_6 & & = 1 \\ 2x_1 & & + x_3 & & & & - x_7 & = 3 \\ & & x_3 & & & & + x_8 & = 3 \\ & & & & & & x_i & \geq 0 \end{array}$$

$\mathcal{B} = \{(1\ 3\ 1\ 3\ 3)^\top\}$ and $\mathcal{C} = \{c = -e_2^{(8)} + t(2e_1^{(8)} + 3e_2^{(8)} - e_3^{(8)}) + s(3e_2^{(8)} - e_3^{(8)}) : t, s \in \mathbb{R}\}$. Consider the point $\overline{x} = (1,1,1,0,0,0,0,2)^\top$. Hence we get $I(\overline{x}) = \{4,5,6,7\}$, $I^0 = \emptyset$ and $T_R(\overline{x}) = \{d : Ad = 0,\ d_i \geq 0 \quad \forall i \in I(\overline{x})\}$. The feasible region of (5) consists of the four faces $x_4 = 0$, $x_5 = 0$, $x_6 = 0$ and $x_7 = 0$ ($t = s = 0$; $t = 1, s = 0$; $t = 0, s = 1$ respectively $t = -\frac{1}{3}, s = \frac{2}{3}$). Obviously we have $T_R(\overline{x}) = \operatorname{cone} T(\overline{x})$. Now delete the second vector in $\mathcal{C}$, that means $\mathcal{C} = \{c = -e_2^{(8)} + t(2e_1^{(8)} + 3e_2^{(8)} - e_3^{(8)}) : t \in \mathbb{R}\}$. Then we also get $I^0 = \emptyset$. That is why the tangent cone of the relaxed problem is the same as above. But the convexified tangent cone $\operatorname{conv} T(\overline{x})$ of (5) is a proper subset of this cone. Because the feasible set consists only of the two faces $x_4 = 0$ and $x_5 = 0$, the cone $\operatorname{conv} T(\overline{x})$ is spanned by the four bold marked vertices where the apex of the cone is $\overline{x}$, see Fig. 4.

Acknowledgements. The authors sincerely thank the anonymous referee, whose comments led to an improvement of the note.

References

1. R.K. Ahuja and J.B. Orlin. Inverse optimization. Operations Research 49:771–783, 2001.
2. J.F. Bard. Practical Bilevel Optimization: Algorithms and Applications. Kluwer Academic Publishers, Dordrecht, 1998.
3. S. Dempe. Foundations of Bilevel Programming. Kluwer Academic Publishers, Dordrecht, 2002.
4. S. Dempe. Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. Optimization 52:333–359, 2003.
5. J. Outrata, M. Kočvara, and J. Zowe. Nonsmooth Approach to Optimization Problems with Equilibrium Constraints. Kluwer Academic Publishers, Dordrecht, 1998.

Second-Order Conditions in $C^{1,1}$ Vector Optimization with Inequality and Equality Constraints

Ivan Ginchev[1], Angelo Guerraggio[2], and Matteo Rocca[3]

[1] Technical University of Varna, Department of Mathematics, Studentska Str. 1
9010 Varna, Bulgaria. ginchev@ms3.tu-varna.acad.bg
[2] University of Insubria, Department of Economics, Via Ravasi 2
21100 Varese, Italy. aguerraggio@eco.uninsubria.it
[3] University of Insubria, Department of Economics, Via Ravasi 2
21100 Varese, Italy. mrocca@eco.uninsubria.it

Summary. The present paper studies the following constrained vector optimization problem: $\min_C f(x)$, $g(x) \in -K$, $h(x) = 0$, where $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^p$ are $C^{1,1}$ functions, $h : \mathbb{R}^n \to \mathbb{R}^q$ is C^2 function, and $C \subset \mathbb{R}^m$ and $K \subset \mathbb{R}^p$ are closed convex cones with nonempty interiors. Two type of solutions are important for the consideration, namely w-minimizers (weakly efficient points) and i-minimizers (isolated minimizers). In terms of the second-order Dini directional derivative second-order necessary conditions a point x^0 to be a w-minimizer and second-order sufficient conditions x^0 to be an i-minimizer of order two are formulated and proved. The effectiveness of the obtained conditions is shown on examples.

1 Introduction

In this paper we deal with the constrained vector optimization problem

$$\min_C f(x), \quad g(x) \in -K, \quad h(x) = 0, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^p$ and $h : \mathbb{R}^n \to \mathbb{R}^q$ are given functions, and $C \subset \mathbb{R}^m$ and $K \subset \mathbb{R}^p$ are closed convex cones with nonempty interiors. The inclusion $g(x) \in -K$ generalizes constraints of inequality type (in fact it is equivalent to $\langle \eta, g(x) \rangle \leq 0$, $\eta \in K'$). This remark explains why the word *inequality* appears in the title of the paper. In the case when f and g are $C^{1,1}$ functions and h is C^2 function we derive second-order optimality conditions for a point x^0 to be a solution of this problem. The paper is thought as a continuation of the investigation initiated by the authors in [8], [9] and [10], where either unconstrained problems or problems with only inequality constraints are studied. Recall that a function is said to be $C^{k,1}$ if it is k-times Fréchet differentiable with locally Lipschitz k-th derivative. The $C^{0,1}$ functions

are the locally Lipschitz functions. The $C^{1,1}$ functions have been introduced in Hiriart-Urruty, Strodiot, Hien Nguen [16] and since then have found various application in optimization. In particular second-order conditions for $C^{1,1}$ scalar problems are studied in [16, 6, 19, 28, 27]. Second-order optimality conditions in vector optimization are investigated in [1, 4, 18, 24, 26], and what concerns $C^{1,1}$ vector optimization in [12, 13, 21, 22, 23]. The given in the present paper approach and results generalize that of [23].

The assumption that f and g are defined on the whole space $\mathbb{R}^n$ is taken for convenience. Since we deal only with local solutions of problem (1), evidently our results generalize straightforward for functions f and g being defined on an open subset of $\mathbb{R}^n$. Usually the solutions of (1) are called points of efficiency. We prefer, like in the scalar optimization, to call them minimizers. In Section 2 we introduce different type of minimizers. Among them in our considerations an important role play the w-minimizers (weakly efficient points) and the i-minimizers (isolated minimizers). When we say first or second-order conditions we mean as usual conditions expressed in suitable first or second-order derivatives of the given functions. Here we deal with the Dini directional derivatives. In Section 2 we define the second-order Dini derivative. In Section 3 we recall after [10] second-order optimality conditions for problems with only inequality constraints. In Section 4 we prove second-order sufficient conditions for $C^{1,1}$ problems with both inequality and equality constraints. Section 5 indicates necessary optimality conditions. Section 6 points out directions for further investigations.

2 Preliminaries

For the norm and the dual parity in the considered finite-dimensional spaces we write $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$. From the context it should be clear to exactly what spaces these notations are applied.

For the cone $M \subset \mathbb{R}^k$ its positive polar cone is $M' = \{\zeta \in \mathbb{R}^k \mid \langle\zeta, \phi\rangle \geq 0 \text{ for all } \phi \in M\}$. The cone M' is closed and convex, and $M'' := (M')' = \operatorname{clco} M$, see Rockafellar [25, Theorem 14.1, page 121].

If $\phi \in \operatorname{clconv} M$ we set $M'[\phi] = \{\zeta \in M' \mid \langle\zeta, \phi\rangle = 0\}$. Then $M'[\phi]$ is a closed convex cone and $M'[\phi] \subset M'$. Consequently its positive polar cone $M[\phi] := (M'[\phi])'$ is a closed convex cone, $M \subset M[\phi]$ and $(M[\phi])' = M'[\phi]$. In this paper we apply the notation $M[\phi]$ for $M = K$ and $\phi = -g(x^0)$.

Given a set $A \subset \mathbb{R}^k$, then the distance from $y \in \mathbb{R}^k$ to A is $d(y, A) = \inf\{\|a - y\| \mid a \in A\}$. The oriented distance from y to A is defined by $D(y, A) = d(y, A) - d(y, \mathbb{R}^k \setminus A)$. The function D is introduced in Hiriart-Urruty [14, 15]. In the case of a convex set A, Ginchev, Hoffmann [11] show that $D(y, A) = \sup_{\|\xi\|=1}(\langle\xi, y\rangle - \sup_{a\in A}\langle\xi, a\rangle)$, which for $A = -C$ and C a closed convex cone gives $D(y, -C) = \sup\{\langle\xi, y\rangle \mid \xi \in C', \|\xi\| = 1\}$.

In terms of the distance function we have

$$K[-g(x^0)] = \{w \in \mathbb{R}^p \mid \limsup_{t \to 0^+} \frac{1}{t}\, d(-g(x^0) + tw, K) = 0\},$$

that is $K[-g(x^0)]$ is the contingent cone [3] of K at $-g(x^0)$.

We call the solutions of problem (1) minimizers. The solutions are understood in a local sense. In any case a solution is a feasible point x^0, that is a point satisfying the constraints $g(x^0) \in -K$, $h(x^0) = 0$.

The feasible point x^0 is said to be a w-minimizer (weakly efficient point) for (1) if there exists a neighbourhood U of x^0, such that $f(x) \notin f(x^0) - \operatorname{int} C$ for all feasible points $x \in U$. The feasible point x^0 is said to be an e-minimizer (efficient point) for (1) if there exists a neighbourhood U of x^0, such that $f(x) \notin f(x^0) - (C \setminus \{0\})$ for all feasible points $x \in U$. We say that the feasible point x^0 is a s-minimizer (strong minimizer) for (1) if there exists a neighbourhood U of x^0, such that $f(x) \notin f(x^0) - C$ for all feasible points $x \in U \setminus \{x^0\}$.

As in [8] it can be proved that the feasible point x^0 is a w-minimizer (s-minimizer) for the vector problem (1) if and only if x^0 is a minimizer (strong minimizer) for the scalar problem

$$\min D(f(x) - f(x^0), -C), \quad g(x) \in -K, \quad h(x) = 0.$$

This observation motivates the following definition. We say that the feasible point x^0 is an isolated minimizer (for short *i-minimizer*) of order k, $k > 0$, for (1) if there exists a neighbourhood U of x^0 and a constant $A > 0$ such that

$$D(f(x) - f(x^0), -C) \geq A \|x - x^0\|^k \quad \text{for all feasible} \quad x \in U. \tag{2}$$

Since any two norms in a finite dimensional real space are equivalent, the notion of an i-minimizer is norm-independent.

Obviously, each i-minimizer is a s-minimizer. Further each s-minimizer is an e-minimizer and each e-minimizer is a w-minimizer (under the assumption $C \neq \mathbb{R}^m$).

The concept of an isolated minimizer for scalar problems is introduced in Auslender [2]. For vector problems it has been extended in Ginchev [7], Ginchev, Guerraggio, Rocca [8], and under the name of strict minimizers in Jiménez [17] and Jiménez, Novo [18]. We prefer the name *isolated minimizer* given originally by A. Auslender.

In the sequel we establish optimality conditions for problem (1) in terms of the second-order Dini derivative (for short *Dini derivative*). For a given $C^{1,1}$ function $\Phi : \mathbb{R}^n \to \mathbb{R}^k$ we define the second-order Dini derivative $\Phi''_u(x^0)$ of Φ at x^0 in direction $u \in \mathbb{R}^n$ by

$$\Phi''_u(x^0) = \operatorname*{Limsup}_{t \to 0^+} \frac{2}{t^2} \left(\Phi(x^0 + tu) - \Phi(x^0) - t\,\Phi'(x^0)u\right).$$

If Φ is twice Fréchet differentiable at x^0 then the Dini derivative is a singleton and can be expressed in terms of the Hessian $\Phi''_u(x^0) = \Phi''(x^0)(u, u)$.

We deal often with the Dini derivative of the function $\Phi : \mathbb{R}^n \to \mathbb{R}^{m+p}$, $\Phi(x) = (f(x), g(x))$. Then we use the notation $\Phi''_u(x^0) = (f(x^0), g(x^0))''_u$. Let us turn attention that always $(f(x^0), g(x^0))''_u \subset f''_u(x^0) \times g''_u(x^0)$, but in general these two sets do not coincide. The following lemma gives some useful properties of the differential quotient.

Lemma 1 ([10]). *Let $\Phi : \mathbb{R}^n \to \mathbb{R}^k$ be a $C^{1,1}$ function and Φ' be Lipschitz with constant L on the ball $\{x \mid \|x - x^0\| \le r\}$, where $x^0 \in \mathbb{R}^n$ and $r > 0$. Then, for $u, v \in \mathbb{R}^m$ and $0 < t < r/\max(\|u\|, \|v\|)$ we have*

$$\| \frac{2}{t^2} \left(\Phi(x^0 + tv) - \Phi(x^0) - t\Phi'(x^0)v\right) - \frac{2}{t^2} \left(\Phi(x^0 + tu) - \Phi(x^0) - t\Phi'(x^0)u\right) \|$$

$$\le L \left(\|u\| + \|v\|\right) \|v - u\| .$$

In particular, for $v = 0$ we get

$$\| \frac{2}{t^2} \left(\Phi(x^0 + tu) - \Phi(x^0) - t\Phi'(x^0)u\right) \| \le L \|u\|^2.$$

3 Inequality Constraints, Sufficient Conditions

Here we consider the problem with only inequality constraints

$$\min_C f(x), \quad g(x) \in -K. \tag{3}$$

After [10] we recall a result establishing second-order sufficient optimality conditions. In the next section it will be applied to treat the problem with both equality and inequality constraints. We put

$$\Delta_I(x^0) = \left\{(\xi, \eta) \in C' \times K'[-g(x^0)] \setminus \{(0, 0)\} \mid \langle \xi, f'(x^0)\rangle + \langle \eta, g'(x^0)\rangle = 0\right\}$$

$$= \left\{(\xi, \eta) \in C' \times K' \mid (\xi, \eta) \ne 0,\ \langle \eta, g(x^0)\rangle = 0,\ \langle \xi, f'(x^0)\rangle + \langle \eta, g'(x^0)\rangle = 0\right\}$$

using the subscript I to underline that Δ_I is a set associated to the problem with only inequality constraints.

Theorem 1 ([10]). *Consider problem (3) with f and g being $C^{1,1}$ functions, and C and K closed convex cones with nonempty interiors. Let x^0 be a feasible point. Suppose that for each $u \in \mathbb{R}^n \setminus \{0\}$ one of the following two conditions is satisfied:*

$$\mathbb{S}'_i: \qquad (f'(x^0)u, g'(x^0)u) \notin -(C \times K[-g(x^0)]),$$

$$\mathbb{S}''_i: \quad (f'(x^0)u, g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus intC \times intK[-g(x^0)])$$
$$\text{and } \forall (y^0, z^0) \in (f(x^0), g(x^0))''_u : \exists (\xi^0, \eta^0) \in \Delta_I(x^0) :$$
$$\langle \xi^0, y^0\rangle + \langle \eta^0, z^0\rangle > 0.$$

Then x^0 is an i-minimizer of order two for problem (3).

Theorem 1 generalizes Theorem 4.2 from Liu, Neittaanmäki, Křížek [23] in the following aspects. Theorem 1 in opposite to [23] concerns arbitrary and not only polyhedral cones C and K. In Theorem 1 the conclusion is that x^0 is an i-minimizer of order two, while in [23] the weaker conclusion is proved that the reference point x^0 is only an e-minimizer.

4 Inequality and Equality Constraints, Sufficient Conditions

In Theorem 2 we establish sufficient conditions for the general problem (1) with both inequality and equality constraints. If the functions f, g, h are at least C^1, we put

$$\Delta(x^0) = \{(\xi,\eta,\zeta) \in C' \times K' \times \mathbb{R}^q \mid (\xi,\eta,\zeta) \neq (0,0,0),\ \langle \eta,\, g(x^0)\rangle = 0,$$

$$\langle \xi,\, f'(x^0)u\rangle + \langle \eta,\, g'(x^0)u\rangle + \langle \zeta,\, h'(x^0)u\rangle = 0 \text{ for } u \in \ker h'(x^0)\}\,.$$

Theorem 2. *Consider problem (1) with f, $g \in C^{1,1}$ and $h \in C^2$, and C and K closed convex cones with nonempty interiors. Let x^0 be a feasible point and let the vectors $h_1'(x^0), \ldots, h_q'(x^0)$, which are the components of $h'(x^0)$, be linearly independent. Let the vectors $\bar{u}^j \in \mathbb{R}^n$, $j = 1,\ldots,q$, be determined by*

$$h_k'(x^0)\bar{u}^j = 0 \quad \text{for} \quad k \neq j, \quad \text{and} \quad h_j'(x^0)\bar{u}^j = 1\,. \tag{4}$$

Suppose that for each $u \in \ker h'(x^0) \setminus \{0\}$ one of the following two conditions is satisfied.

$\mathbb{S}'$:
$$(f'(x^0)u,\, g'(x^0)u) \notin -(C \times K[-g(x^0)])\,,$$

$\mathbb{S}''$:
$$(f'(x^0)u,\, g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus intC \times intK[-g(x^0)])$$
$$\text{and } \forall\,(y^0,z^0) \in (f(x^0),g(x^0))''_u : \exists\,(\xi^0,\eta^0,\zeta^0) \in \Delta(x^0) :$$
$$\langle \xi^0,\, y^0\rangle + \langle \eta^0,\, z^0\rangle + \langle \zeta^0,\, h''(x^0)(u,u)\rangle > 0$$
$$\text{with } \zeta^0 = (\zeta_j^0)_{j=1}^q \text{ satisfying (5), where}$$

$$\zeta_j^0 = -\langle \xi^0,\, f'(x^0)\bar{u}^j\rangle - \langle \eta^0,\, g'(x^0)\bar{u}^j\rangle\,, \quad j = 1,\ldots,q\,. \tag{5}$$

Then x^0 is an i-minimizer of order two for problem (1).

Before going on with the proof we transform our problem (1) to a problem with only inequality constraints. Determine $\bar{u}^1,\ldots,\bar{u}^q \in \mathbb{R}^n$ by (1). For each $j = 1,\ldots,q$, equalities (1) constitute a system of linear equations with respect to the components of $\bar{u}^j$, which due to the linear independence of $h_1'(x^0), \ldots, h_q'(x^0)$ has a solution. Moreover, the vectors $\bar{u}^1,\ldots,\bar{u}^q$ solving this system are linearly independent and $\mathbb{R}^n$ is decomposed into a direct sum $\mathbb{R}^n = L \oplus L'$, where $L = \ker h'(x^0)$ and $L' = \operatorname{lin}\{\bar{u}^1,\ldots,\bar{u}^q\}$. Let $u^1,\ldots,u^{n-q}$ be $n-q$ linearly independent vectors in $L = \ker h'(x^0)$. We consider the system of equations:

$$h_k(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j \bar{u}^j) = 0\,, \quad k = 1, \ldots, q\,. \tag{6}$$

Taking $\tau_1, \ldots, \tau_{n-q}$ as independent variables and $\sigma_1, \ldots, \sigma_q$ as dependent variables, we see that this system satisfies the requirements of the implicit function theorem at the point $\tau_1 = \cdots = \tau_{n-q} = 0$, $\sigma_1 = \cdots = \sigma_q = 0$ (at this point h_k take values $h_k(x^0) = 0$ because x^0 is feasible, and the Jacobian is the unit matrix and hence nondegenerate). The implicit function theorem gives that in a neighbourhood of x^0 given by $|\tau_i| < \bar{\tau}$, $i = 1, \ldots, n-q$, $|\sigma_j| < \bar{\sigma}$, $j = 1, \cdots, q$, this system possesses a unique solution $\sigma_j = \sigma_j(\tau_1, \ldots, \tau_{n-q})$, $j = 1, \ldots, q$. The functions $\sigma_j = \sigma_j(\tau_1, \ldots, \tau_{n-q})$ are C^2 and $\sigma_j(0, \ldots, 0) = 0$.

Lemma 2. *Consider problem (1) with $h \in C^1$, for which $h_1'(x^0), \cdots, h_q'(x^0)$, are linearly independent, and C and K are closed convex cones. Then x^0 is a w-minimizer or i-minimizer of order k for (1) if and only if $\tau^0 = 0$ is respectively a w-minimizer or i-minimizer of order k for the problem*

$$\min_C \bar{f}(\tau_1, \ldots, \tau_{n-q}), \quad \bar{g}(\tau_1, \ldots, \tau_{n-q}) \in -K\,, \tag{7}$$

where

$$\begin{aligned} \bar{f}(\tau_1, \ldots, \tau_{n-q}) &= f(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j(\tau_1, \ldots, \tau_{n-q})\, \bar{u}^j)\,, \\ \bar{g}(\tau_1, \ldots, \tau_{n-q}) &= g(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j(\tau_1, \ldots, \tau_{n-q})\, \bar{u}^j)\,. \end{aligned} \tag{8}$$

Proof. From the implicit function theorem every feasible point x sufficiently close to x^0 admits a representation

$$x = x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j(\tau_1, \ldots, \tau_{n-q})\, \bar{u}^j \tag{9}$$

with $\tau = (\tau_1, \ldots, \tau_{n-q})$ close to $\tau^0 = 0$ and $\sigma_j(\tau_1, \ldots, \tau_{n-q})$ the unique C^1 solution of (6) with value $\sigma^0 = 0$ at $\tau^0 = 0$. Therefore it is obvious that x^0 is a w-minimizer for (1) if and only if τ^0 is a w-minimizer for (7). Suppose now that x^0 is an i-minimizer of order k. Then for some neighbourhood U of x^0 and some $A > 0$ inequality (2) has place. Replacing here x with (9) we get for all τ being close to τ^0 and feasible for (7) the inequality

$$D(\bar{f}(\tau) - \bar{f}(\tau^0), -C) \geq A\, \|x(\tau) - x^0\|^k\,. \tag{10}$$

Expressing $x = x(\tau)$ by (9) and applying the Taylor formula for $\sigma_j(\tau_1, \ldots, \tau_{n-q})$ and the forthcoming expressions for the derivatives we get

$$x(\tau) - x^0 = \sum_{i=1}^{n-q} \tau_i u^i + o(\|\tau\|)\,.$$

With the account that by choice $u^1, \ldots, u^{n-q}$ are linearly independent, we see that close to $\tau^0 = 0$ there exist positive constants A' and A'', such that

$$A' \|\tau - \tau^0\| \leq \|x(\tau) - x^0\| \leq A'' \|\tau - \tau^0\| .$$

These inequalities, together with (10) show that x^0 is an i-minimizer of order k for (1) if and only if $\tau^0 = 0$ is an i-minimizer of order k for (7). □

Now we calculate the derivatives of $\sigma_j(\tau_1, \ldots, \tau_{n-q})$ at $\tau^0 = (0, \ldots, 0)$. We have

$$\sigma_j\big|_{\tau^0} = \sigma_j(0, \ldots, 0) = 0, \quad j = 1 \ \ldots, q . \tag{11}$$

For the first-order derivatives differentiating (6) with respect to τ_i we get

$$h'_k(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j \bar{u}^j)(u^i + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_i} \bar{u}^j) = 0 .$$

For $\tau = \tau^0 = 0$ we get

$$h'_k(x^0)(u^i + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_i}\bigg|_{\tau^0} \bar{u}^j) = 0 ,$$

whence with account of $u^i \in \ker h'(x^0)$ and (1) we obtain

$$\frac{\partial \sigma_j}{\partial \tau_i}\bigg|_{\tau^0} = 0, \quad j = 1, \ldots, q, \quad i = 1, \ldots, n-q . \tag{12}$$

Now we calculate the second-order derivatives:

$$h''_k(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j \bar{u}^j)(u^{i'} + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_{i'}} \bar{u}^j,\ u^{i''} + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_{i''}} \bar{u}^j)$$
$$+ h'_k(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j \bar{u}^j) \sum_{j=1}^{q} \frac{\partial^2 \sigma_j}{\partial \tau_{i'} \partial \tau_{i''}} \bar{u}^j = 0 .$$

For $\tau = \tau^0 = 0$ with account of $u^i \in \ker h'(x^0)$ and (1) we get

$$h''_k(x^0)(u^{i'},\ u^{i''}) + \sum_{j=1}^{q} \frac{\partial^2 \sigma_j}{\partial \tau_{i'} \partial \tau_{i''}}\bigg|_{\tau^0} h'_k(x^0)\, \bar{u}^j = 0 .$$

After all, substituting k with j, we obtain

$$\frac{\partial^2 \sigma_j}{\partial \tau_{i'} \partial \tau_{i''}}\bigg|_{\tau^0} = -h''_j(x^0)(u^{i'},\ u^{i''}), \quad j = 1, \ldots, q, \quad i', i'' = 1, \ldots, n-q . \tag{13}$$

Proof of Theorem 2. According to Lemma 2 to show that x^0 is an i-minimizer of order two for (1) we must show that $\tau^0 = 0$ is an i-minimizer of

order two for the problem with only inequality constraints (7). For this purpose it is enough to check that the sufficient conditions of Theorem 1 applied to problem (7) are satisfied. Since $\bar{g}(\tau^0) = g(x^0)$ we see that x^0 feasible for (1) implies τ^0 feasible for (7). Similarly $K[-\bar{g}(\tau^0)] = K[-g(x^0)]$. Theorem 1 reformulated for problem (7) gives:

Suppose that for each $\tau \in \mathbb{R}^{n-q} \setminus \{0\}$ *one of the following two conditions holds:*

$$\bar{\mathbb{S}}': \qquad (\bar{f}'(\tau^0)\tau,\, \bar{g}'(\tau^0)\tau) \notin -(C \times K[-\bar{g}(\tau^0)])\,,$$

$$\bar{\mathbb{S}}'': \quad (\bar{f}'(\tau^0)\tau,\, \bar{g}'(\tau^0)\tau) \in -(C \times K[-\bar{g}(\tau^0)] \setminus intC \times intK[-\bar{g}(\tau^0)])$$
$$and\ \forall\, (\bar{y}^0, \bar{z}^0) \in (\bar{f}(\tau^0), \bar{g}(\tau^0))''_\tau : \exists\, (\xi^0, \eta^0) \in \bar{\Delta}(0) :$$
$$\langle \xi^0,\, \bar{y}^0 \rangle + \langle \eta^0,\, \bar{z}^0 \rangle > 0\,.$$

Then τ^0 *is an i-minimizer of order two for problem (7). Here*

$$\bar{\Delta}(\tau^0) =$$

$$\{(\xi, \eta) \in C' \times K' \mid (\xi, \eta) \neq (0,0),\ \langle \eta, \bar{g}(\tau^0) \rangle = 0,\ \langle \xi, \bar{f}'(\tau^0) \rangle + \langle \eta, \bar{g}'(\tau^0) \rangle = 0\}.$$

We prove the theorem by showing that conditions $\mathbb{S}'$ and $\mathbb{S}''$ imply respectively $\bar{\mathbb{S}}'$ and $\bar{\mathbb{S}}''$. To show that $\mathbb{S}'$ implies $\bar{\mathbb{S}}'$ we get consecutively:

$$\frac{\partial}{\partial \tau_i} \bar{f}(\tau_1, \ldots, \tau_{n-q}) = \frac{\partial}{\partial \tau_i} f(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j(\tau_1, \ldots, \tau_{n-q})\, \bar{u}^j)$$
$$= f'(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j\, \bar{u}^j)(u^i + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_i}\, \bar{u}^j),$$

$$\frac{\partial}{\partial \tau_i} \bar{f}(0) = f'(x^0)u^i = (f'_1(x^0)u^i,\, \ldots,\, f'_m(x^0)u^i),$$

$$\bar{f}'(0)\,\tau = \sum_{i=1}^{n-q} \frac{\partial}{\partial \tau_i} \bar{f}(0)\,\tau_i = f'(x^0) \sum_{i=1}^{n-q} \tau_i u^i. \tag{14}$$

Similarly

$$\frac{\partial}{\partial \tau_i} \bar{g}(\tau_1, \ldots, \tau_{n-q}) = g'(x^0 + \sum_{i=1}^{n-q} \tau_i u^i + \sum_{j=1}^{q} \sigma_j\, \bar{u}^j)(u^i + \sum_{j=1}^{q} \frac{\partial \sigma_j}{\partial \tau_i}\, \bar{u}^j),$$

$$\frac{\partial}{\partial \tau_i} \bar{g}(0) = (g'_1(x^0)u^i,\, \ldots,\, g'_p(x^0)u^i),$$

$$\bar{g}'(0)\,\tau = \sum_{i=1}^{n-q} \frac{\partial}{\partial \tau_i} \bar{g}(0)\,\tau_i = g'(x^0) \sum_{i=1}^{n-q} \tau_i u^i.$$

Putting

$$u = \sum_{i=1}^{n-q} \tau_i u^i \in \ker h'(x^0) \tag{15}$$

we see that while τ varies in $\mathbb{R}^{n-q} \setminus \{0\}$ the vector u takes all values from $\ker h'(x^0) \setminus \{0\}$. Consequently condition $\bar{\mathbb{S}}'$ is equivalent to $\mathbb{S}'$, that is to

$$(f'(x^0)u,\, g'(x^0)u) \notin -(C \times K[-g(x^0)]) \quad \text{for} \quad u \in \ker h'(x^0) \setminus \{0\}. \tag{16}$$

Next we show that $\mathbb{S}''$ implies $\bar{\mathbb{S}}''$. The above calculations show the equivalence of the first parts of $\mathbb{S}''$ and $\bar{\mathbb{S}}''$, where only first order derivatives appear. Now we compare the second parts of $\mathbb{S}''$ and $\bar{\mathbb{S}}''$. For this purpose we must find first a relation between the Dini derivatives of $(f(x^0),\, g(x^0))''_u$ and $(\bar{f}(\tau^0),\, \bar{g}(\tau^0))''_\tau$. Initially we will consider the case of $f,\, g \in C^2$. Then

$$(f(x^0),\, g(x^0))''_u = (f''_u(x^0), g''_u(x^0)) = (f''(x^0)(u,u),\, g''(x^0)(u,u))$$

is a singleton. Similarly $\bar{f},\, \bar{g} \in C^2$ and

$$(\bar{f}(0),\, \bar{g}(0))''_\tau = (\bar{f}''(0)(\tau,\tau),\, \bar{g}''(0)(\tau,\tau))$$

is a singleton. We have consecutively

$$\frac{\partial^2}{\partial\tau_{i'}\partial\tau_{i''}}\bar{f}(\tau_1,\ldots,\tau_{n-q}) = \frac{\partial^2}{\partial\tau_{i'}\partial\tau_{i''}} f(x^0 + \sum_{i=1}^{n-q}\tau_i u^i + \sum_{j=1}^{q}\sigma_j \bar{u}^j)$$

$$= f''(x^0 + \sum_{i=1}^{n-q}\tau_i u^i + \sum_{j=1}^{q}\sigma_j \bar{u}^j)(u^{i'} + \sum_{j=1}^{q}\frac{\partial\sigma_j}{\partial\tau_{i'}}\bar{u}^j,\, u^{i''} + \sum_{j=1}^{q}\frac{\partial\sigma_j}{\partial\tau_{i''}}\bar{u}^j)$$

$$+ f'(x^0 + \sum_{i=1}^{n-q}\tau_i u^i + \sum_{j=1}^{q}\sigma_j \bar{u}^j)\sum_{j=1}^{q}\frac{\partial^2\sigma_j}{\partial\tau_{i'}\partial\tau_{i''}}\bar{u}^j,$$

$$\frac{\partial^2}{\partial\tau_{i'}\partial\tau_{i''}}\bar{f}(0) = f''(x^0)(u^{i'},\, u^{i''}) - \sum_{j=1}^{q} h''_j(x^0)(u^{i'},\, u^{i''})\, f'(x^0)\bar{u}^j. \tag{17}$$

Therefore for u given by (4) we have

$$\bar{f}''(0)(\tau,\tau) = f''(x^0)(u,u) - \sum_{j=1}^{q} h''_j(x^0)(u,u)\, f'(x^0)\bar{u}^j. \tag{18}$$

Similarly

$$\bar{g}''(0)(\tau,\tau) = g''(x^0)(u,u) - \sum_{j=1}^{q} h''_j(x^0)(u,u)\, g'(x^0)\bar{u}^j. \tag{19}$$

Now we show that when the assumptions on f and g are relaxed from C^2 to $C^{1,1}$ still there exist formulas similar to (18) and (19). In fact the only

reason to consider in advance the case of $f, g \in C^2$ was to elaborate some heuristics. In the relaxed case we show the following result. Let $f, g \in C^{1,1}$ and $h \in C^2$ be such that $h'_1(x^0), \dots, h'_q(x^0)$ are linearly independent. Suppose that $(\bar{y}^0, \bar{z}^0) \in (\bar{f}(0), \bar{g}(0))''_\tau$ and

$$\begin{aligned} \bar{y}^0 &= \lim_k \frac{2}{t_k^2} \left(\bar{f}(t_k\tau) - \bar{f}(0) - t_k \bar{f}'(0)\tau \right), \\ \bar{z}^0 &= \lim_k \frac{2}{t_k^2} \left(\bar{g}(t_k\tau) - \bar{g}(0) - t_k \bar{g}'(0)\tau \right). \end{aligned} \tag{20}$$

Let $u = u(\tau)$ be determined by (4). We will prove that the following limits exist

$$\begin{aligned} y^0 &= \lim_k \frac{2}{t_k^2} \left(f(x^0 + t_k u) - f(x^0) - t_k f'(x^0)u \right), \\ z^0 &= \lim_k \frac{2}{t_k^2} \left(\bar{g}(x^0 + t_k u) - g(x^0) - t_k g'(x^0)u \right), \end{aligned} \tag{21}$$

and satisfy (similarly to (18)–(19)) the relations

$$\begin{aligned} \bar{y}^0 &= y^0 - \sum_{j=1}^{q} h''_j(x^0)(u,u)\, f'(x^0)\bar{u}^j, \\ \bar{z}^0 &= z^0 - \sum_{j=1}^{q} h''_j(x^0)(u,u)\, g'(x^0)\bar{u}^j. \end{aligned} \tag{22}$$

Fix τ. Let now t be a positive real variable and put for brevity $\hat{u} = u + (1/t) \sum_{j=1}^{q} \sigma_j(t\tau)\, \bar{u}^j$. Then

$$\frac{2}{t^2} \left(\bar{f}(t\tau) - \bar{f}(0) - t \bar{f}'(0)\tau \right)$$

$$= \frac{2}{t^2} \left(f(x^0 + t\hat{u}) - f(x^0) - t f'(0)\hat{u} \right) + \frac{2}{t^2} f'(x^0) \sum_{j=1}^{q} \sigma_j(t\tau)\, \bar{u}^j.$$

The Taylor formula with regard to (4), (1) and (13) gives

$$\sigma_j(t\tau) = \frac{1}{2}\sigma''_j(\tau^0)(t\tau, t\tau) + o(t^2) = -\frac{1}{2} t^2 h''_j(x^0)(u,u) + o(t^2),$$

whence

$$\frac{2}{t^2} \left(\bar{f}(t\tau) - \bar{f}(0) - t \bar{f}'(0)\tau \right) = \frac{2}{t^2} \left(f(x^0 + t\hat{u}) - f(x^0) - t f'(0)\hat{u} \right)$$

$$- \sum_{j=1}^{q} h''_j(x^0)(u,u)\, f'(x^0)\bar{u}^j.$$

A similar representation exists for f replaced by g. From these representations and (20) it follows that there exist the limits

$$\hat{y}^0 = \lim_k \frac{2}{t_k^2}\left(f(x^0 + t_k\hat{u}) - f(x^0) - t_k\, f'(x^0)\hat{u}\right),$$
$$\hat{z}^0 = \lim_k \frac{2}{t_k^2}\left(g(x^0 + t_k\hat{u}) - g(x^0) - t_k\, g'(x^0)\hat{u}\right),$$

and

$$\begin{aligned} \bar{y}^0 &= \hat{y}^0 - \sum_{j=1}^{q} h_j''(x^0)(u,u)\, f'(x^0)\bar{u}^j, \\ \bar{z}^0 &= \hat{z}^0 - \sum_{j=1}^{q} h_j''(x^0)(u,u)\, g'(x^0)\bar{u}^j. \end{aligned} \tag{23}$$

Applying now Lemma 1 we get

$$\begin{aligned} &\Big\| \frac{2}{t_k^2}\left(f(x^0 + t_k\hat{u}) - f(x^0) - t_k\, f'(x^0)\hat{u}\right) \\ &- \frac{2}{t_k^2}\left(f(x^0 + t_k u) - f(x^0) - t_k\, f'(x^0)u\right) \Big\| \\ &\le L\left(\|\hat{u}\| + \|u\|\right)\|\hat{u} - u\| = L\left(\|\hat{u}\| + \|u\|\right)\frac{1}{t_k}\Big\|\sum_{j=1}^{q}\sigma_j(t_k\tau)\bar{u}^j\Big\| = o(1). \end{aligned}$$

A similar estimation exists for f replaced by g. In consequence, these inequalities show that there exist the limits (21) and it holds $y^0 = \hat{y}^0$, $z^0 = \hat{z}^0$. These equalities and formulas (10) imply (22).

Now we prove the second part of $\bar{\mathbb{S}}''$ as a consequence of $\mathbb{S}''$. Take $(\bar{y}^0, \bar{z}^0) \in (\bar{f}(\tau^0), \bar{g}(\tau^0))''_\tau$ with $\tau \in \mathbb{R}^{n-q} \setminus \{0\}$ and let (20) be satisfied. Then the limits (21) exist and define $(y^0, z^0) \in (f(x^0), g(x^0))''_u$, where u and τ are related by (4). The latter gives $u \in \ker h'(x^0) \setminus \{0\}$. Since $\mathbb{S}''$ holds, therefore there exists $(\xi^0, \eta^0, \zeta^0) \in \Delta(x^0)$ such that ζ^0 satisfies (5) and $\langle \xi^0, y^0\rangle + \langle \eta^0, z^0\rangle + \langle \zeta^0, h''(x^0)(u,u)\rangle > 0$. Substituting ζ^0 with (5) and applying (22) we get

$$\begin{aligned} 0 &< \langle \xi^0, y^0\rangle + \langle \eta^0, z^0\rangle + \langle \zeta^0, h''(x^0)(u,u)\rangle \\ &= \langle \xi^0, y^0 - \sum_{j=1}^{q} h_j''(x^0)(u,u)\, f'(x^0)\bar{u}^j\rangle + \langle \eta^0, z^0 - \sum_{j=1}^{q} h_j''(x^0)(u,u)\, g'(x^0)\bar{u}^j\rangle \\ &= \langle \xi^0, \bar{y}^0\rangle + \langle \eta^0, \bar{z}^0\rangle. \end{aligned}$$

To demonstrate that the second part of $\bar{\mathbb{S}}''$ is satisfied it remains to show that $(\xi^0, \eta^0) \in \bar{\Delta}(\tau^0)$. This follows from the following observations. We have $(\xi^0, \eta^0) \neq (0, 0)$, since otherwise (5) would give $(\xi^0, \eta^0, \zeta^0) = (0, 0, 0)$. It holds $\langle \eta^0, \bar{g}(\tau^0)\rangle = \langle \eta^0, g(x^0)\rangle = 0$. Finally, for $\tau \in \mathbb{R}^{n-q}$ and u determined by (4) we have

$$\langle \xi^0, \bar{f}'(\tau^0)\tau\rangle + \langle \eta^0, \bar{g}'(\tau^0)\tau\rangle = \langle \xi^0, f'(x^0)u\rangle + \langle \eta^0, g'(x^0)u\rangle = 0. \qquad \square$$

The next example shows that the optimality in particular vector optimization problems can be checked effectively on the base of Theorem 2 and known calculus rules.

Example 1. Consider problem (1), for which $n = 3$, $m = 2$, $p = 1$, $q = 2$, the cones are $C = \mathbb{R}^2_+$ and $K = \mathbb{R}_+$, and the functions f, g, h, are given by

$$f(x_1, x_2, x_3) = (-2x_1^2 - 2x_2^2 + x_3,\ x_1^2 + x_2^2 - x_3),$$

$$g(x_1, x_2, x_3) = x_1|x_1| + x_2|x_2| - x_3,$$

$$h(x_1, x_2, x_3) = (x_1 + x_2,\ x_3).$$

Then the point $x^0 = (0, 0, 0)$ is an i-minimizer of order 2, which can be established on the base of Theorem 2, as it is shown below.

The problem is $C^{1,1}$ and not C^2 because of the function g. We have

$$f(x^0) = (0,0), \quad g(x^0) = 0, \quad h(x^0) = (0,0).$$

The point x^0 is feasible and it holds $C' = \mathbb{R}^2_+$, $K' = \mathbb{R}_+$, $K[-g(x^0)] = \mathbb{R}_+$,

$$f'(x)u = (-4x_1u_1 - 4x_2u_2 + u_3,\ 2x_1u_1 + 2x_2u_2 - u_3),$$

$$g'(x)u = 2u_1|x_1| + 2u_2|x_2| - u_3,$$

$$f'(x^0)u = (u_3,\ -u_3), \quad g'(x^0)u = -u_3,$$

$$h_1'(x^0) = (1,\ 1,\ 0), \quad h_2'(x^0) = (0,\ 0,\ 1).$$

Obviously $h_1'(x^0)$ and $h_2'(x^0)$ are linearly independent, and

$$\ker h'(x^0) = \{u \in \mathbb{R}^3 \mid u_1 + u_2 = 0,\ u_3 = 0\}.$$

$$(f'(x^0)u,\ g'(x^0)u) = (0,\ 0) \in \mathbb{R}^2 \times \mathbb{R} \quad \text{for} \quad u \in \ker h'(x^0),$$

$$\Delta(x^0) = C' \times K' \times \mathbb{R}^2 \setminus \{(0,\ 0,\ 0)\}.$$

For each $u \in \ker h'(x^0) \setminus \{0\}$ condition $\mathbb{S}'$ is not satisfied. We prove that for such u condition $\mathbb{S}''$ holds. We have

$$(f'(x^0)u,\ g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus \operatorname{int} C \times \operatorname{int} K[-g(x^0)]).$$

The second-order derivatives at x^0 are

$$f''_u(x^0) = f''(x^0)(u,u) = (-4u_1^2 - 4u_2^2,\ 2u_1^2 + 2u_2^2),$$

$$g''_u(x^0) = 2u_1|u_1| + 2u_2|u_2|, \quad h''(x^0)(u,u) = (0,\ 0).$$

Turn attention that $(f(x^0),\ g(x^0))''_u = (f''_u(x^0),\ g''_u(x^0))$ is single-valued. The assumption $u \in \ker h'(x^0) \setminus \{0\}$ means $u_1 + u_2 = 0$, $u_3 = 0$. The vectors $\bar{u}^1$, $\bar{u}^2$ satisfying (1) can be chosen as $\bar{u}^1 = (1/2,\ 1/2,\ 0)$, $\bar{u}^2 = (0,\ 0,\ 1)$. According to (5) the vector $\zeta^0 = (\zeta_1^0, \zeta_2^0)$ is expressed by $\xi^0 = (\xi_1^0,\ \xi_2^0)$ and η_0 as $\zeta^0 = (0,\ -\xi_1^0 + \xi_2^0 + \eta^0)$. Now for $y^0 = f''(x^0)(u,u)$, $z^0 = g''_u(x^0)$, $\xi^0 = (0,\ 1) \in C'$, $\eta^0 = 0 \in K'[-g(x^0]$ and $u \in \ker h'(x^0) \setminus \{0\}$ we get

$$\langle \xi^0,\ y^0 \rangle + \langle \eta^0,\ z^0 \rangle + \langle \zeta^0,\ h''(x^0)(u,u) \rangle$$

$$= -4\xi_1^0(u_1^2 + u_2^2) + 2\xi_2^0(u_1^2 + u_2^2) + \eta^0(2u_1|u_1| + 2u_2|u_2|) = 2(u_1^2 + u_2^2) > 0,$$

which shows that condition $\mathbb{S}''$ holds.

5 Necessary Conditions

The following Theorem 3 gives second-order necessary conditions for the problem (3) with only inequality constraints.

Theorem 3 ([10]). *Consider problem (3) with f and g being $C^{1,1}$ functions, and C and K closed convex cones with nonempty interiors. Let x^0 be a w-minimizer for (3). Then for each $u \in \mathbb{R}^n$ the following two conditions hold:*

$$\mathbb{N}_i': \quad (f'(x^0)u,\, g'(x^0)u) \notin -(intC \times intK[-g(x^0)])\,,$$

$$\mathbb{N}_i'': \quad \text{if } (f'(x^0)u,\, g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus intC \times intK[-g(x^0)])$$
$$\text{then } \forall\, (y^0, z^0) \in (f(x^0), g(x^0))_u'' : \exists\, (\xi^0, \eta^0) \in \Delta_I(x^0) :$$
$$\langle \xi^0,\, y^0 \rangle + \langle \eta^0,\, z^0 \rangle \geq 0\,.$$

Here $\Delta_I(x^0)$ has the same meaning as in Theorem 1. Theorem 3 generalizes Theorem 3.1 in Liu, Neittaanmäki, Křížek [23], which states the same thesis under the stronger assumption that C and K are polyhedral cones and C is acute.

The same elimination procedure as in Theorem 2 reduces problem (1) with both equality and inequality constraints to a problem with only inequality constraints to which we can apply Theorem 3. In such a way we obtain the following result:

Theorem 4. *Consider problem (1) with $f, g \in C^{1,1}$ and $h \in C^2$, and C and K closed convex cones with nonempty interiors. Let the vectors $h_1'(x^0), \ldots, h_q'(x^0)$, which are the components of $h'(x^0)$, be linearly independent and let the vectors $\bar{u}^j \in \mathbb{R}^n$ be determined by (1). Suppose that x^0 is a w-minimizer for (1). Then for each $u \in \ker h'(x^0)$ the following two conditions hold:*

$$\mathbb{N}': \quad (f'(x^0)u,\, g'(x^0)u) \notin -(intC \times intK[-g(x^0)])\,,$$

$$\mathbb{N}'': \quad \text{if } (f'(x^0)u,\, g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus intC \times intK[-g(x^0)])$$
$$\text{then } \forall\, (y^0, z^0) \in (f(x^0), g(x^0))_u'' : \exists\, (\xi^0, \eta^0, \zeta^0) \in \Delta(x^0) :$$
$$\langle \xi^0,\, y^0 \rangle + \langle \eta^0,\, z^0 \rangle + \langle \zeta^0,\, h''(x^0)(u,u) \rangle \geq 0$$
$$\text{and } \zeta^0 = (\zeta_j^0)_{j=1}^q \text{ satisfies (5).}$$

Here $\Delta(x^0)$ has the same meaning as in Theorem 2. The next example shows that the finding of the solutions of particular vector optimization problems can be effectively based on Theorem 4 and known calculus rules.

Example 2. Consider problem (1), for which $n = 3$, $m = 2$, $p = 1$, $q = 2$, the cones are $C = \mathbb{R}^2_+$ and $K = \mathbb{R}_+$, and the functions f, g, h, are given by

$$f(x_1, x_2, x_3) = (-2x_1^2 - 2x_2^2 + x_3,\, x_1^2 + x_2^2 - x_3)\,,$$

$$g(x_1, x_2, x_3) = x_1|x_1| + x_2|x_2| - x_3\,,$$

$$h(x_1, x_2, x_3) = (x_1 + x_2,\, 3x_1^2 + 3x_2^2 - 2x_3)\,.$$

Then the point $x^0 = (0,\, 0,\, 0)$ is not a w-minimizer, which can be established on the base of Theorem 4, as it is shown below.

Like in Example 1 we have $f(x^0) = (0,0)$, $g(x^0) = 0$, $h(x^0) = (0,0)$, $C' = \mathbb{R}^2_+$, $K' = \mathbb{R}_+$, $K[-g(x^0)] = \mathbb{R}_+$, $f'(x^0)u = (u_3, -u_3)$, $g'(x^0)u = -u_3$, $h'_1(x^0) = (1, 1, 0)$, $h'_2(x^0) = (0, 0, -2)$. Obviously $h'_1(x^0)$ and $h'_2(x^0)$ are linearly independent, and

$$\ker h'(x^0) = \{u \in \mathbb{R}^3 \mid u_1 + u_2 = 0,\ u_3 = 0\}.$$

$$(f'(x^0)u,\ g'(x^0)u) = (0,\ 0) \in \mathbb{R}^2 \times \mathbb{R} \quad \text{for} \quad u \in \ker h'(x^0)\,,$$

$$\Delta(x^0) = C' \times K' \times \mathbb{R}^2 \setminus \{(0,\ 0,\ 0)\}\,.$$

For each $u \in \ker h'(x^0)$ condition $\mathbb{N}'$ is satisfied. We prove that for some $u \in \ker h'(x^0)$ condition $\mathbb{N}''$ with ζ^0 distinguished by (5) does not hold. Observe that for any such u the statement in the first part of $\mathbb{N}''$ is true

$$(f'(x^0)u,\ g'(x^0)u) \in -(C \times K[-g(x^0)] \setminus \operatorname{int} C \times \operatorname{int} K[-g(x^0)])\,.$$

The second-order derivatives at x^0 are

$$f''_u(x^0) = f''(x^0)(u,u) = (-4u_1^2 - 4u_2^2,\ 2u_1^2 + 2u_2^2)\,,$$

$$g''_u(x^0) = 2u_1|u_1| + 2u_2|u_2|\,, \quad h''(x^0)(u,u) = (0,\ 6u_1^2 + 6u_2^2)\,.$$

Here $(f(x^0),\ g(x^0))''_u = (f''_u(x^0),\ g''_u(x^0))$ is single-valued. The vectors $\bar{u}^1$, $\bar{u}^2$ satisfying (1) can be chosen as $\bar{u}^1 = (1/2,\ 1/2,\ 0)$, $\bar{u}^2 = (0,\ 0,\ -1/2)$. According to (5) the vector $\zeta^0 = (\zeta^0_1,\ \zeta^0_2)$ is expressed by $\xi^0 = (\xi^0_1,\ \xi^0_2)$ and η^0 as $\zeta^0 = (0,\ (1/2)\xi^0_1 - (1/2)\xi^0_2 - (1/2)\eta^0)$. Now for $y^0 = f''(x^0)(u,u)$, $z^0 = g''_u(x^0)$, $\xi^0 \in C'$, $\eta' \in K'[g(x^0)]$, $(\xi^0, \eta^0) \neq (0,0)$ and $u \in \ker h'(x^0) \setminus \{0\}$ we get

$$\langle \xi^0,\ y^0 \rangle + \langle \eta^0,\ z^0 \rangle + \langle \zeta^0,\ h''(x^0)(u,u) \rangle$$

$$= -4\xi^0_1(u_1^2 + u_2^2) + 2\xi^0_2(u_1^2 + u_2^2) + \eta^0(2u_1|u_1| + 2u_2|u_2|) + 6\zeta^0_2(u_1^2 + u_2^2)$$

$$= -\xi^0_1(u_1^2 + u_2^2) - \xi^0_2(u_1^2 + u_2^2) + \eta^0(2u_1|u_1| + 2u_2|u_2| - 3u_1^2 - 3u_2^2)$$

$$\leq -(\xi^0_1 + \xi^0_2 + \eta^0)(u_1^2 + u_2^2) < 0\,,$$

which shows that for any $u \in \ker h'(x^0) \setminus \{0\}$ condition $\mathbb{N}''$ with ζ^0 distinguished by (5) does not hold. Thus, in spite that condition $\mathbb{N}'$ is satisfied for any $u \in \ker h'(x^0)$, there are u for which $\mathbb{N}''$ fails. According to Theorem 4 the point x^0 is not a w-minimizer.

6 Final Comments

A natural question is, whether it is possible to relax the smoothness assumptions for the function h from C^2 to $C^{1,1}$. This problem is reasonable for the sake of the uniformity of the assumptions for all function data in the considered constrained problem (1). Having in mind the formulations of Theorems

2 and 4 it is not difficult to predict the anticipated result for the case of h being only $C^{1,1}$. It is clear by analogy, that the eventual proof should be based on an implicit function theorem for $C^{1,1}$ functions. Implicit function theorems in nonsmooth analysis are investigated by many authors and in many settings. Some variant with application to $C^{1,1}$ optimization gives Kummer [20]. However for our consideration the variant for directionally differentiable functions developed in Demyanov, Rubinov [5, Chapter VI, Section 1] seems to be more suitable. Still, there is a need for some adjustment. For instance, it is important to have calculation rules for the second-order Dini directional derivatives of the implicit function. Therefore, an attempt to move in this direction demands a development of new ideas and will overburden in some sense the present paper. For this reason we postpone the discussion on the possible relaxation of the smoothness assumptions for h.

References

1. B. Aghezzaf. Second-order necessary conditions of the Kuhn-Tucker type in multiobjective programming problems. Control Cybernet. 28(2):213–224, 1999.
2. A. Auslender. Stability in mathematical programming with nondifferentiable data. SIAM J. Control Optim., 22: 239–254, 1984.
3. J.-P. Aubin and H. Frankowska. Set-Valued Analysis. Birkhäuser, Boston, 1990.
4. S. Bolintinéanu and M. El Maghri. Second-order efficiency conditions and sensitivity of efficient points. J. Optim. Theory Appl., 98(3):569–592, 1998.
5. V. F. Demyanov and A. M. Rubinov. Constructive Nonsmooth Analysis. Peter Lang, Frankfurt am Main, 1995.
6. P. G. Georgiev and N. Zlateva. Second-order subdifferentials of $C^{1,1}$ functions and optimality conditions. Set-Valued Anal. 4(2): 101–117, 1996.
7. I. Ginchev. Higher order optimality conditions in nonsmooth vector optimization. In: A. Cambini, B. K. Dass, L. Martein (Eds.), "Generalized Convexity, Generalized Monotonicity, Optimality Conditions and Duality in Scalar and Vector Optimization", J. Stat. Manag. Syst. 5(1–3): 321–339, 2002.
8. I. Ginchev, A. Guerraggio, and M. Rocca. First-order conditions for $C^{0,1}$ constrained vector optimization. In: F. Giannessi, A. Maugeri (eds.), Variational analysis and applications, Proc. Erice, June 20– July 1, 2003, Kluwer Acad. Publ. & Springer, Dordrecht–Berlin, 2005, to appear.
9. I. Ginchev, A. Guerraggio, and M. Rocca. From scalar to vector optimization. Appl. Math., to appear.
10. I. Ginchev, A. Guerraggio, and M. Rocca. Second-order conditions in $C^{1,1}$ constrained vector optimization. In: J.-B Hiriart-Urruty, C. Lemarechal, B. Mordukhovich, Jie Sun, Roger J.-B. Wets (eds.), Variational Analysis, Optimization, and their Applications, Math. Program., Series B, to appear.
11. I. Ginchev and A. Hoffmann. Approximation of set-valued functions by single-valued one. Discuss. Math. Differ. Incl. Control Optim., 22: 33–66, 2002.
12. A. Guerraggio and D. T. Luc. Optimality conditions for $C^{1,1}$ vector optimization problems. J. Optim. Theory Appl., 109(3):615–629, 2001.
13. A. Guerraggio and D. T. Luc. Optimality conditions for $C^{1,1}$ constrained multiobjective problems. J. Optim. Theory Appl., 116(1):117–129, 2003.

14. J.-B. Hiriart-Urruty. New concepts in nondifferentiable programming. Analyse non convexe, Bull. Soc. Math. France 60:57–85, 1979.
15. J.-B. Hiriart-Urruty. Tangent cones, generalized gradients and mathematical programming in Banach spaces. Math. Oper. Res. 4:79–97, 1979.
16. J.-B. Hiriart-Urruty, J.-J Strodiot, and V. Hien Nguen: Generalized Hessian matrix and second order optimality conditions for problems with $C^{1,1}$ data. Appl. Math. Optim. 11:169–180, 1984.
17. B. Jiménez. Strict efficiency in vector optimization. J. Math. Anal. Appl. 265: 264–284, 2002.
18. B. Jiménez and V. Novo. First and second order conditions for strict minimality in nonsmooth vector optimization. J. Math. Anal. Appl. 284:496–510, 2003.
19. D. Klatte and K. Tammer. On the second order sufficient conditions to perturbed $C^{1,1}$ optimization problems. Optimization 19:169–179, 1988.
20. B. Kummer. An implicit function theorem for $C^{0,1}$ equations and parametric $C^{1,1}$ optimization. J. Math. Anal. Appl. 158:35–46, 1991.
21. L. Liu. The second-order conditions of nondominated solutions for $C^{1,1}$ generalized multiobjective mathematical programming. J. Syst. Sci. Math. Sci. 4(2):128–138, 1991.
22. L. Liu and M. Křížek. The second-order optimality conditions for nonlinear mathematical programming with $C^{1,1}$ data. Appl. Math. 42:311–320, 1997.
23. L. Liu, P. Neittaanmäki, and M. Křížek. Second-order optimality conditions for nondominated solutions of multiobjective programming with $C^{1,1}$ data. Appl. Math. 45:381–397, 2000.
24. C. Malivert. First and second order optimality conditions in vector optimization. Ann. Sci. Math. Québec 14:65–79, 1990.
25. R. T. Rockafellar. Convex Analysis. Princeton University Press, Princeton, 1970.
26. S. Wang. Second-order necessary and sufficient conditions in multiobjective programming. Numer. Funct. Anal. Opim. 12:237–252, 1991.
27. X. Q. Yang. Second-order conditions in $C^{1,1}$ optimization with applications. Numer. Funct. Anal. Optim. 14:621–632, 1993.
28. X. Q. Yang and V. Jeyakumar. Generalized second-order directional derivatives and optimization with $C^{1,1}$ functions. Optimization 26:165–185, 1992.

Benson Proper Efficiency in Set-Valued Optimization on Real Linear Spaces*

E. Hernández[1], B. Jiménez[2] and V. Novo[3]

[1] Departamento de Matemática Aplicada, E.T.S.I. Industriales, UNED, c/ Juan del Rosal 12, 28040 Madrid, Spain. ehernandez@ind.uned.es
[2] Departamento de Economía e Historia Económica, Universidad de Salamanca, Facultad de Economía y Empresa, Campus Miguel de Unamuno, s/n, 37007 Salamanca, Spain. bjimen1@encina.pntic.mec.es
[3] Departamento de Matemática Aplicada, E.T.S.I. Industriales, UNED, c/ Juan del Rosal 12, 28040 Madrid, Spain. vnovo@ind.uned.es

Summary. In this work, a notion of cone-subconvexlikeness of set-valued maps on linear spaces is given and several characterizations are obtained. An alternative theorem is also established for this kind of set-valued maps. Using the notion of vector closure introduced recently by Adán and Novo, we also provide, in this framework, an adaptation of the proper efficiency in the sense of Benson for set-valued maps. The previous results are then applied to obtain different optimality conditions for this Benson-vectorial proper efficiency by using scalarization and multiplier rules.

1 Introduction

In the last decades, there has been an increasing interest in vector optimization problems with set-valued objectives or constraints. See, for instance, [7, 8, 10, 12, 13, 14, 15, 16] and references therein. This kind of optimization problems with set-valued maps are closely related to stochastic programming, control theory and economic theory.

In this work, we introduce a new concept of proper efficiency in the sense of Benson for an optimization problem with set-valued maps on real linear spaces, and we characterize this concept of proper efficiency. We introduce this Benson vectorial proper efficiency by using concepts and results given by Adán and Novo [1, 2, 3, 4, 5]. We extend the notion of cone-subconvexlikeness of set-valued maps on linear spaces and give several characterizations. We establish separation theorems and an alternative theorem for solid cones. We also analyze the behaviour of a cone-subconvexlike set-valued map via a positive linear operator. We prove scalarization theorems and characterize the

* This research was partially supported by Ministerio de Ciencia y Tecnología (Spain), project BFM2003-02194.

Benson-vectorial proper efficiency in optimization problems of set-valued maps with cone-subconvexlikeness. Lastly, using a new generalized Slater constraint qualification, we obtain a Lagrange multiplier rule of algebraic type for vector optimization problems with set-valued maps.

2 Notations and Preliminaries

Throughout this work, we will assume always, unless stated specifically otherwise, that Y is a real linear space partially ordered by a convex cone $K \subset Y$ and A is a nonempty subset of Y. Let $\text{cone}(A)$, $\text{co}(A)$, $\text{aff}(A)$, $\text{span}(A)$ and $L(A) = \text{span}(A - A)$ denote the generated cone, convex hull, affine hull, linear hull and associated linear subspace of A, respectively. In this section, we recall some algebraic concepts and known results.

The core (algebraic interior) and the intrinsic core (relative algebraic interior) of A are defined, respectively, as follows:

$$\text{cor}(A) = \{y \in A : \ \forall v \in Y, \ \exists t > 0, \ \forall \alpha \in [0, t], \ y + \alpha v \in A\},$$
$$\text{icr}(A) = \{y \in A : \ \forall v \in L(A), \ \exists t > 0, \ \forall \alpha \in [0, t], \ y + \alpha v \in A\}.$$

We say that A is solid (relatively solid) if $\text{cor}(A) \neq \emptyset$ ($\text{icr}(A) \neq \emptyset$). It is clear that if $\text{cor}(A) \neq \emptyset$ then $\text{cor}(A) = \text{icr}(A)$ because $L(A) = Y$.

It is well-known that for finite dimensional spaces there exist sets which are not solid but they are relatively solid, for example, any segment, ray or line in $\mathbb{R}^2$. At the end of this section we show an example in infinite dimension (see Example 1).

The algebraic closure of a set A is defined by

$$\text{lin}(A) = A \cup \{y \in Y : \ \exists a \in A, \ [a, y) \subset A\}.$$

Except for solid convex sets, this concept is not satisfactory as a substitute for topological closure. In order to solve this problem, Adán and Novo [4] have introduced a weaker closure of algebraic type, which was called vector closure. This vector closure coincides with the algebraic closure for convex sets, and coincides with the topological closure for solid convex sets.

Definition 1. *Let A be a nonempty subset of Y. The vector closure of A is the set* $\text{vcl}(A) = \{y \in Y : \ \exists v \in Y, \ \forall t > 0, \ \exists \alpha \in (0, t], \ y + \alpha v \in A\}$.

It is clear that $y \in \text{vcl}(A)$ if and only if there exist $v \in Y$ and a sequence $\lambda_n \to 0^+$ such that $y + \lambda_n v \in A$ for all n. The set A is called vectorially closed if $A = \text{vcl}(A)$.

We say that a cone K is pointed if $K \cap (-K) = \{0\}$. It is well-known that for a convex cone K, whose relative algebraic interior is non-empty, the following conditions hold:

(i) $\text{icr}(K) \cup \{0\}$ is a convex cone,
(ii) $\text{icr}(K) + K = \text{icr}(K)$,

(*iii*) $\operatorname{icr}(\operatorname{icr}(K)) = \operatorname{icr}(\operatorname{icr}(K) \cup \{0\}) = \operatorname{icr}(K)$.

Denote by Y' the algebraic dual of Y and by A^+ the positive dual cone of A, that is,

$$A^+ = \{\varphi \in Y' \colon \varphi(a) \geq 0,\ \forall a \in A\}$$

and $A^{+s} = \{\varphi \in Y' \colon \varphi(a) > 0,\ \forall a \in A \setminus \{0\}\}$ is the strict positive dual of A. A^+ is a vectorially closed convex cone and

$$[\operatorname{cone}(A)]^+ = \operatorname{cone}(A^+) = [\operatorname{conv}(A)]^+ = \operatorname{conv}(A^+) = A^+.$$

Other properties that will be used and appear in [4, 5] are the following:

(*i*) $A, B \subset Y,\ A \subset B \Rightarrow \operatorname{vcl}(A) \subset \operatorname{vcl}(B)$,
(*ii*) $[\operatorname{vcl}(\operatorname{cone}(\operatorname{conv}(A)))]^+ = A^+ = \operatorname{vcl}(A^+)$.

If Y is a topological vector space, the interior and closure of a set A are denoted by $\operatorname{int}(A)$ and $\operatorname{cl}(A)$, respectively. It is easy to check the following inclusions

$$A \subset \operatorname{lin}(A) \subset \operatorname{vcl}(A) \subset \operatorname{cl}(A).$$

To illustrate the notions above we give an example in infinite dimension.

Example 1. Let Y be the vector space of all sequences of real numbers, let S be the subspace of Y of all convergent sequences:

$$S = \{a = (a_n) \in Y :\ \exists \lim a_n = \alpha \in \mathbb{R}\},$$

and let K be the subset of S of all sequences with nonnegative limit:

$$K = \{(a_n) \in S :\ \lim a_n \geq 0\}.$$

It is clear that K is a nonpointed convex cone with $K \cap (-K) = c_0$, the linear space of the sequences converging to zero. Furthermore, the vector space generated by K is S, i.e., $L(K) = K - K = S$, and it is easy to check that K is vectorially closed. Its intrinsic core is

$$\operatorname{icr}(K) = \{a \in K :\ \lim a_n > 0\}.$$

Indeed, let $a \in K$ such that $\alpha = \lim a_n > 0$, and let us see that $a \in \operatorname{icr}(K)$, i.e., that $\forall v \in S = L(K)$, $\exists t_0 > 0$ such that $a + tv \in K$, $\forall t \in (0, t_0]$. As the sequence $v = (v_n) \in S$ there exists $\lim v_n = \beta$. Then $\lim(a_n + tv_n) = \alpha + t\beta \geq 0$ for all $t \in (0, t_0]$ if we choose

$$t_0 = \begin{cases} 1 & \text{if } \beta \geq 0 \\ -\alpha/\beta & \text{if } \beta < 0. \end{cases}$$

Now pick $a \in \operatorname{icr}(K)$. As $\operatorname{icr}(K) \subset K$ we have $\lim a_n \geq 0$. Suppose that $\lim a_n = 0$. Let $v = (v_n)$ defined by $v_n = 1$ for all $n \in \mathbb{N}$. Since $-v \in S$ and $a \in \operatorname{icr}(K)$ there exists $t_0 > 0$ such that $a + t(-v) \in K$, $\forall t \in (0, t_0]$. This implies that $-t_0 = \lim(a_n + t_0(-v_n)) \geq 0$, which is a contradiction.

The following cone separation theorem is due to Adán and Novo [5, Theorem 2.2].

Theorem 1. *Let M, K be two vectorially closed and relatively solid convex cones in Y. Let K^+ be solid. If $M \cap K = \{0\}$ then there exists a linear functional $\varphi \in Y'\backslash\{0\}$ such that $\forall k \in K$, $m \in M$, $\varphi(k) \geq 0 \geq \varphi(m)$ and furthermore $\forall k \in K\backslash\{0\}$, $\varphi(k) > 0$.*

Throughout this work, we assume that, unless indicated otherwise, X is a set, Y and Z are linear spaces, $K \subset Y$ and $D \subset Z$ are pointed relatively solid convex cones, and $F\colon X \longrightarrow 2^Y$ and $G\colon X \longrightarrow 2^Z$ are set-valued maps with domain X. The image of a subset A of X under F is denoted by $F(A) = \cup_{x\in A} F(x)$.

Consider the following unconstrained (P) and constrained (CP) vector optimization problems with set-valued maps:

$$(P) \quad \begin{cases} K - \operatorname{Min} F(x) \\ \text{subject to } x \in X, \end{cases}$$

$$(CP) \quad \begin{cases} K - \operatorname{Min} F(x) \\ \text{subject to } x \in X,\ G(x) \cap (-D) \neq \emptyset. \end{cases}$$

The feasible set of (CP) is defined by

$$\Omega = \{x \in X\colon\ G(x) \cap (-D) \neq \emptyset\}. \tag{1}$$

In [5], Adán and Novo have introduced the following concept of proper efficient point of a set $S \subset Y$ in the framework of vector optimization problems in partially ordered real linear spaces.

Definition 2. *The set of Benson-vectorial (BeV) proper efficient points of $S \subset Y$ is defined by*

$$BeV(S) = \{y \in S\colon\ \operatorname{vcl}(\operatorname{cone}(S - y + K)) \cap (-K) = \{0\}\}.$$

If we assume that Y is a topological linear space, and in this definition we replace the vector closure by the topological closure, we obtain the usual Benson (Be) proper efficiency defined in [6]. Because of $\operatorname{vcl}(S) \subset \operatorname{cl}(S)$, it is clear that $\operatorname{Be}(S) \subset \operatorname{BeV}(S)$.

For a vector optimization problem with set-valued maps, we introduce the following concept of proper efficient solution.

Definition 3. *A point $x \in X$ is called a Benson-vectorial (BeV) proper efficient solution of problem (P) if there exists*

$$y \in F(x) \cap BeV(F(X)).$$

The pair (x, y) is called a Benson-vectorial proper minimizer of (P).

3 Cone-Subconvexlike Set-Valued Maps

It is well-known that convexity plays an important role in optimization theory. In this section, we propose the following notion of cone-subconvexlikeness for set-valued maps on linear spaces. As we shall see presently, this concept is weaker than other concepts of cone-subconvexlikeness for set-valued maps.

Let X be a set, let $F\colon X \longrightarrow 2^Y$ be a set-valued map with $\mathrm{dom}(F) = X$ and let $K \subset Y$ be a relatively solid convex cone.

Definition 4. *F is said to be K-subconvexlike on X if $\exists k_0 \in \mathrm{icr}(K)$ such that $\forall x, x' \in X$, $\forall \alpha \in (0,1)$, $\forall \varepsilon > 0$,*

$$\varepsilon k_0 + \alpha F(x) + (1-\alpha)F(x') \subset F(X) + K.$$

Proposition 1. *The following statements are equivalent:*
(a) F is K-subconvexlike on X.
(b) $\forall k \in \mathrm{icr}(K)$, $\forall x, x' \in X$, $\forall \alpha \in (0,1)$,

$$k + \alpha F(x) + (1-\alpha)F(x') \subset F(X) + \mathrm{icr}(K).$$

(c) $\forall x, x' \in X$, $\forall \alpha \in (0,1)$, $\exists k \in K$ such that $\forall \varepsilon > 0$

$$\varepsilon k + \alpha F(x) + (1-\alpha)F(x') \subset F(X) + K. \tag{2}$$

(d) $F(X) + \mathrm{icr}(K)$ is a convex set.

Proof. The implications (b) $\Rightarrow$ (a) $\Rightarrow$ (c) are clear. Let us see (c) $\Rightarrow$ (b). Let $k \in \mathrm{icr}(K)$, $x, x' \in X$, $\alpha \in (0,1)$. Then, by assumption, $\exists k' \in K$ such that $\forall \varepsilon > 0$ condition (2) holds (with k' instead of k). As $k \in \mathrm{icr}(K) = \mathrm{icr}(\mathrm{icr}(K))$, for $-k' \in L(K) = L(\mathrm{icr}(K) = \mathrm{span}(K - K)$ there exists $\varepsilon_0 > 0$ such that $k_0 = k + \varepsilon_0(-k') \in \mathrm{icr}(K)$. So,

$$\begin{aligned} k + \alpha F(x) + (1-\alpha)F(x') &= [\varepsilon_0 k' + \alpha F(x) + (1-\alpha)F(x')] + k_0 \\ &\subset F(X) + K + k_0 \subset F(X) + \mathrm{icr}(K) \end{aligned}$$

(the last inclusion is true because $K + \mathrm{icr}(K) \subset \mathrm{icr}(K)$).
(b) $\Rightarrow$ (d) Let $u, u' \in F(X) + \mathrm{icr}(K)$, $\alpha \in (0,1)$. Then, $u = y + k$, $u' = y' + k'$ with $y \in F(x)$, $y' \in F(x')$, $k, k' \in \mathrm{icr}(K)$, $x, x' \in X$. Therefore

$$\alpha u + (1-\alpha)u' = \alpha k + (1-\alpha)k' + \alpha y + (1-\alpha)y'.$$

As $\mathrm{icr}(K)$ is a convex set, $k_0 = \alpha k + (1-\alpha)k' \in \mathrm{icr}(K)$. So,

$$\alpha u + (1-\alpha)u' \in k_0 + \alpha F(x) + (1-\alpha)F(x') \subset F(X) + \mathrm{icr}(K).$$

(d) $\Rightarrow$ (b) Let $k \in \mathrm{icr}(K)$, $x, x' \in X$, $\alpha \in (0,1)$, $y \in F(x)$, $y' \in F(x')$, then

$$k + \alpha y + (1-\alpha)y' = \alpha(y+k) + (1-\alpha)(y'+k) \in F(X) + \mathrm{icr}(K)$$

because $F(X) + \mathrm{icr}(K)$ is a convex set by assumption, and $y + k, y' + k \in F(X) + \mathrm{icr}(K)$. □

Remark 1. Of course, we may define that F is K-subconvexlike on X in the sense of Li ([12], given for a solid cone), if $\exists k_0 \in \operatorname{icr}(K)$, $\forall x, x' \in X, \forall \alpha \in (0,1)$, $\forall \varepsilon > 0$, $\exists x'' \in X$ such that

$$\varepsilon k_0 + \alpha F(x) + (1-\alpha)F(x') \subset F(x'') + K.$$

However, this notion is more restrictive than Definition 4 (see Example 2). When $\operatorname{cor}(K) \neq \emptyset$, K-subconvexlikeness in the sense of Li becomes exactly Definition 2.2 in Li [12] (see also [14, Definition 1.2]).

Example 2. Consider $X = [-2,0]$, $F : X \longrightarrow 2^{\mathbb{R}^2}$ defined by $F(x) = [(-2,1),(0,2+x)]$ and $K = \mathbb{R}_+ \times \{0\}$. It follows that $F(X) + \operatorname{icr}(K)$ is a convex set so F is K-subconvexlike on X. However, F is not K-subconvexlike on X in the sense of Li. Indeed, given $k_0 \in \operatorname{icr}(K)$ if $x = 0$, $x' = -2$, $\alpha = \frac{1}{5}$ and $\varepsilon = 1$ then $\forall x'' \in X$

$$\varepsilon k_0 + \alpha F(x) + (1-\alpha)F(x') \not\subset F(x'') + K$$

as can easily be checked.

The previous proposition can be considered an extension of Lemmas 3.1 and 3.2 in Li [13] which are valid in a topological linear space Y provided with a convex cone K whose interior is nonempty.

In order to simplify the notations we introduce a new definition.

Definition 5. *A set-valued map $F\colon X \longrightarrow 2^Y$ is said to be relatively solid K-subconvexlike on X if the following conditions hold:*

(i) F is K-subconvexlike on X,

(ii) $\operatorname{icr}(F(X) + \operatorname{icr}(K)) \neq \emptyset$.

Remark 2. If Y is finite dimensional, condition (ii) is always true whenever F is K-subconvexlike because $F(X) + \operatorname{icr}(K)$ is a convex set.

Example 3. Let Y and K be the sets of Example 1. Let A be the convex cone

$$A = \{(a_n) \in Y : \ a_n \geq 0, \ \forall n \in \mathbb{N}\}$$

and $C = A + \operatorname{icr}(K) = \{(a_n) \in Y : \ \liminf a_n > 0\}$. We have that C is a convex set and $\operatorname{icr}(C) = \emptyset$. So $F\colon A \longrightarrow 2^Y$, given by $F(x) = x + K$, is K-subconvexlike on A but is not relatively solid K-subconvexlike on A. However, if we consider $F\colon K \longrightarrow 2^Y$ is relatively solid K-subconvexlike on K.

In Theorem 3 we establish an alternative theorem for K-subconvexlike set-valued maps with K solid. Previously, in Theorem 2 we establish a partial result of alternative type when K is only a relatively solid cone.

Theorem 2. *Let $K \subset Y$ be a relatively solid convex cone. Assume that $F : X \to 2^Y$ is relatively solid K-subconvexlike on X. If there is no $x \in X$ such that*

$$F(x) \cap (-\operatorname{icr}(K)) \neq \emptyset, \tag{3}$$

then $\exists \varphi \in K^+ \setminus \{0\}$ such that $\varphi(y) \geq 0 \quad \forall y \in F(X)$.

Proof. The set $F(X) + \operatorname{icr}(K)$ is convex by Proposition 1. From (3) it follows that $0 \notin F(X) + \operatorname{icr}(K)$. So, $0 \notin \operatorname{icr}(F(X) + \operatorname{icr}(K))$. Using the support theorem [9, Theorem 6.C], there exists $\varphi \in Y' \setminus \{0\}$ such that

$$\varphi(y + k) \geq 0 \quad \forall y \in F(X),\ \forall k \in \operatorname{icr}(K) \tag{4}$$

(and φ is strictly positive on $\operatorname{icr}(F(X) + \operatorname{icr}(K))$). With standard reasonings, from (4) it follows that $\varphi(k) \geq 0\ \forall k \in K$, i.e., $\varphi \in K^+$. If $\exists y \in F(X)$ such that $\varphi(y) < 0$, choosing $k \in \operatorname{icr}(K)$ with $\varphi(k)$ small enough we obtain that $\varphi(y + k) < 0$, which contradicts (4). Hence, $\varphi(y) \geq 0$ for all $y \in F(X)$. □

Theorem 3. *Let K be a solid convex cone. If F is K-subconvexlike on X, then exactly one of the following systems is consistent:*

(i) $\exists x \in X$ such that $F(x) \cap (-\operatorname{cor}(K)) \neq \emptyset$.

(ii) $\exists \varphi \in K^+ \setminus \{0\}$ such that $\forall y \in F(X)$, $\varphi(y) \geq 0$.

Proof. By [4, Proposition 6.(iii)], $\operatorname{cor}(F(X) + \operatorname{cor}(K)) = F(X) + \operatorname{cor}(K)$, and consequently, condition (ii) in Definition 5 is satisfied. Therefore, by Theorem 2, not (i) $\Rightarrow$ (ii). If we assume that both (i) and (ii) are satisfied, then there exist $x \in X$, $y \in F(x) \cap (-\operatorname{cor}(K))$ and $\varphi \in K^+ \setminus \{0\}$ such that $\varphi(y) \geq 0$. But, since $y \in -\operatorname{cor}(K)$ and $\varphi \in K^+ \setminus \{0\}$, we deduce that $\varphi(y) < 0$ and by Theorem 2.2 in [12] this is a contradiction. □

Remark 3. This theorem is slightly more general than Theorem 2.1 of Li [14] because the notion of K-subconvexlikeness of this author is more restrictive than our notion, even when $\operatorname{cor}(K) \neq \emptyset$ (see Remark 1). If we consider that Y is a topological vector space then Theorem 3 collapses into Lemma 3.3 in [13]. Indeed, when Y is a topological vector space and $\operatorname{int}(K) \neq \emptyset$, then $\operatorname{int}(K) = \operatorname{cor}(K)$ and the linear functional φ satisfying condition (ii) is continuous because we can apply Theorem 3.7 in [17] since the open set $\operatorname{int}(K)$ is contained in the set $\{y \in Y :\ \varphi(y) > 0\}$ [12, Lemma 2.2] as $\varphi \in K^+ \setminus \{0\}$. Let us note that if $\operatorname{cor}(K) = \emptyset$ and $\operatorname{icr}(K) \neq \emptyset$, then both (i) (with $\operatorname{icr}(K)$ instead of $\operatorname{cor}(K)$) and (ii) can be true. For instance, in $\mathbb{R}^2$, $K = \mathbb{R}_+ \times \{0\}$, $X = \{(x, 0) \colon x \in (0, 1]\}$, $F(x, 0) = (x, 0) - K$ and $\varphi(x, y) = y$.

Lemma 1. *Let S_1 be a relatively solid convex set of Y and $S_2 \subset Y$. If $S_1 \subset S_2$ and $\operatorname{vcl}(S_1) = \operatorname{vcl}(S_2)$, then $\operatorname{icr}(S_1) = \operatorname{icr}(S_2)$.*

Proof. One has $\operatorname{aff}(S_1) = \operatorname{aff}(S_2)$ because by assumption $\operatorname{vcl}(S_1) = \operatorname{vcl}(S_2)$ and for any set $S \subset Y$, $\operatorname{aff}(S) = \operatorname{aff}(\operatorname{vcl}(S))$. Hence, as $S_1 \subset S_2$ we deduce that $\operatorname{icr}(S_1) \subset \operatorname{icr}(S_2)$. On the other hand, $S_2 \subset \operatorname{vcl}(S_2) = \operatorname{vcl}(S_1)$ and as S_2 and $\operatorname{vcl}(S_1)$ have the same affine hull, we get that $\operatorname{icr}(S_2) \subset \operatorname{icr}(\operatorname{vcl}(S_1)) = \operatorname{icr}(S_1)$. The last equality is true by Proposition 4(i) in [4]. Consequently, the conclusion follows. □

Proposition 2. *Let S be a relative solid convex subset of Y and $\varphi : Y \to Z$ a linear map. Then $\varphi(\operatorname{icr}(S)) = \operatorname{icr}(\varphi(S))$.*

Proof. Firstly let us see that

$$\varphi(\mathrm{icr}(S)) \subset \mathrm{icr}(\varphi(S)). \tag{5}$$

(as a consequence, $\varphi(S)$ is relatively solid). It is obvious that $\varphi(L(S)) = L(\varphi(S))$. Take $a \in \mathrm{icr}(S)$ and let us prove that $\varphi(a) \in \mathrm{icr}(\varphi(S))$. Given $w \in L(\varphi(S))$, there exists $v \in L(S)$ satisfying $\varphi(v) = w$. As $a \in \mathrm{icr}(S)$, for $v \in L(S)$ there exists $t_0 > 0$ such that $a + tv \in S\ \forall t \in (0, t_0]$. From here,

$$\varphi(a) + tw = \varphi(a) + t\varphi(v) \in \varphi(S) \quad \forall t \in (0, t_0],$$

and therefore, $\varphi(a) \in \mathrm{icr}(\varphi(S))$. Now, the reverse inclusion: $\mathrm{icr}(\varphi(S)) \subset \varphi(\mathrm{icr}(S))$. For this aim, let us see that $\varphi(S)$ and $\varphi(\mathrm{icr}(S))$ have the same vector closure. We have that

$$\varphi(\mathrm{vcl}(S)) \subset \mathrm{vcl}(\varphi(S)). \tag{6}$$

Indeed, choose $b \in \mathrm{vcl}(S)$, then there exists $v \in Y$ such that $\forall \alpha' > 0\ \exists \alpha \in (0, \alpha']$ such that $b + \alpha v \in S$. Hence, $\varphi(b) + \alpha\varphi(v) \in \varphi(S)$. This means that $\varphi(b) \in \mathrm{vcl}(\varphi(S))$. The following inclusions are clear taking into account (6):

$$\varphi(S) \subset \varphi(\mathrm{vcl}(S)) = \varphi(\mathrm{vcl}(\mathrm{icr}(S))) \subset \mathrm{vcl}(\varphi(\mathrm{icr}(S))) \subset \mathrm{vcl}(\varphi(S)).$$

From this chain, we select the following:

$$\varphi(S) \subset \mathrm{vcl}(\varphi(\mathrm{icr}(S))) \subset \mathrm{vcl}(\varphi(S)).$$

Taking vector closure and using that $\mathrm{vcl}(\mathrm{vcl}(B)) = \mathrm{vcl}(B)$, if B is a relative solid convex set, by [4, Proposition 3(iii)] (as $\varphi(\mathrm{icr}(S)) = \varphi(\mathrm{icr}(\mathrm{icr}(S))) \subset \mathrm{icr}(\varphi(\mathrm{icr}(S)))$, by condition (5) and as S is a relative solid, $\varphi(\mathrm{icr}(S))$ is a relative solid too) we have that:

$$\mathrm{vcl}(\varphi(S)) \subset \mathrm{vcl}(\varphi(\mathrm{icr}(S))) \subset \mathrm{vcl}(\varphi(S)).$$

Therefore, $\mathrm{vcl}(\varphi(S)) = \mathrm{vcl}(\varphi(\mathrm{icr}(S)))$, and by Lemma 1,

$$\mathrm{icr}(\varphi(S)) = \mathrm{icr}(\varphi(\mathrm{icr}(S))) \subset \varphi(\mathrm{icr}(S)).$$

Using (5), we have the conclusion. □

Next we analyze the postcomposition of a K-subconvexlike set-valued map with a positive linear map.

Let $\mathcal{L}(Y, Z)$ be the set of all linear maps φ from Y to Z, and let $\mathcal{L}_+(Y, Z)$ be the subset of positive linear maps, i.e.,

$$\mathcal{L}_+(Y, Z) = \{\varphi \in \mathcal{L}(Y, Z) :\ \varphi(K) \subset D\}.$$

Proposition 3. *Let $F : X \longrightarrow 2^Y$ be K-subconvexlike on X. If $\varphi \in \mathcal{L}_+(Y, Z)$, then $\varphi \circ F$ is D-subconvexlike on X.*

Proof. By Proposition 1(c), $\forall x, x' \in X$, $\forall \alpha \in (0,1)$, $\exists k \in K$ such that $\forall \varepsilon > 0$ we have

$$\varepsilon k + \alpha F(x) + (1-\alpha)F(x') \subset F(X) + K,$$

and therefore,

$$\varepsilon\varphi(k) + \alpha(\varphi \circ F)(x) + (1-\alpha)(\varphi \circ F)(x') \subset (\varphi \circ F)(X) + \varphi(K) \subset (\varphi \circ F)(X) + D.$$

As $\varphi(k) \in D$, statement (c) of Proposition 1 is satisfied for $\varphi \circ F$ and consequently, $\varphi \circ F$ is D-subconvexlike on X. □

Corollary 1. *Let $(F,G) : X \longrightarrow 2^{Y \times Z}$ be $K \times D$-subconvexlike on X.*
(i) If $\varphi \in K^+$ then $(\varphi \circ F, G)$ is $\mathbb{R}_+ \times D$-subconvexlike on X.
(ii) If $\psi \in \mathcal{L}_+(Z,Y)$ then $F + \psi \circ G$ is K-subconvexlike on X.

Proof. It is enough to apply Proposition 3 to (F,G) and the positive linear function $(y,z) \in Y \times Z \mapsto (\varphi(y), z)$ in part (i), and to the positive linear function $(y,z) \in Y \times Z \mapsto y + \psi(z)$ in part (ii). □

4 Benson-Vectorial Proper Efficiency

In this section we analyze different optimality conditions for Benson-vectorial proper efficiency, by using a pointed relatively solid convex cone and K-subconvexlike set-valued maps. Firstly, we establish a necessary condition and a sufficient condition through scalarization. Then, we obtain optimality conditions by using multiplier rules of algebraic type.

Now X is a set, Y is a linear space and $K \subset Y$ is a pointed relatively solid convex cone.

Let $\varphi \in \mathcal{L}(Y, \mathbb{R})$. We can associate to problem (P) the following scalar optimization problem with a set-valued map:

$$(SP_\varphi) \quad \begin{cases} \text{Min } (\varphi \circ F)(x) \\ \text{subject to } x \in X. \end{cases}$$

Definition 6. *If $x_0 \in X$, $y_0 \in F(x_0)$ and*

$$\varphi(y_0) \leq \varphi(y) \quad \forall y \in F(x), \ \forall x \in X,$$

then x_0 is called a minimal solution of problem (SP_φ), and (x_0, y_0) is called a minimizer of problem (SP_φ).

Theorem 4. *Let $\varphi \in K^{+s}$. If (x_0, y_0) is a minimizer of (SP_φ) then (x_0, y_0) is a Benson-vectorial proper minimizer of (P).*

Proof. Assume that (x_0, y_0) is not a Benson-vectorial proper minimizer. Then there exists

$$y \in \text{vcl}[\text{cone}(F(X) - y_0 + K)] \cap (-K) \quad \text{with} \quad y \neq 0.$$

Then $y \in -K$ and, since $\varphi \in K^{+s}$, we have that

$$\varphi(y) < 0. \tag{7}$$

On the other hand, as $y \in \text{vcl}[\text{cone}(F(X) - y_0 + K)]$, due to the definition of vcl, there exist $v \in Y$ and a sequence $\lambda_n \to 0^+$ such that $y + \lambda_n v \in \text{cone}(F(X) - y_0 + K)$ for all n. So, there exist sequences $\{\alpha_n\} \subset \mathbb{R}^+$, $\{y_n\} \subset F(X)$ and $\{k_n\} \subset K$ such that $y + \lambda_n v = \alpha_n(y_n - y_0 + k_n)$. Since φ is linear, we deduce

$$\varphi(y) + \lambda_n \varphi(v) = \alpha_n(\varphi(y_n) - \varphi(y_0) + \varphi(k_n)). \tag{8}$$

By hypothesis (x_0, y_0) is a minimizer of (SP_φ) and $\varphi \in K^{+s}$ so we have that $\varphi(y) \geq \varphi(y_0)$ for all $y \in F(X)$ and $\varphi(k_n) \geq 0$ for all n. From this and (8) it follows that for all n

$$\varphi(y) + \lambda_n \varphi(v) \geq 0.$$

As $\lambda_n \to 0^+$, we get $\varphi(y) \geq 0$, which contradicts (7). Therefore (x_0, y_0) is a Benson-vectorial proper minimizer of (P). □

As a consequence of the previous result, if we consider a topological linear space Y and we replace the vector closure by the topological closure and the relative algebraic interior by the topological interior, the previous proof is valid too. Therefore, the result above is an extension of Theorem 4.1 in Li [13].

To establish sufficient conditions we need some convexity properties and the following lemma.

Lemma 2. *Let S be a relatively solid convex set of Y. Then*

$$\text{icr}(S) \subset \text{icr}(\text{cone}(S)). \tag{9}$$

Proof. Firstly, let us prove that

$$L(\text{cone}(S)) = \text{aff}(S \cup \{0\}) = \begin{cases} L(S) & \text{if } 0 \in \text{aff}(S) \\ L(S) + \mathbb{R}s_0 & \text{if } 0 \notin \text{aff}(S), \end{cases} \tag{10}$$

where s_0 is an arbitrary element of S and $\mathbb{R}s_0$ is the linear subspace generated by s_0. Indeed, the statement is obvious when $0 \in \text{aff}(S)$. Thus, assume that $0 \notin \text{aff}(S)$. The linear subspace $L(S) + \mathbb{R}s_0$ is the smallest affine variety which contains $S \cup \{0\}$ because:

1) $S \subset L(S) + s_0 \subset L(S) + \mathbb{R}s_0$ and $\{0\} \subset L(S) + \mathbb{R}s_0$.

2) If V is an affine variety containing $S \cup \{0\}$, then $\text{aff}(S) = L(S) + s_0 \subset V$ and V is a linear subspace of Y. So, $L(S) \subset V - s_0 = V$ and $\mathbb{R}s_0 \subset V$ since $s_0 \in S \subset V$. Therefore, $L(S) + \mathbb{R}s_0 \subset V$.

Secondly, let us see equation (9). Let $a \in \text{icr}(S)$, we have to prove that $\forall u \in L(\text{cone}(S))$,

$$\exists t_0 > 0 \text{ such that } a + tu \in \text{cone}(S) \quad \forall t \in (0, t_0]. \tag{11}$$

Taking into account equation (10), it is enough to prove (11) in the following cases: (i) $u \in L(S)$, (ii) $u = s_0$ and (iii) $u = -s_0$.

(i) Let $u \in L(S)$. As $a \in \text{icr}(S)$, then there is $t_0 > 0$ such that $a + tu \in S \subset \text{cone}(S)$ $\forall t \in (0, t_0]$, i.e., (11) is satisfied.

(ii) Now, $u = s_0$. Then, as $a, s_0 \in \text{cone}(S)$ we have $a + ts_0 \in \text{cone}(S)$ $\forall t \geq 0$ since $\text{cone}(S)$ is a convex cone.

(iii) Finally, $u = -s_0$. As $a \in \text{icr}(S) \subset S$ and $s_0 \in S$ (so $a - s_0 \in L(S)$), there exists $\gamma > 0$ such that

$$s_1 := s_0 + (1 + \gamma)(a - s_0) = a + \gamma(a - s_0) \in S.$$

The equation $a + t(-s_0) = \rho s_1$ in the unknown (t, ρ) has solution (t_0, ρ_0) where $t_0 = \gamma/(1+\gamma) > 0$ and $\rho_0 = 1/(1+\gamma) > 0$. Hence $a + t_0(-s_0) = \rho_0 s_1 \in \text{cone}(S)$, and therefore $[a, a + t_0(-s_0)] \subset \text{cone}(S)$ (i.e., (11) is true). □

Theorem 5. *Assume that K is vectorially closed and $\text{cor}(K^+) \neq \emptyset$. Let F be relatively solid K-subconvexlike on X. If (x_0, y_0) is a Benson-vectorial proper minimizer of (P) then there exists $\varphi \in K^{+s}$ such that (x_0, y_0) is a minimizer of (SP_φ).*

Proof. Since (x_0, y_0) is a Benson-vectorial proper minimizer then

$$-\text{vcl}[\text{cone}(F(X) - y_0 + K)] \cap K = \{0\}. \tag{12}$$

As $\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))] \subset \text{vcl}[\text{cone}(F(X) - y_0 + K)]$, then

$$-\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))] \cap K = \{0\}. \tag{13}$$

Let us see that $\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))]$ is a vectorially closed relatively solid convex cone. It is clear that $\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))]$ is a cone. Because of F is relatively solid K-subconvexlike on X, $\text{icr}[F(X) + \text{icr}(K)] \neq \emptyset$ and $F(X) + \text{icr}(K)$ is a convex set, then $\text{icr}[F(X) - y_0 + \text{icr}(K)] \neq \emptyset$ [5, Proposition 2.1(ii)] and $F(X) - y_0 + \text{icr}(K)$ is convex too. Therefore, $\text{cone}(F(X) - y_0 + \text{icr}(K))$ is convex and applying Lemma 2 we obtain that

$$\text{icr}[\text{cone}(F(X) - y_0 + \text{icr}(K))] \neq \emptyset.$$

Applying Proposition 3(iii)-(iv) in [4], we obtain that $\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))]$ is vectorially closed and convex. On the other hand, by Proposition 4(i) in [4], $\text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))]$ is a relatively solid set. Under these conditions we can apply the separation Theorem 1, so taking into account condition (13), there exists $\varphi \in K^{+s} \setminus \{0\}$ such that

$$\varphi(v) \geq 0 \quad \text{for all } v \in \text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))].$$

Since $F(X) - y_0 + \text{icr}(K) \subset \text{vcl}[\text{cone}(F(X) - y_0 + \text{icr}(K))]$ we have

$$\varphi(y) - \varphi(y_0) + \varphi(k) \geq 0 \quad \text{for all } y \in F(X) \text{ and } k \in \operatorname{icr}(K).$$

Due to $\varphi \in K^{+s}$ and $\lambda k \in \operatorname{icr}(K)$ for all $\lambda > 0$, it follows that

$$\varphi(y) - \varphi(y_0) \geq 0 \quad \text{for all } y \in F(X).$$

Therefore (x_0, y_0) is a minimizer of (SP_φ). □

From the theorems above we obtain the following corollary, which gives us a characterization of Benson-vectorial proper minimizers under K-subconvexlikeness.

Corollary 2. *Let K^+ be solid and K be vectorially closed. Let F be relatively solid K-subconvexlike on X. Then (x_0, y_0) is a Benson-vectorial proper minimizer of (P) if and only if (x_0, y_0) is a minimizer of (SP_φ) for some $\varphi \in K^{+s}$.*

Therefore if we consider a topological linear space Y and $\operatorname{int}(K) \neq \emptyset$ then Theorem 5 and Corollary 2 can be considered extensions of Theorem 4.2 and Corollary 4.1 in Li [13].

Finally, we give a generalized Slater constraint qualification in order to obtain a Lagrange multiplier rule of algebraic type for constrained vector optimization problems with set-valued maps.

Definition 7. *We say that the optimization problem (CP) satisfies the generalized Slater constraint qualification if there exists $x \in X$ such that $G(x) \cap -\operatorname{icr}(D) \neq \emptyset$.*

Theorem 6. *Let $cor(K^+) \neq \emptyset$. Suppose that (F, G) is relatively solid $K \times D$-subconvexlike on X, F is relatively solid K-subconvexlike on Ω and* $\operatorname{aff}(\operatorname{icr}(D)) = \operatorname{aff}(\operatorname{icr}[G(X) + \operatorname{icr}(D)])$. *If (CP) satisfies the generalized Slater constraint qualification and (x_0, y_0) is a Benson-vectorial proper minimizer of (CP) then there exists $T \in \mathcal{L}_+(Z, Y)$ such that $0 \in T(G(x_0))$ and (x_0, y_0) is a Benson-vectorial proper minimizer of the unconstrained problem*

$$\begin{aligned} &K - Min\ (F + T \circ G)(x) \\ &\textit{subject to } x \in X. \end{aligned}$$

Proof. Since F is relatively solid K-subconvexlike on Ω, we can apply Theorem 5 to problem (CP), then there exists a linear functional $\varphi \in K^{+s}$ such that (x_0, y_0) is a minimizer of the scalar problem

$$\operatorname{Min}\{\varphi[F(x)] \colon\ x \in \Omega\},$$

i.e.

$$\varphi(y) \geq \varphi(y_0) \quad \text{for all } y \in F(\Omega). \tag{14}$$

Let $H \colon X \longrightarrow 2^{\mathbb{R} \times Z}$ be the set-valued map defined by

$$H(x) = [\varphi(F(x)) - \varphi(y_0)] \times G(x) = \varphi(F(x)) \times G(x) - (\varphi(y_0), 0).$$

As a consequence of (14) we have

$$H(X) \cap -\mathrm{icr}(\mathbb{R}_+ \times D) = \emptyset. \tag{15}$$

Since (F, G) is $K \times D$-subconvexlike on X then, by Corollary 1(i), we have that $H = (\varphi \circ (F - y_0), G) = (\varphi \circ F - \varphi(y_0), G)$ is $\mathbb{R}_+ \times D$-subconvexlike on X. From here and by $\mathrm{icr}[(F, G)(X) + \mathrm{icr}(K \times D)] \neq \emptyset$, applying Proposition 2 we obtain that

$$\mathrm{icr}[(\varphi \circ (F - y_0), G)(X) + \mathrm{icr}(\mathbb{R}_+ \times D)] \neq \emptyset.$$

Thus, H is relatively solid $\mathbb{R}_+ \times D$-subconvexlike on X. Together with (15), by Theorem 2 applied to H, we obtain that there exists $(r, \psi) \in \mathbb{R}_+ \times D^+ \backslash \{(0,0)\}$ such that

$$r[\varphi(F(x) - y_0)] + \psi[G(x)] \geq 0 \quad \text{for all } x \in X \tag{16}$$

and (see the proof of Theorem 2)

$$(r, \psi)(y', z') > 0 \text{ for all } (y', z') \in \mathrm{icr}[(\varphi \circ (F - y_0), G)(X) + \mathrm{icr}(\mathbb{R}_+ \times D)]. \tag{17}$$

We note that $r > 0$. Otherwise, if $r = 0$ then from condition (17) it results

$$\psi(\mathrm{icr}[G(X) + \mathrm{icr}(D)]) > 0. \tag{18}$$

As a consequence of the generalized Slater constraint qualification, $0 \in G(X) + \mathrm{icr}(D)$ so $\mathrm{icr}(D) \subset G(X) + \mathrm{icr}(D)$. On the other hand, by hypothesis, $\mathrm{aff}(\mathrm{icr}(D)) = \mathrm{aff}(\mathrm{icr}[G(X) + \mathrm{icr}(D)])$, therefore

$$\mathrm{icr}(D) = \mathrm{icr}(\mathrm{icr}(D)) \subset \mathrm{icr}[G(X) + \mathrm{icr}(D)]$$

and by (18) we obtain that

$$\psi(\mathrm{icr}(D)) > 0. \tag{19}$$

Again, because of the generalized Slater constraint qualification, there exists some $x' \in X$ and $z' \in G(x') \cap -\mathrm{icr}(D) \neq \emptyset$ and, consequently, by (19), $\psi(z') < 0$ and by (16), $\psi(z') \geq 0$, which is a contradiction. Thus, $r > 0$. Since $x_0 \in \Omega$ and $\psi \in D^+$ then there exists $z' \in G(x_0) \cap -D$ such that $\psi(z') \leq 0$. Taking $x = x_0$ and $y_0 \in F(x_0)$ in (16) we have that $\psi(z') \geq 0$, so $\psi(z') = 0$. Hence,

$$0 \in \psi[G(x_0)]. \tag{20}$$

As $r \neq 0$ and $\varphi \in K^{+s}$, we can choose $k \in K$ such that $r\varphi(k) = 1$. We define the operator $T\colon Z \longrightarrow Y$ by

$$T(z) = \psi(z)k. \tag{21}$$

It is clear that $T(D) \subset K$, i.e., $T \in \mathcal{L}_+(Z,Y)$. By (20), $0 \in T(G(x_0))$ and consequently

$$y_0 \in F(x_0) \subset F(x_0) + T(G(x_0)).$$

Now, from (16) and (21) we have that for all $x \in X$

$$r\varphi[F(x)+T(G(x))] = r\varphi[F(x)]+\psi[G(x)]r\varphi(k) = r\varphi[F(x)]+\psi[G(x)] \geq r\varphi(y_0)$$

If we divide this inequality by $r > 0$ we obtain that (x_0, y_0) is a minimizer of the scalar problem

$$K - \text{Min}\{[\varphi \circ (F + T \circ G)](x) \colon\ x \in X\}.$$

According to Theorem 4, (x_0, y_0) is a Benson-vectorial proper minimizer of the unconstrained optimization problem

$$K - \text{Min}\{(F + T \circ G)(x) \colon\ x \in X\}. \qquad \square$$

Remark 4. It is easy to check that the condition $\text{aff}(\text{icr}(D)) = \text{aff}(\text{icr}(G(X) + \text{icr}(D))$ is weaker than $\text{cor}(D) \neq \emptyset$. Indeed if $\text{cor}(D) \neq \emptyset$ then

$$\text{aff}(\text{cor}(D)) = \text{aff}[\text{cor}(G(X) + \text{cor}(D))] = Z.$$

Theorem 7. *Consider problem (CP). Assume $\text{cor}(K^+) \neq \emptyset$. Let (F, G) be a $K \times D$-subconvexlike set-valued map on X. If there exists a positive linear operator $T \in \mathcal{L}_+(Z,Y)$ and a pair (x_0, y_0) with $x_0 \in \Omega$ and $y_0 \in F(x_0)$ such that:*

(i) (x_0, y_0) is a Benson-vectorial proper minimizer of the problem

$$K - Min\ (F + T \circ G)(x) \quad \textit{subject to } x \in X,$$

(ii) $0 \in T(G(x_0))$ and

(iii) $\text{icr}[(F + T \circ G)(X) + \text{icr}(K)] \neq \emptyset$.

Then (x_0, y_0) is a Benson-vectorial proper minimizer of problem (CP).

Proof. Since (F, G) is $K \times D$-subconvexlike on X by Corollary 1(ii) $F + T \circ G$ is K-subconvexlike on X. Moreover, by assumption (iii), $F + T \circ G$ is relatively solid K-subconvexlike on X. So, applying Theorem 5 there exists $\varphi \in K^{+s}$ such that for all $x \in X$

$$\varphi(F(x) + T(G(x))) \geq \varphi(y_0)$$

Hence,

$$\varphi(F(x)) + \varphi(T(G(x))) \geq \varphi(y_0) \text{ for all } x \in X \tag{22}$$

Therefore, if $x \in \Omega$, there exists $z \in G(x)$ such that $z \in -D$. On the other hand, as $T \in \mathcal{L}_+(Z,Y)$, $T(z) \in -K$ and $\varphi \in K^{+s}$, we obtain $\varphi(T(z)) \leq 0$. From this, according to (22) and taking $z \in G(x)$, for each $y \in F(x)$ we obtain

$$\varphi(y) \geq \varphi(y) + \varphi(T(z)) \geq \varphi(y_0).$$

Hence, for all $y \in F(\Omega)$, one has $\varphi(y) \geq \varphi(y_0)$. As $y_0 \in F(x_0) \subset F(\Omega)$, applying Theorem 4, (x_0, y_0) is a Benson-vectorial proper minimizer of the problem (CP). $\square$

Once again our results extend Theorems 5.1 and 5.2 in Li [13] which are given in the framework of topological linear spaces with solid cones.

Acknowledgements. The authors are grateful to the anonymous referee for his helpful comments and suggestions which led to the present improved version of the paper.

References

1. M. Adán and V. Novo. Partial and generalized subconvexity in vector optimization problems, J. Convex Anal. 8:583-594, 2001.
2. M. Adán and V. Novo. Optimality conditions for vector optimization problems with generalized convexity in real linear spaces. Optimization 51:73-91, 2002.
3. M. Adán and V. Novo. Efficient and weak efficient points in vector optimization with generalized cone convexity. Appl. Math. Lett. 16:221-225, 2003.
4. M. Adán and V. Novo. Weak efficiency in vector optimization using a clousure of algebraic type under cone-convexlikeness. Eur. J. Oper. Res. 149:641-653, 2003.
5. M. Adán and V. Novo. Proper efficiency in vector optimization on real linear spaces J. Optim. Theory Appl. 121:515-540, 2004.
6. H.P. Benson. Efficiency and proper efficiency in vector maximization with respect to cones. J. Math. Anal. Appl. 93:273-289, 1983.
7. H.W. Corley. Existence and langrangian duality for maximizations of set-valued functions. J. Optim. Theory Appl. 54:489-501, 1987.
8. H.W. Corley. Optimality conditions for maximizations of set-valued functions. J. Optim. Theory Appl. 58:1-10, 1988.
9. R.B. Holmes. Geometric Functional Analysis and its Applications. Springer-Verlag, New York, 1975.
10. Y.W. Huang. Optimality conditions for vector optimization with set-valued maps. Bull. Austral. Math. Soc. 66:317-330, 2002.
11. V. Jeyakumar. A generalization of a minimax theorem of Fan via a theorem of the alternative. J. Optim. Theory Appl. 48:525-533, 1986.
12. Z. Li. The optimality conditions for vector optimization of set-valued maps J. Math. Anal. Appl. 237:413-424, 1999.
13. Z.F. Li. Benson proper efficiency in the vector optimization of set-valued maps. J. Optim. Theory Appl. 98:623-649, 1998.
14. Z.F. Li. A theorem of the alternative and its application to the optimization of set-valued maps. J. Optim. Theory Appl. 100:365-375, 1999.
15. Z.F. Li ZF and G.Y. Chen. Lagrangian multipliers, saddle points, and duality in vector optimization of set-valued maps. J. Math. Anal. Appl. 215:297-316, 1997.
16. L.J. Lin. Optimization of set-valued functions. J. Math. Anal. Appl. 186:30-51, 1994.
17. J. Van Tiel. Convex analysis. An Introductory Text. John Wiley & Sons. Chichester, 1984.

Some Results About Proximal-Like Methods

A. Kaplan[1] and R. Tichatschke[2]

[1] Dept. of Mathematics, University of Trier, D-54286 Trier, Germany.
Al.Kaplan@tiscali.de
[2] Dept. of Mathematics, University of Trier, D-54286 Trier, Germany.
tichat@uni-trier.de

Summary. We discuss some ideas for improvement, extension and application of proximal point methods and the auxiliary problem principle to variational inequalities in Hilbert spaces. These methods are closely related and will be joined in a general framework, which admits a consecutive approximation of the problem data including applications of finite element techniques and the ε-enlargement of monotone operators. With the use of a "reserve of monotonicity" of the operator in the variational inequality, the concepts of weak- and elliptic proximal regularization are developed. Considering Bregman-function-based proximal methods, we analyze their convergence under a relaxed error tolerance criterion in the subproblems. Moreover, the case of variational inequalities with non-paramonotone operators is investigated, and an extension of the auxiliary problem principle with the use of Bregman functions is studied. To emphasize the basic ideas, we renounce all the proofs and sometimes precise descriptions of the convergence results and approximation techniques. Those can be found in the referred papers.

1 Introduction

Let $(V, \|\cdot\|)$ be a Hilbert space with the topological dual V' and the duality pairing $\langle\cdot,\cdot\rangle$ between V and V'.

The variational inequality

$$(VI) \qquad \text{find } u^* \in K \text{ and } q^* \in \mathcal{Q}(u^*): \quad \langle \mathcal{F}(u^*) + q^*, u - u^*\rangle \geq 0 \ \ \forall\, u \in K \tag{1}$$

is considered, assuming that $K \subset V$ is a a convex closed set, $\mathcal{Q}: V \to 2^{V'}$ is a maximal monotone operator and $\mathcal{F}: K \to V'$ is a weakly continuous operator with certain monotonicity properties.

Sometimes, in order to use notions and facts originating from convex optimization, we turn to the problem

$$(CP) \qquad \min\{J(u) : u \in K\},$$

where J is a proper convex, lower semicontinuous functional. It presents a particular case of VI (1) with $\mathcal{F} := \mathbf{0}$ and $\mathcal{Q} := \partial J$.

In the sequel, $\{K^k\}$ is a family of convex closed sets approximating K, $K^k \subset V$, and $\{\mathcal{Q}^k\}$ is a family of operators approximating $\mathcal{Q}$. Usually, it is supposed that $\mathcal{Q}^k$ is maximal monotone, or that

$$\mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k},$$

$\mathcal{Q}_\epsilon$ means the ϵ-enlargement of $\mathcal{Q}$.

In proximal-like methods and in the auxiliary problem principle (APP), we use a regularizing functional h of Bregman type with zone[3] S, or its generalization defined by (7).

The following general scheme for solving VI (1) is considered: at step $k+1$, having a current iterate u^k ($u^1 \in K \cap S$ is arbitrarily chosen) the point u^{k+1} is calculated by solving the problem

$$\begin{aligned}(P_\delta^k) \qquad & \text{find } u^{k+1} \in K^k \cap \bar{S}, \quad q^k \in \mathcal{Q}^k(u^{k+1}) : \\ & \langle \mathcal{F}(u^k) + q^k + \mathcal{L}^k(u^{k+1}) - \mathcal{L}^k(u^k) \\ & \quad + \chi_k(\nabla h(u^{k+1}) - \nabla h(u^k)), u - u^{k+1}\rangle \\ & \quad \geq -\delta_k \|u - u^{k+1}\| \quad \forall\, u \in K^k \cap \bar{S}.\end{aligned} \tag{2}$$

Here, $\bar{S}$ denotes the closure of S and $\{\delta_k\}$ and $\{\chi_k\}$ are non-negative sequences with

$$\lim_{k\to\infty} \delta_k = 0, \qquad 0 < \chi_k \leq \bar{\chi} < \infty.$$

The choice of the family of monotone operators $\{\mathcal{L}^k\}$, $\mathcal{L}^k : K \to V'$, depends on the particular method under consideration.

Special cases of scheme (2) are:

- classical proximal point method (PPM):
  $\mathcal{F} := \mathbf{0}$ (i.e. $\mathcal{F}$ is included in $\mathcal{Q}$), $\mathcal{L}^k := \mathbf{0}$, $h(u) := \frac{1}{2}\|u\|^2$, $S := V$;
- generalized (Bregman-function-based) proximal method (GPM):
  $\mathcal{F} := \mathbf{0}$; $\mathcal{L}^k := \mathbf{0}$, h - Bregman function;
- APP:
  $\mathcal{Q} := \mathbf{0}$; $\mathcal{Q}^k := \mathbf{0}$, $h \in C^1(V)$ - strongly convex, $S := V$.
  The APP supposes additionally that the operators $\mathcal{F} - \mathcal{L}^k$ fulfills a kind of co-coercivity condition on K (see, e.g., assumption F2 in Section 5).

To give a hint at the nature of $\mathcal{L}^k$, let us consider the following variant of the APP:

$$\begin{aligned}&\text{find } u^{k+1} \in K \subset \mathbb{R}^n : \\ &\langle \mathcal{F}(u^k) + \mathcal{L}^k(u^{k+1}) - \mathcal{L}^k(u^k) + \chi_k(u^{k+1} - u^k), u - u^{k+1}\rangle \geq 0 \quad \forall\, u \in K,\end{aligned}$$

[3] I.e., h satisfies the conditions B1-B7 in Section 7.1 below, if V is a finite-dimensional space. In case V is infinite-dimensional, see [3], [25].

which arises from scheme (2) if

$$V := \mathbb{R}^n,\ \mathcal{Q}^k \equiv \mathcal{Q} := \mathbf{0},\ K^k \equiv K,\ S := \mathbb{R}^n,\ h(u) := \frac{1}{2}\|u\|^2,\ \delta_k \equiv 0.$$

In this case, depending on the choice of $\mathcal{L}^k$, we obtain:

- $\mathcal{L}^k \equiv \mathbf{0}$ - gradient projection type method;
- $\mathcal{L}^k : u \mapsto (\nabla \mathcal{F}(u^k) - \chi_k I)u$ - Newton method;
- $\mathcal{L}^k : u \mapsto \nabla \mathcal{F}(u^k)u$ - regularized Newton method;
- $\mathcal{L}^k : u \mapsto (D - \chi_k I)u$, with $D > \mathbf{0}$ - projection methods.

Also SOR, quasi-Newton methods etc. can be embedded in APP scheme (see [36], [33]).

Remark 1. The alternative $\mathcal{F} := \mathbf{0}$ or $\mathcal{Q}^k \equiv \mathbf{0}$ is not exhaustive for real numerical methods. For instance, Lions-Mercier splitting method (cf. [12], [32])

$$u^{k+1} = (\mathcal{I} + \chi^{-1}\mathcal{Q})^{-1}(\mathcal{I} - \chi^{-1}\mathcal{F})(u^k)$$

for Problem (1) with $K = V$ is a particular case of scheme (2) with $\mathcal{F} \neq \mathbf{0}$, $\mathcal{Q}^k \equiv \mathcal{Q} \neq \mathbf{0}$ and $\chi_k \equiv \chi > 0$.

Proceeding from the general framework (2) and the convergence results in [22, 25, 26, 27, 28], we revise here some ideas originally developed for the improvement of certain proximal-like methods and applications to some classes of problems. On this way, these ideas can be extended to a wide class of proximal methods as well as to the APP. For instance, using a Bregman function h with a zone $S \subset K$ in the APP, new methods with unconstrained auxiliary problems and good chances for decomposition can be constructed.

The paper is structured as follows. In Section 2 properties and applications of weak regularization are discussed. Sections 3 and 6 deal with conditions on data approximation in the subproblems. These conditions admit, in particular, the use of finite element techniques for space approximation and the use of ϵ-enlargements for operator approximations. Section 4 is devoted to multi-step regularization techniques, which we suggest especially for handling ill-posed infinite-dimensional and semi-infinite problems. The combination of the PPM and the APP is studied in Section 5. In Section 7 we analyze convergence of Bregman-function -based proximal methods in the case of non-paramonotone operators $\mathcal{Q}$ as well as the use of a weakened error tolerance criterion. Also an extension of the APP with Bregman functions is considered. The final Section 8 contains the description of the elliptic proximal regularization on the example of a parabolic variational inequality.

2 Weak Regularization

In the classical PPM as well as in the APP the regularizing functional h is supposed to be strongly convex in V.

The following example (cf. [18]), as well as numerical experiments in [35] for ill-posed control problems, in [38] for Bingham fluids and in [42] for Signorini and contact problems in elasticity, show that, under a certain reserve of monotonicity of the operator $\mathcal{Q}$, the use of a functional h with weaker convexity properties can yield an essential acceleration of the numerical process.

Note that in this example $\mathcal{Q}$ is neither strongly monotone nor even strictly monotone.

Example 1. Let $\Omega := \{(x,y) : -\frac{\pi}{2} < x,y < \frac{\pi}{2}\}$, $V := H^1(\Omega)$. VI (1) is considered with

$$K := V, \quad \mathcal{F} := \mathbf{0}, \quad \mathcal{Q} : u \mapsto -\Delta u - f,$$

f is given by $f(x,y) := 2\sin x \sin y$. Obviously this is equivalent to the Neumann problem

$$-\Delta u = 2\sin x \sin y, \qquad \frac{\partial u}{\partial n}|_{\partial\Omega} = 0,$$

whose solution set is

$$U^* := \{\sin x \sin y + d : \ d \in \mathbb{R}\}.$$

Applying the exact PPM with $u^1 := 0$, $\chi_k \equiv 1$, we obtain the sequence $\{u^k\} \subset H^2(\Omega)$, where u^{k+1} is the unique solution of the boundary value problem

$$-2\Delta u + u = 2\sin x \sin y - \Delta u^k + u^k, \qquad \frac{\partial u}{\partial n}|_{\partial\Omega} = 0. \tag{3}$$

It is not difficult to verify (by means of an immediate substitution) that

$$u^k = a_k \sin x \sin y,$$

where $(1 - a_k) = \frac{3}{5}(1 - a_{k-1})$, $a_1 = 0$. With $u^* = \sin x \sin y \in U^*$ one gets

$$u^* - u^k = \left(\frac{3}{5}\right)^{k-1} \sin x \sin y.$$

But, replacing in this method the classical regularizing functional

$$h : u \mapsto \frac{1}{2}\|u\|^2_{H^1(\Omega)} \qquad \text{by} \qquad h : u \mapsto \frac{1}{2}\|u\|^2_{L^2(\Omega)},$$

a sequence $\{v^k\} \subset H^2(\Omega)$, $v^1 := u^1 = 0$, is generated, where v^{k+1} is the unique solution of the problem

$$-\Delta v + v = 2\sin x \sin y + v^k, \qquad \frac{\partial v}{\partial n}|_{\partial\Omega} = 0. \tag{4}$$

Here $v^k = b_k \sin x \sin y$, holds, with $(1 - b_k) = \frac{1}{3}(1 - b_{k-1})$, $b_1 := a_1 = 0$. Hence,

$$u^* - v^k = \left(\frac{1}{3}\right)^{k-1} \sin x \sin y.$$

To compare numerically these solutions:

$$\frac{\|u^* - v^6\|}{\|u^* - u^6\|} \approx 0.053, \qquad \frac{\|u^* - v^{11}\|}{\|u^* - u^{11}\|} \approx 0.0028, \qquad \text{etc.}$$

Moreover, under identical finite element approximations in problems (3) and (4), the conditioning of the discretized problems for (4) is, at least, not worse.

Remark 2. For the first acquaintance, one can think of weak regularization as the use of a regularizing functional

$$h : u \mapsto \frac{1}{2}\|u\|_H^2,$$

where $H \supset V$ is a Hilbert space and $\|\cdot\|_H$ is weaker than $\|\cdot\|_V$. However, such functional h is not strongly convex in V, and the standard assumptions on VI (1), guaranteeing solvability of the regularized problems in the classical PPM, are insufficient in this case.

This insufficiency becomes clear by means of the following example.

Example 2. Let $V := H^1(0,1)$ and $\varphi \in C[0,1] \setminus H^1(0,1)$ be a given function with $1 \le \varphi(x) \le 2\ \forall x \in [0,1]$. Define the convex closed set

$$K := \{u \in H^1(0,1) : \max_{0 \le x \le 1} u(x) - \min_{0 \le x \le 1} u(x) \le 1\}$$

and the convex continuous functional

$$J : u \mapsto \int_0^1 \max^2\{0, u + \varphi\} dx,$$

and consider CP

$$\min\{J(u) : u \in K\},$$

which is equivalent to VI (1) with $\mathcal{F} := \mathbf{0}$, $\mathcal{Q} := \partial J$.

Here, the operator $\mathcal{Q}$ is maximal monotone, and the solution set U^* is non-empty ($u^* \equiv -2 \in U^*$). Applying the classical PPM, unique solvability of the subproblems and weak convergence (in $H^1(0,1)$) of the iterates to some $u^* \in U^*$ are guaranteed. But, using

$$h : u \mapsto \frac{1}{2}\|u\|_{L^2(0,1)}^2,$$

the regularized problem reads as follows

$$\min\{J(u) + \frac{\chi}{2}\|u - a\|_{L^2(0,1)}^2 : u \in K\}. \tag{5}$$

Now, we show that this problem, given for instance with $\chi := 2$ and $a := 10$, is not solvable. Indeed, it is easy to see that the problem

$$\min\{J(u) + \|u - 10\|^2_{L^2(0,1)} : \ u \in K_C\}, \tag{6}$$

where

$$K_C := \{u \in C[0,1] : \ \max_{0\le x\le 1} u(x) - \min_{0\le x\le 1} u(x) \le 1\},$$

has a unique solution

$$u_C = \frac{1}{2}(10 - \varphi) \notin H^1(0,1).$$

Due to the compactness of the embedding of $H^1(0,1)$ into $C[0,1]$, it holds

$$\begin{aligned} &\inf \ \{J(u) + \|u - 10\|^2_{L^2(0,1)} : \ u \in K\} \\ &= \min \ \{J(u) + \|u - 10\|^2_{L^2(0,1)} : \ u \in K_C\}. \end{aligned}$$

Assuming that problem (5) is solvable, let $\tilde{u}$ be its solution. From the last equation, we obtain that $\tilde{u}$ is also a solution of problem (6). Hence $\tilde{u} = u_C$ has to be, but this is impossible, because $\tilde{u} \in H^1(0,1)$ and $u_C \notin H^1(0,1)$.

The natural requirement for the use of weak regularization is to guarantee solvability of weakly regularized problems and the same quality of convergence as in related methods where a strongly convex (in V) functional h is applied.

In Example 1 we face the situation that, although $h : u \mapsto \frac{1}{2}\|u\|^2_{L^2(\Omega)}$ is not strongly convex (in V) and the operator $\mathcal{Q}$ is even not strictly monotone (in V), the sum $\mathcal{Q} + \nabla h$ is strongly monotone. This suffices, and weak regularization with similar properties can be realized for a series of elliptic VIs with semi-coercive operators (Signorini- and contact problems for elastic bodies, etc).

To be more precise, conditions on the functional h in methods with weak regularization can be described as follows. Throughout the paper, we assume that $\mathcal{B} : V \to V'$ is a linear, continuous, symmetric and monotone operator. Moreover, we suppose that for all k either

$$\mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k}$$

holds and $\mathcal{Q} - \mathcal{B}$ is monotone, or $\mathcal{Q}^k - \mathcal{B}$ is monotone, i.e. the operator $\mathcal{B}$ is a reserve of monotonicity of $\mathcal{Q}$ or $\mathcal{Q}^k$, respectively.

Then, we will say that a convex functional h provides weak regularization, if the functional

$$\eta : u \mapsto \frac{1}{2}\langle \mathcal{B}u, u\rangle + h(u) \tag{7}$$

is strongly convex and of Bregman type (with zone S).

This description includes also *regularization on a subspace* of V ([22, Section 2]).

For the GPMs, considered in [24, 25, 28], the assumption on strong convexity of η can be replaced by the weaker assumption B8 from Section 7.1 below (function D in B8 has to be taken with η in place of h).

In case $S := V$ the standard assumptions on h include also that ∇h is Lipschitz continuous. Considering $S \neq V$, this assumption contradicts to the definition of Bregman functions, and it can be avoided by means of additional requirements on $\mathcal{Q}$ or $\mathcal{F} + \mathcal{Q}$.

Remark 3. It should be noted that, as against from the classical proximal mapping, even in the case of problem (CP) and $h : u \mapsto \frac{1}{2}\|u\|_H^2$, the "weak" proximal mapping

$$a \mapsto \arg\min_{u \in K}\{J(u) + \frac{\chi}{2}\|u - a\|_H^2\}$$

is (in general) not non-expansive in V, and moreover, does not possess the Feyer property. The last property is basic in the convergence analysis of PPMs with $h(u) := \frac{1}{2}\|u\|_V^2$. Therefore, in [17, 22, 23] essentially different techniques for the convergence analysis are used.

3 Conditions on Set and Operator Approximation

In the literature, if regularization methods include an approximation of the set K, usually it is supposed that $\{K^k\}$ converges to K "sufficiently" fast in the Hausdorff or Mosco sense, for example:

$$\mathrm{dist}_H(K^k, K) \leq c\varphi_k, \qquad \sum \frac{\varphi_k}{\chi_k} < \infty.$$

However, this type of assumptions is not very realistic when dealing with VIs in mathematical physics. Indeed, constructing $\{K^k\}$, $K^k = K_{\Delta_k}$, by means of the finite element method (FEM) on a sequence of triangulations with parameter $\Delta_k \to 0$, as well as by related finite difference methods (FDM), we meet the following typical situation:

(i) for arbitrary $v \in K$ and $v^k := \arg\min_{z \in K^k} \|v - z\|$, the relation

$$\lim_{k \to \infty} \|v - v^k\| = 0 \qquad (\text{ approx. } K \text{ by } K^k)$$

takes place *without any estimate for the rate of convergence*;

(ii) for an important class of variational inequalities, the solutions possess a better regularity than arbitrary elements from K. Then, for $v \in U^*$ (solution set) it holds

$$\|v - v^k\| \leq c(v)\Delta_k^{\beta_1}, \ \beta_1 > 0 \qquad (\text{ approx. } U^* \text{ by } K^k);$$

(iii) for an arbitrary bounded sequence $\{w^k\}$, $w^k \in K^k$, the estimate

$$\min_{v \in K} \|v - w^k\| \leq c\Delta_k^{\beta_2}, \ \beta_2 > 0 \quad (\text{ approx. } K^k \text{ by } K)$$

is valid (of course, $c = 0$ fits the case $K^k \subset K$, but this inclusion is not guaranteed, in general).

Thus, because of the weak property (i), the FEM does not provide the required qualities of Hausdorff or Mosco approximations of K. Considering (i)-(iii) as conditions, together with

$$\sum \frac{\varphi_k}{\chi_k} < \infty, \text{ where } \varphi_k := \max\{\Delta_k^{\beta_1}, \Delta_k^{\beta_2}\},$$

we deal with quite different requirements on the type of approximation.

In [23], for VI (1) with $\mathcal{F} \equiv \mathbf{0}$, we study the inexact PPM with weak regularization. This method is a partial case of scheme (2), when $\mathcal{L}^k \equiv \mathbf{0}$, $S := V$ and h is a quadratic regularizing functional. Conditions on approximation of K are generalizations of the properties (i)-(iii) above.

If, in particular, $\mathcal{Q}$ is Lipschitz continuous, $\mathcal{Q}^k \equiv \mathcal{Q}$, the approximation of K possesses the properties (i)-(iii) and

$$\sum \frac{\varphi_k}{\chi_k} < \infty, \qquad \sum \frac{\delta_k}{\chi_k} < \infty, \tag{8}$$

then convergence results in [23] permit one to state immediately weak convergence of the iterates to a solution u^* of the problem. Thus, the same type of convergence as for the exact PPM is guaranteed.

It is also worth mentioning that, in a series of VIs in mathematical physics, the operator $\mathcal{Q}$ can be split up into the sum

$$\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2,$$

where $\mathcal{Q}_1$ is Lipschitz-continuous and monotone, and $\mathcal{Q}_2$ is a monotone operator of a special structure, like:

- $\mathcal{Q}_2(u) = \partial\left(\int_\Omega |\nabla u| d\Omega\right)$ – the Bingham problem
  (flow of a viscous plastic fluid in a cylindrical pipe of cross section Ω);
- $\mathcal{Q}_2(u) = \partial\left(\int_\Gamma g|u_t| d\Gamma\right)$ – the Signorini problem
  (unilateral contact of an elastic body and a rigid support: Γ - contact part of the boundary, u_t - tangential component of the displacement vector u, and $g > 0$ - given function).

If solutions of these problems are of class H^2 (see regularity results for the Bingham problem in [13] and for the Signorini problem in [14]), then the properties (i)–(iii) of finite element approximations can be satisfied. Using

$$\mathcal{Q}^k = \mathcal{Q}_1 + \mathcal{Q}_2^k,$$

where (r_k – smoothing parameter)

$$\mathcal{Q}_2^k(u) := \text{grad}\left(\int_\Omega \sqrt{|\nabla u|^2 + r_k^2}d\Omega\right),$$

respectively,

$$\mathcal{Q}_2^k(u) := \text{grad}\left(\int_\Gamma g\sqrt{u_t^2 + r_k^2}d\Gamma\right), \tag{9}$$

convergence of the above mentioned method under (8) and $r_k > 0$, $\sum \frac{r_k}{\chi_k} < \infty$ can be also easily concluded from [23].

An approximation of $\mathcal{Q}$ with the use of the ϵ-enlargement concept will be discussed below in Section 6.

4 Multi-Step Regularization (MSR)

The main approximation techniques in mathematical physics are finite element - and finite difference methods on a sequence of triangulations/grids. For their success it is very important to obtain on coarse grids a relative good approximation of the sought solution. Then, continuing on finer grids (i.e. dealing with discretized problems of larger dimension), the required exactness of the solution can be achieved with reasonable numerical expense.

However, this approach meets complications in the case of ill-posed problems:

- convergence of regularization methods is slow; if discretization is improved (in a standard manner) after each proximal step, we get in a short time a large-scale problem while the current iterate can be still far from the sought solution;
- attempts to obtain on coarse grids a better approximation of the sought solution by means of a priori given numbers of proximal steps in each fixed discretized problem are not encouraging. Indeed, doing many proximal steps in the discretized problem, we are moving to its solution (if it exists). But, the original problem is ill-posed, hence, this solution can be far from what we are looking for.

The approach of multi-step regularization (cf. [17, 22] for proximal methods with quadratic functional h possessing property (7)) is developed to reduce these complications.

Multi-step-regularization: Considering VI (1) with $\mathcal{F} := \mathbf{0}$, at the k-th outer step, we deal with the approximated problem

$$\text{find } z \in K^k,\ q^k \in \mathcal{Q}^k(z): \qquad \langle q^k, u - z\rangle \geq 0 \ \ \forall\, u \in K^k. \tag{10}$$

Starting with iterate $u^{k,0} := u^k \in K^{k-1}$, obtained at the end of the $(k-1)$-st outer step, we make inner iterations ($s = 1, 2, ...$)

$$\text{find } u^{k,s} \in K^k,\ q^{k,s} \in \mathcal{Q}^k(u^{k,s}): \tag{11}$$
$$\langle q^{k,s} + \chi_k \left(\nabla h(u^{k,s}) - \nabla h(u^{k,s-1})\right), u - u^{k,s}\rangle \geq -\delta_k \|u - u^{k,s}\| \ \ \forall\ u \in K^k.$$

Here $\{\chi_k\}$ is as before, $\delta_k \geq 0\ \forall\ k$ and $\sum \frac{\delta_k}{\chi_k} < d_0$, with given d_0. Inner iterations terminate as soon as

$$\|\nabla h(u^{k,s}) - \nabla h(u^{k,s-1})\| < \theta_k, \tag{12}$$

with a given sequence $\{\theta_k\}$. Then, we put $s := s(k)$, $u^{k+1} := u^{k,s(k)}$, refine the approximation replacing K^k, $\mathcal{Q}^k$ by K^{k+1}, $\mathcal{Q}^{k+1}$ and set $k := k + 1$.

Of course, up to now, this description is not better than saying "we do certain number of proximal steps in each problem (10)". The success of multi-step regularization depends mainly on the appropriate choice of the parameters θ_k.

To design such a choice, let us consider the case that $\mathcal{Q}^k \equiv \mathcal{Q}$ is Lipschitz-continuous,

$$K \subset B_r,\ K^k \subset B_r \qquad (B_r := \{u \in V : \|u\| \leq r\}),$$

and the approximation of K by K^k meets the conditions (i)—(iii) from Section 3 together with

$$\sum \sqrt{\varphi_k/\chi_k} < \infty, \qquad \varphi_k := \max\{\Delta_k^{\beta_1}, \Delta_k^{\beta_2}\}.$$

Then, if $h : u \mapsto \frac{1}{2}\|u\|_V^2$ is used, the parameter θ_k has to satisfy

$$\frac{1}{4r}\left(4\mu\frac{\varphi_k}{\chi_k} - (\theta_k - \delta_k)^2\right) + \delta_k < 0, \tag{13}$$

where

$$\mu \geq \sup_{u \in K} \|\mathcal{Q}(u)\|_{V'}. \tag{14}$$

For convergence results in the general case of an unbounded set K, a multi-valued operator $\mathcal{Q}$ and a weak regularization, see [22], Theorems 3.11, 3.15 and Remark 3.14.

Remark 4. However, from (13) and (14) it is clear that the use of multi-step regularization requires some additional information about the problem. The calculation of μ can be very difficult (if possible at all). In the general situation with an unbounded set K and a multi-valued operator $\mathcal{Q}$, we have also to know the radius of the ball B_r such that $U^* \cap B_{r/4} \neq \emptyset$. Relation (14) has to be replaced by

$$\mu \geq \sup_{u \in K \cap B_r} \sup_{y \in \mathcal{Q}(u)} \|y\|_{V'}.$$

To justify rule (13), it provides the strong Feyer property for the inner iterates $u^{k,1}, ..., u^{k,s(k)-1}$ w.r.t. some equivalent norm $|||\cdot|||$ of the space V, i.e. for each $u^* \in U^*$ and each k it holds

$$|||u^{k,l} - u^*||| < |||u^{k,l-1} - u^*|||, \quad l = 1, ..., s(k) - 1.$$

If, for instance, weak regularization with $h : u \mapsto \frac{1}{2}\|Pu\|_H^2$ is applied, where P is an orthoprojector on a closed subspace of V, then $|||\cdot|||$ is defined by

$$|||u|||^2 := \|\mathcal{B}u\|^2 + \|Pu\|_H^2.$$

Moreover, the rules for improving an approximation on the outer steps and the parameters δ_k do not depend on the number of inner iterations.

In [19] the convergence of MSR-methods is investigated for two elliptic variational inequalities in elasticity theory: the two body contact problem (without friction) and the Signorini problem. In both cases the approximation of K is performed by the FEM and the approximation of the multi-valued operator $\mathcal{Q}$ in the Signorini problem is like (9).

In [1] a comparison of MSR-methods and diagonal processes (with improving the discretization after each proximal step) is given for convex semi-infinite optimization problems in finance and engineering, whereas in [35] numerical experiments for the optimal control of elliptic systems are described.

5 Extension of the APP

The APP is usually studied for VIs with single-valued operators. Formally it corresponds to (1), (2) with $S := V$, $\mathcal{F} := \mathcal{F} + \mathcal{Q}$ and $\mathcal{Q}^k \equiv \mathbf{0}$.

For variational inequalities with multi-valued operators, the known results for the APP, even in case $\mathcal{L}^k \equiv \mathbf{0}$, require strong monotonicity of the operator and a very special rule for the choice of $\{\chi_k\}$ (cf. [8]):

$$\sum \chi_k^{-1} = \infty, \qquad \sum \chi_k^{-2} < \infty.$$

This is like step-size rules in subgradient methods and is not very efficient.

In our extension of the APP, described by (2), we take

- $q^k \in \mathcal{Q}^k(u^{k+1})$ (at the sought iterate),
- $\mathcal{F}(u^k)$ (at the previous iterate).

The idea to use an additive part of the operator at the sought iterate is not new in the APP. Already in the first paper on the APP, considering a problem

$$\min\{J_1(u) + J_2(u) : \ u \in K\}, \tag{15}$$

where J_1 is a convex, Gâteaux-differentiable functional and J_2 is a proper convex, lower semicontinuous functional, Cohen [7] (Algorithm 2.1) constructed auxiliary optimization problems, which are equivalent to (2) with

$$\mathcal{F} := \nabla J_1, \quad \mathcal{Q}^k \equiv \partial J_2, \quad K^k \equiv K, \quad h : u \mapsto \frac{1}{2}\|u\|_V^2, \quad \delta_k \equiv 0$$

and $\mathcal{L}^k$ a symmetric monotone operator.

In implicit form, combinations of the APP and proximal-like methods can be found in [43] (corresponds to (2) with $V := \mathbb{R}^n$, $\mathcal{Q}^k \equiv \mathcal{Q}$, $\mathcal{L}^k \equiv \mathcal{L}$, $\chi_k \equiv \chi$, $K^k \equiv K$) and [36] ($V := \mathbb{R}^n$, $K^k \subset K$, $\delta_k \equiv 0$). Both schemes do not allow that $\mathcal{Q}$ is a multi-valued, non-symmetric operator. Also h is supposed to be strongly convex.

Such an approach is of special interest for constructing decomposition methods. To avoid needless formalization, let us turn to the optimization problem (15). Suppose that

$$\min\{J_2(u) : \ u \in K\} \tag{16}$$

splits up into independent problems on the sets $K_1, ..., K_s$, $K = \Pi_{i=1}^s K_i$ is the Cartesian product. Then the auxiliary problems with $\mathcal{L}^k \equiv 0$ can be rewritten as

$$\min\{\langle \nabla J_1(u^k), u - u^k\rangle + J_2(u) + \frac{\chi_k}{2}\|u - u^k\|^2 : \ u \in K\},$$

and obviously, they can be decomposed in the same manner as (16). For VI (1) and scheme (2) such decomposition is described in [20].

Another important fact concerns VI (1) with a symmetric operator $\mathcal{Q}$: using symmetric $\mathcal{Q}^k$ and $\mathcal{L}^k$, the exact problems (2) are convex optimization problems (even if $\mathcal{F}$ is not symmetric, and hence VI (1) cannot be transformed into an optimization problem).

These properties are lost if we consider *proximal methods* and $\mathcal{F}$ is not of a special structure.

In [21, 26], the convergence of scheme (2)(with $S := V$) is considered under conditions on $\mathcal{Q}$, which are typical for PPMs. The conditions on the approximation of K by K^k and $\mathcal{Q}$ by $\mathcal{Q}^k$ are very similar to those mentioned above, i.e. again based on properties (i)–(iii) of the FEM. The choice of the regularizing functional h fulfils the requirements to methods with weak regularization, i.e. the standard assumption (for the APP) on the strong convexity of h is relaxed taking into account a reserve of monotonicity of the operator $\mathcal{Q}$.

Let $\hat{K}$ be a convex set, $\hat{K} \supset K \cup \left(\cup_{k=1}^{\infty} K^k\right)$. Typically, the operators $\mathcal{L}^k$ in the APP can be described by $\mathcal{L}^k = \mathcal{L}_{y|y=u^k}$, using an appropriate family $\{\mathcal{L}_y\}$ of monotone operators $\mathcal{L}_y : V \to V'$ parameterized by $y \in \hat{K}$. The operators $\mathcal{L}_y$ are also supposed to be Lipschitz-continuous on $\hat{K}$ with a common Lipschitz constant.

Then, the assumptions concerning the operator $\mathcal{F}$ are formulated in our papers [21, 26] as follows:

F1 $\mathcal{F} : V \to V'$ is single-valued and weakly continuous on $\hat{K}$, and the functional $u \mapsto \langle \mathcal{F}(u), u\rangle$ is weakly lower semicontinuous on $\hat{K}$;

F2 there exists $\gamma > 0$ such that, for all $u \in U^*$, $v \in \hat{K}$, the inequality

$$\langle \mathcal{F}(v) - \mathcal{F}(u) - \mathcal{L}_v(v) + \mathcal{L}_v(u), v - u\rangle + \langle \mathcal{F}(u) + q^*, z(v) - u\rangle \geq \gamma \|\mathcal{F}(v) - \mathcal{L}_v(v) - \mathcal{F}(u) + \mathcal{L}_v(u)\|_{V'}^2 \quad (17)$$

is valid. Here $q^* \in \mathcal{Q}(u)$ satisfies

$$\langle q^*, v - u\rangle \geq 0 \quad \forall\, v \in K,$$

and $z(v) := \arg\min_{y \in K} \|v - y\|$.

Obviously, assumption F2 is weaker than the rather usual requirement (for the APP) that the operators $\mathcal{F} - \mathcal{L}_y$, $y \in \hat{K}$, are co-coercive with modulus γ.

Example 3. Let $V := \mathbb{R}^1$, $\mathcal{Q}(u) := u + 4$, $K := [-1, 1]$, $\hat{K} := [-2, 2]$, $\mathcal{L}_y \equiv \mathbf{0}$ and

$$\mathcal{F}(u) := \begin{cases} u + 2 & \text{if } u < -1 \\ u^2 & \text{if } u \geq -1. \end{cases}$$

Here we meet the situation that F2 is fulfilled (with $\gamma = 1$), although the operator $\mathcal{F}$ is not co-coercive and even not pseudo-monotone (in Karamardian's sense). It is also worth to mention that in this example $\mathcal{F} + \mathcal{Q}$ is not monotone on K.

Under the assumptions traced above, the main convergence result in [26] (Theorem 3.1) establishes weak convergence of the iterates (2) to a solution $u^* \in U^*$ if

$$\frac{c}{\gamma} \leq \chi_k \leq \bar{\chi} \text{ (with some } c > 0), \qquad \sum \varphi_k < \infty, \qquad \sum \delta_k < \infty.$$

In the particular case $\mathcal{F} := \mathbf{0}$, $\mathcal{L}^k \equiv \mathbf{0}$, the values γ in (17) may be arbitrary large, and we recover the convergence results for the PPM with weak regularization mentioned in Section 3.

6 The Use of ϵ-Enlargements

Recently, a series of proximal-like methods has been suggested, in which the operator in the VI is approximated by its ϵ-enlargement, with $\epsilon \to 0$ within the iteration process (cf. [4, 5, 30, 37, 39]).

The ϵ-enlargement $\mathcal{T}_\epsilon$ of a monotone operator $\mathcal{T} : V \to V'$ is defined as

$$\mathcal{T}_\epsilon(u) = \{v \in V' : \langle w - v, z - u\rangle \geq -\epsilon, \quad \forall\, z \in V,\ w \in \mathcal{T}(z)\}.$$

For its properties, in case $\mathcal{T}$ is maximal monotone, we refer to [4].

However, the treatment of the subproblems in these methods meets serious difficulties, because the verification of the inclusion $q \in \mathcal{T}_\epsilon(u)$ for a maximal monotone operator $\mathcal{T}$, as well as the calculation of an element $q \in \mathcal{T}_\epsilon(u)$ with

certain properties, may be very complicated. Moreover, for an operator $\mathcal{T}$ with separable structure, $\mathcal{T}_\epsilon$ is usually not separable. For instance, the operator

$$\mathcal{T} : u \mapsto \partial(|u_1| + |u_2|) \equiv \partial_{u_1}(|u_1|) \times \partial_{u_2}(|u_2|)$$

is evidently separable, but calculating

$$\mathcal{T}_\epsilon(1,1) = \{u \in [-1,1] \times [-1,1] : \ u_1 + u_2 \geq 2 - \epsilon\},$$

we see that $\mathcal{T}_\epsilon$ is not separable.

This prevents, for instance, from the use of the ϵ-enlargement in decomposition methods.

If the operator $\mathcal{T}$ has a reserve of monotonicity like

$$\langle \tau(u) - \tau(v), u - v\rangle \geq \langle \mathcal{B}(u-v), u-v\rangle$$
$$\forall\, u, v \in D(\mathcal{T}),\ \forall\, \tau(u) \in \mathcal{T}(u),\ \tau(v) \in \mathcal{T}(v),$$

one cannot guarantee that

$$\langle \hat{\tau}(u) - \tau(v), u - v\rangle \geq \beta\langle \mathcal{B}(u-v), u-v\rangle - \alpha\epsilon \tag{18}$$
$$\forall\, u \in D(\mathcal{T}_\epsilon),\ v \in D(\mathcal{T}),\ \forall\, \hat{\tau}(u) \in \mathcal{T}_\epsilon(u),\ \tau(v) \in \mathcal{T}(v),$$

is valid with some α and $\beta = 1$, moreover, the existence of appropriate $\beta > 0$ and α is not clear at all. However, in order to use weak regularization, we need a relation like (18) for the operators $\mathcal{Q}^k$.

Now we describe some simple ideas to construct $\mathcal{Q}^k$, $\mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k}$, which

- inherits separability of $\mathcal{Q}$ and all continuity properties of the ϵ-enlargement;
- its treatment is simpler than that of $\mathcal{Q}_{\epsilon_k}$;
- possesses an ϵ- reserve of monotonicity in the sense of relation (18).

Suppose that a continuous operator $\hat{\mathcal{Q}} : V \to V'$ is chosen such that both $\mathcal{Q} - \hat{\mathcal{Q}}$ and $\hat{\mathcal{Q}} - \mathcal{B}$ are monotone. Then, defining

$$\mathcal{Q}^k := \hat{\mathcal{Q}} + (\mathcal{Q} - \hat{\mathcal{Q}})_{\epsilon_k}, \tag{19}$$

we get $\mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k}$, and the continuity properties of $\mathcal{Q}^k$ depend mainly on those of $(\mathcal{Q} - \hat{\mathcal{Q}})_{\epsilon_k}$. Inequality (18) is valid with

$$\mathcal{T} := \mathcal{Q}, \quad \epsilon := \epsilon_k, \quad \mathcal{T}_\epsilon := \mathcal{Q}^k, \quad \alpha = \beta = 1.$$

Of course, the use of $\hat{\mathcal{Q}} := \mathcal{B}$ is possible and leads to the same conclusions. But, as we explain in [25], the choice $\hat{\mathcal{Q}} \neq \mathcal{B}$ may be preferable for the treatment of $\mathcal{Q}^k$.

In the above mentioned Signorini- and Bingham problems, the multivalued part of the operator is of special structure: it is the subdifferential of a convex positive homogeneous functional. For these and other variational inequalities with a similar property, the following result (cf. [25], Lemmata A.1 and A.2) is helpful to handle $\mathcal{Q}^k$.

Lemma 1. *Let j be a convex, positive homogeneous, lower semicontinuous functional and $\mathcal{T} := \partial j$. Then, for each $\epsilon > 0$,*

$$\mathcal{T}_\epsilon(u) = \{y \in \mathcal{T}(0) : \ \langle y, u\rangle \geq -\epsilon + j(u)\}$$

holds, and

$$\mathcal{T}_\epsilon(u) = \partial_\epsilon j(u) \quad \forall\ u \in dom j$$

(∂_ϵ denotes the ϵ-subdifferential).

Examples on the use of this lemma, including the special case of the Bingham problem, are given in [25], too.

6.1 Application of ϵ-Enlargement in Signorini Problems

The operator of the plane Signorini problem is $\mathcal{Q} : \ u \mapsto Au - l + \partial j(u)$, where

$$\langle \mathcal{A}u, v\rangle := \int_\Omega a_{mnpq} e_{mn}(u) e_{pq}(v) d\Omega \quad \forall\ u, v \in [H^1(\Omega)]^2$$

$$j(u) := \int_{\Gamma_c} g|u_t| d\Gamma. \tag{20}$$

In this description the summation is taken (by Einstein's summation convention) over terms with repeating indices ($m, n, p, q = 1, 2$);

- $\Omega \subset \mathbb{R}^2$ is a bounded domain, Γ_c is part of the boundary of Ω;
- $g \in L_\infty(\Gamma_c)$ is a given non-negative function;
- $a_{mnpq} \in L_\infty(\Omega)$ are given functions with symmetry property

$$a_{mnpq} = a_{nmpq} = a_{pqmn},$$

and there exists a constant $c_0 > 0$ such that for all symmetric matrices $[\sigma_{m,n}]_{m,n=1,2}$

$$a_{mnpq}(x)\sigma_{mn}\sigma_{pq} \geq c_0 \sigma_{mn}\sigma_{mn} \quad \text{a.e. on} \quad \Omega; \tag{21}$$

- $e_{mn} = \frac{1}{2}\left(\frac{\partial u_m}{\partial x_n} + \frac{\partial u_n}{\partial x_m}\right)$ are components of the strain tensor;
- u_t is the tangential component of displacement u.

The Signorini problem can be described in the form of VI (1) where

$$V := [H^1(\Omega)]^2; \quad \mathcal{F} := \mathbf{0} \quad \text{and} \quad K := V.$$

Due to (21), we can choose the operator $\mathcal{B}$ as

$$\langle \mathcal{B}u, v\rangle := c_0 \int_\Omega e_{mn}(u) e_{mn}(v) d\Omega \quad \forall\ u, v \in V.$$

Then, according to the *second Korn inequality* (see [14, Section 2.2]), the functional

$$h : u \mapsto \frac{1}{2}\|u\|^2_{[L_2(\Omega)]^2}$$

realizes weak regularization.

To construct the operators $\mathcal{Q}^k := \hat{\mathcal{Q}} + (\mathcal{Q} - \hat{\mathcal{Q}})_{\epsilon_k}$, one can take $\hat{\mathcal{Q}} = \mathcal{B}$. Then we have to deal with the ϵ-enlargement of the operator

$$\mathcal{Q} - \hat{\mathcal{Q}} : \qquad u \mapsto (\mathcal{A} - \mathcal{B})u - l + \partial j(u).$$

But taking $\hat{\mathcal{Q}} : \ u \mapsto \mathcal{A}u - l$, then the both operators $\mathcal{Q} - \hat{\mathcal{Q}} = \partial j$ and $\hat{\mathcal{Q}} - \mathcal{B}$ are monotone, and the functional j satisfies Lemma 1. In this case the treatment of $(\mathcal{Q} - \hat{\mathcal{Q}})_\epsilon$ is much simpler than taking $\hat{\mathcal{Q}} = \mathcal{B}$.

In Section 3 an approximation of the operator ∂j by the single-valued operator ∇j_k was mentioned, where $j_k(u) := \int_\Gamma g\sqrt{u_t^2 + r_k^2}d\Gamma$.

Because

$$j_k(u) - r_k \int_\Gamma g d\Gamma \leq j(u) \leq j_k(u) \quad \forall\ u \in V,$$

one gets

$$\nabla j_k(u) \in \partial_{\epsilon_k} j(u) \equiv (\partial j(u))_{\epsilon_k} \quad \forall\ u \in V,$$

with $\epsilon_k = r_k \int_\Gamma g d\Gamma$.

6.2 Application of ϵ-Enlargement in Decomposition

Assume now, that $V := V_1 \times V_2$, where V_1, V_2 are closed subspaces with norms induced by $\|\cdot\|$, and their duals V_1' and V_2'. Let

$$\mathcal{Q}(v) := (\mathcal{Q}_1(v_1), \mathcal{Q}_2(v_2)), \qquad \mathcal{Q}_i : V_i \to 2^{V_i'}, \quad i = 1, 2.$$

If $\mathcal{Q}$ is maximal monotone in V, then $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are maximal monotone in V_1 and V_2, respectively. Considering their ϵ-enlargements $\mathcal{Q}_{1,\epsilon}$ and $\mathcal{Q}_{2,\epsilon}$ in V_1 and V_2, respectively, we obtain

$$\mathcal{Q} \subset (\mathcal{Q}_{1,\epsilon}, \mathcal{Q}_{2,\epsilon}) \subset \mathcal{Q}_{(2\epsilon)},$$

and the operator $\hat{\mathcal{Q}}_\epsilon := (\mathcal{Q}_{1,\epsilon}, \mathcal{Q}_{2,\epsilon})$ (in distinction from $\mathcal{Q}_\epsilon$) has the same separable structure as $\mathcal{Q}$.

Of course, the "recipes" considered above to construct $\mathcal{Q}^k$ are compatible in a straightforward manner. For instance, if $\mathcal{Q}$ has a separable structure and $\mathcal{Q} - \mathcal{B}$ is monotone, then one can construct a separable $\mathcal{Q}^k$ such that it inherits the continuity properties of $\mathcal{Q}_{\epsilon_k}$ and fulfils (18).

7 Bregman-Function-Based Methods

There are numerous publications dedicated to Bregman-function-based proximal methods in *finite dimensional* spaces. In a Hilbert space, the exact method was studied in [3], whereas in [21, 25] we prove convergence of inexact versions, including successive approximations of K, the use of ϵ-enlargements of $\mathcal{Q}$ and weak regularization as described in Sections 2 and 6.

But conditions on the approximation of K require $K^k \supset K$. This inclusion emerges when dealing with convex semi-infinite problems, but it is not valid, for instance, if K^k is obtained by applying usual discretization techniques to elliptic variational inequalities. In order to use our convergence analysis in this case, an approximation of K has to be inserted into the algorithm for solving the subproblems.

In the sequel we assume that $S \not\supset K$. This includes the main case $S \subset K$, but some additional assumptions on the operator $\mathcal{Q}$ are needed[4] - even for the exact GPM with strongly convex h.

If $\mathcal{Q}$ is not symmetric, the paramonotonicity and pseudo-monotonicity (in the sense of Brezis-Lions) of $\mathcal{Q}$ are supposed (see [3, 24, 25]). In case $V := \mathbb{R}^n$, Solodov/Svaiter [40] have shown that the pseudo-monotonicity requirement can be omitted, but their arguments are finite-dimensional in essence.

Paramonotonicity of a (monotone) operator $\mathcal{Q}$ means that the relation

$$\langle z - z', v - v' \rangle = 0 \qquad \text{with } z \in \mathcal{Q}(v),\ z' \in \mathcal{Q}(v')$$

implies $z \in \mathcal{Q}(v'),\ z' \in \mathcal{Q}(v)$.

This condition is rather restrictive, in particular, a maximal monotone operator associated with a Lagrangian of a smooth convex programming problem is paramonotone only if all constraints are not active ([15, 28]).

7.1 Bregman-Function-Based Proximal Method

In this part, we consider the application of the GPM to variational inequalities with *non-paramonotone operators* of particular type.

With this goal, let us recall the conditions defining Bregman functions in $\mathbb{R}^m$:

B1 $S \subset \mathbb{R}^m$ is a convex, open set;
B2 h is continuous and strictly convex in $\bar{S}$;
B3 h is continuously differentiable on S;
B4 Given any $z \in \bar{S}$ and a scalar α, the set $\{v \in S : D(z, v) \leq \alpha\}$ is bounded, where the distance function D is defined by

$$D(z, v) := h(z) - h(v) - \langle \nabla h(v), z - v \rangle;$$

[4] An exclusion is the proximal-like method in [9], where, however, a very unusual strategy for the choice of $\{\chi_k\}$ in the PPM is used, forcing in particular that $\chi_k \to 0$.

B5 If $\{v^k\} \subset S$ converges to v, then $D(v, v^k) \to 0$;
B6 If $\{z^k\} \subset \bar{S}$, $\{v^k\} \subset S$, $v^k \to v$, $\{z^k\}$ is bounded and $D(z^k, v^k) \to 0$, then $z^k \to v$.

From [40] it is known that condition B6 is a corollary of B1-B3.

A special assumption is required in order to guarantee the solvability of the regularized problem in S, for instance, the so-called *zone coerciveness* condition

B7 $\forall\, p \in \mathbb{R}^m, \exists\, v \in S:\ \nabla h(v) = p.$

Finally, we introduce one more assumption

B8 $\forall\, z \in \bar{S},\ \exists\, \alpha(z) > 0,\ c(z):\ D(z, v) \geq \alpha(z)\|z - v\| - c(z)\ \ \forall\, v \in S,$

which is certainly weaker than the strong convexity of h and implies B4. B8 means geometrically that $D(z, \cdot)$ lies above the translated and scaled second order cone given by the function

$$v \mapsto a(z)\|v - z\| - c(z).$$

As it will be explained later on, B8 permits one, in particular, to weaken the error tolerance criteria in Bregman-function-based proximal methods. At the same time, this condition is very mild: among the known Bregman functions only

$$h : u \mapsto \sum_{i=1}^{m} (u_i - u_i^{\kappa}), \qquad \kappa \in (0, 1),$$

does not satisfy B8. Moreover, B8 is evident if S is a bounded set.

In [28] we study the convergence of GPM for two classes of variational inequalities. The first one arises from problem

$$\min\{f(x):\ g_i(x) \leq 0,\ i = 1, ..., m\},$$

where f, $g_1, ..., g_m$ are supposed to be convex and continuous on a Hilbert space X. We assume that the set $X^* \times \Lambda^*$ of the saddle points of the Lagrangian

$$L(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

on $X \times \mathbb{R}^m_+$ is non-empty. The inclusion $(x^*, \lambda^*) \in X^* \times \Lambda^*$ is equivalent to the variational inequality

$$\text{find } (x^*, \lambda^*) \in X \times \mathbb{R}^m_+,\ q(x^*, \lambda^*) \in \partial_x L(x^*, \lambda^*):$$
$$\langle q(x^*, \lambda^*), x - x^* \rangle + \langle -g(x^*), \lambda - \lambda^* \rangle \geq 0 \quad \forall\, (x, \lambda) \in X \times \mathbb{R}^m_+, \quad (22)$$

where ∂_x denotes the partial subdifferential with respect to x and $\langle \cdot, \cdot \rangle$ stands for the duality pairing between X and X' as well as for the inner product in $\mathbb{R}^m$.

The multi-valued operator

$$\mathcal{Q} : (x, \lambda) \mapsto (\partial_x L(x, \lambda), -g(x))$$

is monotone on $X \times \mathbb{R}^m_+$, but as it was mentioned, it is not paramonotone except for "an exotic case".

To solve VI (22) we consider the following method. Let be

- $\{X^k\}$ a subsequence of closed subspaces of X approximating X;
- $h : \mathbb{R}^m_+ \to \mathbb{R}$ a separable Bregman function with zone $S := \mathbb{R}^m_{++}$ satisfying B1-B8;
- $\{\chi_k\}$, $\{\delta_k\}$, $\{\epsilon_k\}$ non-negative sequences satisfying

$$0 < \chi_k \le \bar{\chi}_k < \infty, \quad \sum \frac{\delta_k}{\chi_k} < \infty, \quad \sum \frac{\epsilon_k}{\chi_k} < \infty; \tag{23}$$

- $\mathcal{I} : X \to X'$ the canonical isometry operator;
- $\partial_{x,\epsilon} L(x, \lambda) := \partial_\epsilon f(x) + \sum_{i=1}^m \lambda_i \partial_\epsilon g_i(x)$.

Method: Starting with $(x^1, \lambda^1) \in X \times S$, at the $(k+1)$-st step the current iterate $(x^k, \lambda^k) \in X^{k-1} \times S$ is used to solve

$$\begin{aligned} &\text{find } (x^{k+1}, \lambda^{k+1}) \in X^k \times \bar{S},\ q \in \partial_{x,\epsilon_k} L(x^{k+1}, \lambda^{k+1}) : \\ &\quad \langle q + \chi_k(\mathcal{I}x^{k+1} - \mathcal{I}x^k), x - x^{k+1} \rangle \\ &\quad + \langle -g(x^{k+1}) + \chi_k(\nabla h(\lambda^{k+1}) - \nabla h(\lambda^k)), \lambda - \lambda^{k+1} \rangle \\ &\qquad \ge -\delta_k \left(\|x - x^{k+1}\|_X + \|\lambda - \lambda^{k+1}\|_{\mathbb{R}^m} \right) \quad \forall (x, \lambda) \in X^k \times \bar{S}. \end{aligned} \tag{24}$$

Conditions B1-B3 and B7, B8 provide the existence of (x^{k+1}, λ^{k+1}) satisfying (24) and the inclusion $\lambda^{k+1} \in S$. Obviously, (24) can be considered as a particular case of scheme (2) with $\mathcal{F} := \mathbf{0}$, $\mathcal{L}_k := \mathbf{0}$, the Bregman function

$$(x, \lambda) \mapsto \frac{1}{2}\|x\|^2_X + h(\lambda) \qquad \text{in place of } h,$$

and $K^k := X^k \times \mathbb{R}^m_+$, $\mathcal{Q}^k : (x, \lambda) \mapsto (\partial_{x,\epsilon_k} L(x, \lambda),\ -g(x))$. In [28], Theorem 3.2, weak convergence of the iterates (24) to some $(x^*, \lambda^*) \in X^* \times \Lambda^*$ is proved under (23) and the following conditions:

A1 for each $(x^*, \lambda^*) \in X^* \times \Lambda^*$ and each i: $\lambda^*_i = 0 \Rightarrow g_i(x^*) < 0$;
A2 $\lim_{k\to\infty} \|x - P_{X^k}(x)\|_X = 0 \ \ \forall x \in X$ (P_{X^k} – orthoprojector onto X^k);
A3 $\forall x^* \in X^*$, $\exists\ c_1(x^*)$, $\{\varphi_k\}$:

$$\|x^* - P_{X^k}(x^*)\|_X \le c_1(x^*)\varphi_k \ \ \forall\ k, \quad \sum \frac{\varphi_k}{\chi_k} < \infty.$$

The strict complementarity assumption A1 is the only additional condition on the operator $\mathcal{Q}$, whereas A2 and A3 are well matched with the properties (i)-(iii) of a finite element approximation.

The convergence analysis uses essentially the following modification of Proposition 4 in Iusem/Kallio [16] (which was proved for a VI with single-valued, monotone and continuous operator):

Lemma 2. *Suppose that VI (22) satisfies A1 and* $(\tilde{x}, \tilde{\lambda}) \in X \times \mathbb{R}^m_+$ *fulfills*

$$0 \in \partial_x L(\tilde{x}, \tilde{\lambda}), \quad g_i(\tilde{x})\tilde{\lambda}_i = 0, \quad i = 1, ..., m.$$

Moreover, let

$$J(\tilde{\lambda}) \subset J(\lambda^*) \quad \textit{for some} \quad \lambda^* \in \Lambda^*,$$

where $J(\lambda) := \{i : \lambda_i = 0\}$. *Then* $(\tilde{x}, \tilde{\lambda}) \in X^* \times \Lambda^*$.

The second class of VIs considered in [28] is the complementarity problem

$$\text{find } u^* \in \mathbb{R}^m_+, \; q \in \mathcal{Q}(u^*): \qquad \langle q, u - u^* \rangle \geq 0 \quad \forall\, u \in \mathbb{R}^m_+. \tag{25}$$

Here $\mathcal{Q} : \mathbb{R}^m \to 2^{\mathbb{R}^m}$ is a monotone, upper semicontinuous operator with nonempty, convex and compact image $\mathcal{Q}(u)$ for $u \in \mathbb{R}^m_+$. Thus, the sum $\mathcal{Q} + \mathcal{N}_{\mathbb{R}^m_+}$, where $\mathcal{N}_{\mathbb{R}^m_+}$ denotes the normality operator for $\mathbb{R}^m_+$, is maximal monotone (see, for instance, [11]).

We take in this case

$$\mathcal{Q}^k := \mathcal{Q}_{\epsilon_k} \qquad \text{or} \qquad \mathcal{Q}^k := (\mathcal{Q} + \mathcal{N}_{\mathbb{R}^m_+})_{\epsilon_k}$$

and a separable Bregman function h with zone $S := \mathbb{R}^m_{++}$, satisfying B1-B8.

The GPM

$$\begin{aligned} &\text{find } u^{k+1} \in \bar{S}, \; q^k \in \mathcal{Q}^k(u^{k+1}): \\ &\langle q^k + \chi_k \left(\nabla h(u^{k+1}) - \nabla h(u^k)\right), u - u^{k+1} \rangle \geq -\delta_k \|u - u^{k+1}\| \;\; \forall\, u \in \bar{S} \end{aligned} \tag{26}$$

is studied under the same conditions on $\{\chi_k\}$, $\{\epsilon_k\}$ and $\{\delta_k\}$ as above.

Convergence of the iterates (26) to some $u \in U^*$ is proved in [28] assuming a modified strict complementarity condition:

$$u \in U^* : \; u_i = 0 \;\Rightarrow\; \tau_i(u) > 0, \tag{27}$$

where $\tau_i(u) := \min_{q \in \mathcal{Q}(u)} q_i$. If $\mathcal{Q}$ is single-valued, (27) coincides with the standard notion of strict complementarity.

The convergence results from [28] can be trivially extended to method (24), where the operator $\partial_{x,\epsilon_k} L$ is replaced by $\mathcal{Q}^k_x$ with $\partial_x L \subset \mathcal{Q}^k_x \subset \partial_{x,\epsilon_k} L$, as well as to method (26) with $\mathcal{Q}^k$ satisfying $\mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k}$. All recipes from Section 6 can be used to construct $\mathcal{Q}^k$.

7.2 Extended APP with Bregman Functions

To our knowledge, Bregman functions h with zone $S \neq V$ have not been used in connection with the APP. In different variants of the APP, the operator ∇h is supposed to be Lipschitz continuous on K or on some set $\hat{K} \supset K$. This excludes the use of Bregman-like functions with zone $S \not\supset K$, and in particular with $S \subset K$. Exactly, Bregman functions with zone $S \subset K$ provide a full

"interior point effect", i.e. with a certain precaution the auxiliary problems can be treated as unconstrained ones.

Now, we consider scheme (2) allowing

$$S \subset K, \quad \text{in particular} \quad K := K_1 \cap K_2, \; S := \text{int} K_1$$

as special cases. The convergence of the extended APP in form (2), where h is a Bregman function with a zone S, was proved in [27] under the following special assumptions:

- $V := \mathbb{R}^m, \quad S \cap K \cap D(\mathcal{Q}) \neq \emptyset, \quad U^* \cap \bar{S} \neq \emptyset;$
- h is strongly convex with modulus κ;
- $K^k \equiv K, \quad \mathcal{Q} \subset \mathcal{Q}^k \subset \mathcal{Q}_{\epsilon_k}$;
- $\mathcal{F} + \mathcal{Q}$ is paramonotone in $\bar{S} \cap K$, and $v^k \in D(\mathcal{Q}) \cap K \cap S$, $v^k \to \bar{v}$, $q(v^k) \in \mathcal{Q}(v^k)$ implies that $\{q(v^k)\}$ is a bounded sequence;
- the operators $\mathcal{F} - \mathcal{L}^k$ satisfy the co-coercivity condition F2 in Section 5 (with $\hat{K} = K$);
- the operators $\mathcal{L}^k$ are Lipschitz-continuous with a common constant;
- $\frac{1}{4\gamma\kappa} < \chi_k < \bar{\chi} < \infty$, $\sum \max\{0, \chi_k - \chi_{k+1}\} < \infty$, $\sum \epsilon_k < \infty$, $\sum \delta_k < \infty$ (γ is the modulus of co-coercivity in F2).

The other conditions are quite traditional for the exact PPM and concern the operator $\mathcal{Q}$ and the set K only.

Basing on the proof of Lemma 3.2 in [26], the convergence analysis in [27] can be easily adapted to method (2) with weak regularization, when h is defined by (7).

7.3 Weakened Error Tolerance Criteria in Proximal Methods

The criterion for the approximate calculation of the iterates used in the general scheme (2) is not suitable for a straightforward application, but it permits one to apply the convergence results obtained in [22, 23, 27] to related algorithms with more practicable criteria.

In [10] Eckstein has analyzed different accuracy conditions for iterates in Bregman-function-based proximal methods.

For VI (1) with $V := \mathbb{R}^m$ and $\mathcal{F} := \mathbf{0}$, the inclusion

$$0 \in \chi_k^{-1} \mathcal{Q}(u^{k+1}) + \nabla h(u^{k+1}) - \nabla h(u^k) + e^{k+1} \tag{28}$$

is studied, where h is a Bregman function with zone $S = \text{int} K$ and $0 < \chi_k \leq \bar{\chi} < \infty$. Convergence $u^k \to u \in U^*$ has been established under standard assumptions on $\mathcal{Q}$ (paramonotonicity, etc.; see A1-A3 in [10, Section 3]) and the conditions

$$\sum_{k=1}^{\infty} \|e^k\| < \infty, \qquad \sum_{k=1}^{\infty} \langle e^k, u^k \rangle < \infty. \tag{29}$$

Eckstein explains that the relations (28), (29) are easier to check than in other inexact schemes.

At the same time, using an element $q \in \mathcal{Q}(u^{k+1})$, which transforms (28) into an equality, we obviously get

$$\langle q + \chi_k \left(\nabla h(u^{k+1}) - \nabla h(u^k)\right), u - u^{k+1}\rangle \geq -\chi_k \|e^{k+1}\| \|u - u^{k+1}\| \quad \forall\, u \in K.$$

But this corresponds to scheme (2) with $\mathcal{Q}^k \equiv \mathcal{Q}$ and $\delta_k := \chi_k \|e^{k+1}\|$. Due to the additional assumption B8 on Bregman function, our convergence analysis in [29] ensures convergence of the iterates (28) exclusively under

$$\sum_{k=1}^{\infty} \|e^k\| < \infty.$$

Remark 5. Let us recall that B8 does not cause any complication for the choice of an appropriate Bregman function. Moreover, in [29] the validity of condition B8 is proved for entropy-like and logarithmic-quadratic distance functions which are not of Bregman type. We also show there that the GPM in form (28) with these distance functions converges under the weaker error criterion $\sum \|e^k\| < \infty$, than in former papers [2, 41]. Recently, a new concept of a relative error criterion has been introduced in [6, 40]. The corresponding algorithms include a correction of the inexact proximal iteration by means of an extragradient-like step.

8 Elliptic Proximal Regularization

Starting with papers of Lions [31] and Olejnik [34], elliptic regularization is a useful tool for the theoretical and numerical treatment of parabolic and degenerate elliptic boundary value problems. The main idea is that the original problem is approximated by a family of non-degenerated elliptic problems. This was done exclusively by means of the Browder-Tikhonov regularization concept. On this way the regularization parameter tends to zero causing a certain instability of the real numerical process.

In [23], we have introduced elliptic regularization by following the scheme of proximal methods. Now, this approach will be considered for a *parabolic variational inequality.*

Let $\Omega \subset \mathbb{R}^m$ be an open set with a sufficiently smooth boundary and

$$Z := L^2(0,T; H_0^1(\Omega)), \qquad H := L^2(0,T; L^2(\Omega)),$$

where $L^2(0,T;W)$ denotes the space of measurable on $]0,T[$ functions $v : t \mapsto v(t) \in W$ with

$$\|v\|_{L^2(0,T;W)} = \left(\int_0^T \|v(t)\|_W^2 dt\right)^{1/2} < \infty;$$

$Z' := L^2(0,T;H^{-1}(\Omega))$ denotes the dual space of Z, and H is identified with its dual.

Introducing the linear and unbounded (in Z) operator $\Lambda := \frac{d}{dt}$ with the domain

$$D(\Lambda) := \left\{ v \in Z : \frac{dv}{dt} \in Z',\ v(0) = 0 \right\},$$

one can consider $D(\Lambda)$ as a Hilbert space endowed with the graph-norm

$$\|v\|_{D(\Lambda)} := \left(\|v\|^2 + \|\Lambda v\|^2_{Z'} \right)^{1/2}.$$

Then

$$D(\Lambda) \subset Z \subset H \subset Z' \subset D(\Lambda)',$$

and each space is dense in the next one.

Take now $V := D(\Lambda)$ and let $\mathcal{A} : H_0^1(\Omega) \to H^{-1}(\Omega)$ be an elliptic (with constant α) operator. Let the operator $A : V \to V'$ be a linear, bounded and monotone operator defined by

$$(Av)(t) := \mathcal{A}v(t) \quad \text{a.e. on } [0,T],$$

with $D(A) = V$. Given $f \in V'$ and a convex, closed set $K \subset D(\Lambda)$, we consider the VI

$$\text{find } u \in K : \quad \langle Au, v-u \rangle + \langle \Lambda u, v-u \rangle \geq \langle f, v-u \rangle \quad \forall\, v \in K. \tag{30}$$

The operator Λ is closed and $\Lambda \geq 0$ on V. Therefore, Λ is maximal monotone, and because $D(A) = V = D(\Lambda)$, the sum $A + \Lambda$ is maximal monotone, too. Further we suppose that VI (30) is solvable.

We make use of the regularizing operator $\Lambda^* J^{-1} \Lambda$, where $\Lambda^* = -\frac{d}{dt}$ with

$$D(\Lambda^*) = \left\{ v \in Z : \frac{dv}{dt} \in Z', v(T) = 0 \right\}$$

is the conjugate operator of Λ, and $J : v \to -\Delta_x v$ is the duality mapping between Z and Z' (Δ_x denotes the Laplacian w.r.t. space variables x).

Basing on the principle of proximal regularization, the following iteration process for VI (30) is studied in [23]:

$$\begin{aligned} &\text{find } u^{k+1} \in K : \\ &\langle Au^{k+1} + \Lambda u^{k+1} + \chi_k \Lambda^* J^{-1} \Lambda (u^{k+1} - u^k), v - u^{k+1} \rangle \\ &\qquad \geq \langle f, v - u^{k+1} \rangle \quad \forall\, v \in K \end{aligned} \tag{31}$$

$(0 < \chi_k \leq \bar{\chi} < \infty)$.

It corresponds to scheme (2) with

$$\mathcal{F} := \mathbf{0},\ \mathcal{L}^k \equiv \mathbf{0},\ K^k \equiv K,\ \delta_k \equiv 0,$$

$$\mathcal{Q}^k \equiv \mathcal{Q} : v \mapsto Av + \Lambda v - f,$$
$$h : v \mapsto \frac{1}{2}\langle \Lambda v, J^{-1}\Lambda v\rangle.$$

The operator

$$v \mapsto (A + \Lambda + \chi_k \Lambda^* J^{-1}\Lambda)v - f$$

is elliptic in $\Omega\times]0,T[$ (i.e. with respect to space and time variables), and

$$\langle A(u-v), u-v\rangle + \langle \Lambda^* J^{-1}\Lambda(u-v), u-v\rangle \geq \min\{\alpha, 1\}\|u-v\|^2$$

holds for $u, v \in V$.

The last inequality allows to conclude that the choice of h satisfies the principle of weak regularization. The weak convergence (in V) of the iterates (31) to a solution of VI (30) follows immediately from the convergence analysis in [23].

If $\mathcal{A}$ is a *degenerate* elliptic operator, a similar result can be attained by using

$$h : v \mapsto \frac{1}{2}\langle \Lambda v, J^{-1}\Lambda v\rangle + \frac{1}{2}\langle -\Delta_x v, v\rangle.$$

References

1. L. Abbe. A logarithmic barrier approach and its regularization applied to convex semi-infinite programming problems. PhD Thesis, University of Trier, 2001.
2. A. Auslender, M. Teboulle, and S. Ben-Tiba. A logarithmic-quadratic proximal method for variational inequalities. Computational Optimization and Applications 12: 31-40, 1999.
3. R. Burachik and A. Iusem. A generalized proximal point algorithm for the variational inequality problem in Hilbert space. SIAM J. Optim., 8: 197-216, 1998.
4. R. Burachik, A. Iusem, and B. Svaiter. Enlargements of maximal monotone operators with application to variational inequalities. Set-Valued Analysis 5: 159-180, 1997.
5. R. Burachik, C. Sagastizábal, B. Svaiter. Bundle methods for maximal monotone operators. In: Thera M, Tichatschke R (eds) Ill-posed Variational Problems and Regularization Techniques. LNEMS 477, Springer, pp. 49–64, 1999.
6. R. Burachik and B. Svaiter. A relative error tolerance for a family of generalized proximal point methods. Math. Oper. Res. 26: 816–831, 2001.
7. G. Cohen. Auxiliary problem principle and decomposition of optimization problems. JOTA 32: 277–305, 1980.
8. G. Cohen. Auxiliary problem principle extended to variational inequalities. JOTA 59: 325–333, 1998.
9. P. da Silva, J. Eckstein, and C. Humes. Rescaling and stepsize selection in proximal methods using separable generalized distances. SIAM J. Optim. 12: 238–261, 2001.
10. J. Eckstein. Approximate iterations in Bregman-function-based proximal algorithms. Math. Programming 83: 113–123, 1998.

11. F. Facchinei and J.S. Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer, New York, 2003.
12. D. Gabay. Applications of the method of multipliers to variational inequalities. In: Fortin M, Glowinski R (eds) Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems. North-Holland, Amsterdam, pp. 299–331, 1983.
13. R. Glowinski. Numerical Methods for Nonlinear Variational Problems. Springer, New York, 1984.
14. I. Hlaváček, J. Haslinger, I. Nečas, and J. Lovišek. Numerical Solution of Variational Inequalities. Springer, Berlin, 1988.
15. A. Iusem. On some properties of paramonotone operators. J. Conv. Analysis 5: 269–278, 1998.
16. A. Iusem and M. Kallio. An interior point method for constrained saddle point problems. Computational and Applied Mathematics (to appear)
17. A. Kaplan and R. Tichatschke. Stable Methods for Ill-Posed Variational Problems - Prox-Regularization of Elliptic Variational Inequalities and Semi-Infinite Optimization Problems. Akademie Verlag, Berlin, 1994.
18. A. Kaplan and R. Tichatschke. Proximal point methods in examples. Forschungsbericht 96-20, Mathematik/Informatik, Universität Trier, 1996.
19. A. Kaplan and R. Tichatschke. Prox-regularization and solution of ill-posed elliptic variational inequalities. Applications of Mathematics 42: 111–145, 1997.
20. A. Kaplan and R. Tichatschke. Auxiliary problem principle and proximal point methods. J. Global Optimization 17: 201–224, 2000.
21. A. Kaplan and R. Tichatschke. Auxiliary problem principle and the approximation of variational inequalities with non-symmetric multi-valued operators. CMS Conference Proc. 27: 185–209, 2000.
22. A. Kaplan and R. Tichatschke. Proximal point approach and approximation of variational inequalities. SIAM J. Control Optim. 39: 1136–1159, 2000.
23. A. Kaplan and R. Tichatschke. A general view on proximal point methods to variational inequalities in Hilbert spaces - iterative regularization and approximation. J. Nonlinear and Convex Analysis 2: 305–332, 2001.
24. A. Kaplan and R. Tichatschke. Convergence analysis of non-quadratic proximal methods for variational inequalities in Hilbert spaces. J. Global Optimization 22: 119–136, 2002.
25. A. Kaplan and R. Tichatschke. Proximal-based regularization methods and a successive approximation of variational inequalities in Hilbert spaces. Control and Cybernetics 31: 521–544, 2002.
26. A. Kaplan and R. Tichatschke. Extended auxiliary problem principle to variational inequalities with multi-valued operators. Optimization 53: 223–252, 2004.
27. A. Kaplan and R. Tichatschke. Extended auxiliary problem principle using Bregman distances. Optimization 53: 603–623, 2004.
28. A. Kaplan and R. Tichatschke. Interior proximal method for variational inequalities: Case of non-paramonotone operators. Set-Valued Analysis 12: 357–382, 2004
29. A. Kaplan and R. Tichatschke. On inexact generalized proximal methods with a weakened error tolerance criterion. Optimization 53: 3–17, 2004.
30. K. Kiwiel. A projection-proximal bundle method for convex nondifferentiable minimization. In: M. Thera and R. Tichatschke (eds), Ill-posed Variational Problems and Regularization Techniques , LNEMS 477, Springer, pp. 137–150, 1999.

31. J.L. Lions. Equations différentielles opérationelles dans les espaces de Hilbert. C.I.M.E. Varenna, Juillet, 1963.
32. P.L. Lions and B. Mercier. Splitting algorithm for the sum of two nonlinear operators. SIAM J. Numer. Anal. 16: 964–979, 1979.
33. S. Makler-Scheimberg, V.N. Nguyen, and J.J. Strodiot. Family of perturbation methods for variational inequalities. JOTA 89: 423–452, 1996.
34. O. Olejnik. On second order linear equations with nonnegative characteristic form. Matem. Sb. 69: 111–140, 1966.
35. S. Rotin. Convergence of the proximal-point method for ill-posed control problems with partial differential equations. PhD Thesis, University of Trier, 1999.
36. G. Salmon, V.H. Nguyen, and J.J. Strodiot. Coupling the auxiliary problem principle and the epiconvergence theory for solving general variational inequalities. JOTA 104: 629–657, 2000.
37. G. Salmon, V.H. Nguyen, and J.J. Strodiot. A perurbed and inexact version of the auxiliary problem method for solving general variational inequalities with a multivalued operator. In: V.H. Nguyen, J.J. Strodiot, and P. Tossings (eds) Optimization, LNEMS 481, Springer, Berlin, pp. 396–418, 2000.
38. H. Schmitt. Normschwächere Prox-Regularisierungen. PhD Thesis, University of Trier, 1996.
39. M. Solodov and B. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. Set-Valued Analysis 7: 323–345, 1999.
40. M. Solodov and B. Svaiter. An inexact hybrid generalized proximal point algorithm and some new results on the theory of Bregman functions. Math. Oper. Res. 25: 214–230, 2000.
41. M. Teboulle. Convergence of proximal-like algorithms. SIAM J. Optim. 7: 1069–1083, 1997.
42. T. Voetmann. Numerical solution of variational inequalities using weak regularization with Bregman functions. PhD Thesis, University of Trier, 2005, to appear.
43. D.L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. SIAM J. Optimization 6: 714–726, 1996.

Application of the Proximal Point Method to a System of Extended Primal-Dual Equilibrium Problems

Igor V. Konnov[1]

Department of Applied Mathematics, Kazan University, Kazan, Russia.
ikonnov@ksu.ru

Summary. We consider a general system of equilibrium type problems which can be viewed as an extension of Lagrangean primal-dual equilibrium problems. We propose to solve the system by an inexact proximal point method, which converges to a solution under monotonicity assumptions. In order to make the method implementable, we suggest to make use of a dual descent algorithm and utilize gap functions for ensuring satisfactory accuracy of certain auxiliary problems. Some examples of applications are also given.

1 Introduction

Let Ω be a nonempty, closed and convex set in a real l-dimensional space R^l and let $\Psi : \Omega \times \Omega \to R$ be an equilibrium bifunction, i.e., $\Psi(z, z) = 0$ for each $z \in \Omega$. Then one can define the general *equilibrium problem* (EP for short) which is to find an element $z^* \in \Omega$ such that

$$\Psi(z^*, z) \geq 0 \quad \forall z \in \Omega. \tag{1}$$

EP is known to represent a very common and suitable format for various problems arising in Mathematical Physics, Economics, Operations Research and many other fields; moreover, it is closely related with other well-known general problems in Nonlinear Analysis, such as fixed point, saddle point, variational inequality, complementarity, and optimization problems; see e.g. [1, 3, 6] and references therein. For instance, if we set

$$\Psi(z', z) = \langle T(z'), z - z' \rangle,$$

where $T : \Omega \to R^l$ is a given mapping, then EP (1) reduces to the variational inequality problem.

Recently, a system of extended primal-dual variational inequalities was considered in [12, 13] and it was shown that various economic equilibrium

problems, transportation equilibrium problems and some classes of inverse optimization problems can be reduced to such a system. In [12, 13], several dual type algorithms were suggested for solving this system under strong monotonicity assumptions on the cost mappings. At the same time, an inverse EP, which is close to a system of primal-dual variational inequalities, was also considered in [11]. In this paper, we shall consider a new general problem, which involves the problems discussed in [11]-[13]. Namely, the *system of extended primal-dual equilibrium problems* (SEP for short) is the problem of finding a pair $(x^*, y^*) \in X \times Y$ such that

$$\begin{cases} \Phi(x^*, x) + \langle y^*, H(x) - H(x^*) \rangle \geq 0 & \forall x \in X, \\ \Gamma(y^*, y) - \langle H(x^*), y - y^* \rangle \geq 0 & \forall y \in Y; \end{cases} \tag{2}$$

where X and Y are nonempty convex subsets of R^n and R^m_+, respectively; $\Phi : X \times X \to R$ and $\Gamma : R^m_+ \times R^m_+ \to R$ are equilibrium bifunctions, such that $\Phi(x, \cdot)$ and $\Gamma(y, \cdot)$ are convex for all $x \in X$ and $y \in R^m_+$; and $H : X \to R^m$ is a mapping with convex components $H_i : X \to R$ for $i = 1, \ldots, m$. It is clear that the particular case of SEP (2) with

$$\Gamma(y^*, y) = \langle b, y - y^* \rangle, \quad Y = R^m_+$$

is nothing but an analogue of the Karush-Kuhn-Tucker optimality conditions for the following constrained EP: Find $x^* \in K$ such that

$$\Phi(x^*, x) \geq 0 \quad \forall x \in K$$

where $K = \{x \in X \mid H(x) \leq b\}$. Next, if we set

$$\Phi(x', x) = \langle F(x'), x - x' \rangle \quad \text{and} \quad \Gamma(y', y) = \langle G(y'), y - y' \rangle$$

in (2), where $F : X \to R^n$ and $G : R^m_+ \to R^m$ are given mappings, then SEP (2) reduces to the system of extended primal-dual variational inequalities from [12, 13]. At the same time, SEP (2) falls into the format of EP (1) if we set $z = (x, y)$, $z' = (x', y')$, $l = n + m$,

$$\Psi(z', z) = \Phi(x', x) + \langle y', H(x) - H(x') \rangle + \Gamma(y', y) - \langle H(x'), y - y' \rangle, \tag{3}$$

$$\Omega = X \times Y. \tag{4}$$

Clearly, SEP (2) allows one to model and investigate very broad classes of problems, since just EP gives the most general formulation for different types of equilibria. Moreover, the systems in [12, 13] were considered under strong monotonicity assumptions on either F or G, whereas many problems arising in applications can provide only monotonicity properties. Motivated by these facts, we intend to consider SEP (2) under monotonicity alone and suggest an inexact version of the proximal point method for this system. We also present an approach to make this method implementable for SEP (2). It is based upon an application of a dual descent algorithm and utilization of gap functions for ensuring satisfactory accuracy of certain auxiliary problems. In addition, we describe several examples of equilibrium type problems which can be solved by the method suggested.

2 Proximal Point Method

In this section, we describe an application of the proximal point method for SEP (2) and present its convergence result. First we recall some definitions of monotonicity properties for mappings and bifunctions; see e.g. [3, 10, 17].

Let V be a nonempty convex subset of a finite-dimensional space E. A mapping $Q : V \to E$ is said to be

(i) *monotone*, if, for all $u, u' \in V$, we have $\langle Q(u) - Q(u'), u - u' \rangle \geq 0$;

(ii) *strongly monotone* with modulus $\alpha > 0$, if, for all $u, u' \in V$, we have

$$\langle Q(u) - Q(u'), u - u' \rangle \geq \alpha \|u - u'\|^2 ;$$

(iii) *co-coercive* with modulus $\beta > 0$, if, for all $u, u' \in V$, we have

$$\langle Q(u) - Q(u'), u - u' \rangle \geq \beta \|Q(u) - Q(u')\|^2 .$$

We recall that a mapping $Q : V \to E$ is said to be locally Lipschitz continuous if it is Lipschitz continuous on each bounded subset of V. Similarly, we say that a mapping $Q : V \to E$ is locally co-coercive if it is co-coercive on each bounded subset of V.

Next, an equilibrium bifunction $h : V \times V \to R$ is said to be

(i) *monotone*, if, for all $u, u' \in V$, we have $h(u, u') + h(u', u) \leq 0$;

(ii) *strongly monotone* with modulus $\alpha > 0$, if, for all $u, u' \in V$, we have

$$h(u, u') + h(u', u) \leq -\alpha \|u - u'\|^2 .$$

In the case where $h(u, v) = \langle Q(u), v - u \rangle$, the bifunction h is monotone (strongly monotone with modulus α) if and only if so is Q.

We now introduce the blanket assumptions of this paper:

(A1) X is a nonempty, convex and closed subset of R^n, Y is a nonempty, convex and closed subset of R^m_+.

(A2) $\Phi : X \times X \to R$ is a continuous and monotone equilibrium bifunction, such that $\Phi(x, \cdot)$ is convex for each $x \in X$.

(A3) $H : \tilde{X} \to R^m$ is a continuous mapping on an open set $\tilde{X}$ containing X, such that each component $H_i : \tilde{X} \to R$ is convex for $i = 1, \ldots, m$.

(A4) $\Gamma : R^m_+ \times R^m_+ \to R$ is a monotone and continuous equilibrium bifunction, such that $\Gamma(y, \cdot)$ is convex for each $y \in R^m_+$.

So, we do not impose the strong monotonicity assumptions on the cost bifunctions Φ and Γ. We first construct a convergent iterative sequence by the inexact *proximal point method* (PPM for short) which was suggested for EPs in [14] and can be written as follows.

(PPM) Choose a point $z^0 = (x^0, y^0) \in X \times Y$, a number $\theta > 0$, and a non-negative sequence $\{\varepsilon_k\}$ such that

$$\sum_{k=0}^{\infty} \varepsilon_k < \infty. \tag{5}$$

For each number $k = 1, 2, \ldots$, we have a point $z^{k-1} = (x^{k-1}, y^{k-1})$, and we compute a point $z^k = (x^k, y^k) \in X \times Y$ such that

$$\left\| z^k - w^k \right\| \leq \varepsilon_k, \tag{6}$$

where $w^k = (u^k, v^k) \in X \times Y$ is a solution of the following auxiliary SEP:

$$\begin{cases} \Phi(u^k, x) + \theta \langle u^k - x^{k-1}, x - u^k \rangle + \langle v^k, H(x) - H(u^k) \rangle \geq 0 & \forall x \in X, \\ \quad \Gamma(v^k, y) + \theta \langle v^k - y^{k-1}, y - v^k \rangle - \langle H(u^k), y - v^k \rangle \geq 0 & \forall y \in Y. \end{cases} \tag{7}$$

The kth iteration has been completed.

Thus, each iterate z^k is an approximation of the exact solution w^k of the auxiliary SEP (7) with the accuracy ε_k. The convergence result for PPM can be formulated as follows.

Theorem 1. *Let (A1)-(A4) hold and let SEP (2) be solvable. Then PPM generates an iteration sequence $\{z^k\}$ which is well-defined and converges to a solution of SEP (2).*

Proof. By definition, SEP (2) is equivalent to EP (1), where Ψ and Ω are defined in (3) and (4), respectively. Next, the auxiliary SEP (7) can be equivalently rewritten as follows: Find $w^k \in \Omega$ such that

$$\Psi(w^k, z) + \theta \langle w^k - z^{k-1}, z - w^k \rangle \geq 0 \quad \forall z \in \Omega. \tag{8}$$

Due to (A2)-(A4), $\Psi : \Omega \times \Omega \to R$ is a continuous equilibrium bifunction, moreover, $\Psi(z', \cdot)$ is convex for each $z' \in \Omega$. Next, for all $z, z' \in \Omega$, we have

$$\begin{aligned} &\Psi(z', z) + \Psi(z, z') \\ &= \Phi(x', x) + \langle y', H(x) - H(x') \rangle + \Gamma(y', y) - \langle H(x'), y - y' \rangle \\ &+ \Phi(x, x') + \langle y, H(x') - H(x) \rangle + \Gamma(y, y') - \langle H(x), y' - y \rangle \\ &= [\Phi(x', x) + \Phi(x, x')] + [\Gamma(y', y) + \Gamma(y, y')] \\ &+ \langle y' - y, H(x) - H(x') \rangle - \langle H(x') - H(x), y - y' \rangle \\ &= [\Phi(x', x) + \Phi(x, x')] + [\Gamma(y', y) + \Gamma(y, y')] \leq 0, \end{aligned}$$

since Φ and Γ are monotone. It means that Ψ is also monotone and that the cost bifunction in (8) is strongly monotone. Hence, EP (8) has a unique solution (cf. [10, Prop. 2.1.16]). We see that all the assumptions of Theorem 2.1 in [14] are satisfied, therefore, $\{z^k\}$ converges to a point $z^* = (x^*, y^*) \in \Omega$, which is a solution of EP (1), (3), (4), or equivalently, of SEP (2). □

Remark 1. In [14], the convergence result for the general PPM was established under more general conditions. Namely, if each auxiliary EP (8) is solvable,

and the solution set of EP (1) coincides with that of the dual EP: Find $z^* \in \Omega$ such that

$$\Psi(z, z^*) \leq 0 \quad \forall z \in \Omega,$$

then PPM generates a sequence which is convergent to a solution of the initial EP (1). The latter condition is satisfied under generalized monotonicity assumptions on Ψ. We now use the monotonicity assumptions for Φ and Γ for the sake of simplicity of exposition. Note that various versions of PPM for generalized monotone variational inequalities were investigated in [2, 4]. Also, in [8], a splitting type method, which contains PPM as a particular case, was presented for monotone variational inequalities. At the same time, its convergence result remains valid under the more general assumption that the solution set of the initial and the dual problems coincide.

Thus, together with the convergence property, we have shown that SEP (7) has always a unique solution. However, the main problem for every PPM is to indicate a way for its implementation. In fact, finding even an approximate solution to the auxiliary problem (7) is not a trivial task. We will present an approach to make this method implementable, which is based upon application of a dual descent algorithm and of gap functions for ensuring satisfactory accuracy of solution of auxiliary problems.

3 Dual Auxiliary Procedure

In this section, we describe a dual auxiliary procedure for finding a solution of SEP (7); or equivalently, EP (8), (3), (4). First we define an auxiliary mapping $F_k : R^m_+ \to R^m$ as follows:

$$F_k(y) = -H(\tilde{u}),$$

where $\tilde{u} = X_k(y) \in X$ is determined as the unique solution to the auxiliary EP:

$$\Phi(\tilde{u}, x) + \theta\langle \tilde{u} - x^{k-1}, x - \tilde{u}\rangle + \langle y, H(x) - H(\tilde{u})\rangle \geq 0 \quad \forall x \in X.$$

Under (A1)-(A3), the mapping F_k is well-defined and single-valued since the cost bifunction in the above EP is strongly monotone and continuous (cf. [10, Prop. 2.1.16]. Some monotonicity and continuity properties of the so defined mapping F_k are given in the next lemma.

Lemma 1. *Let (A1)-(A3) hold. Then,*
(i) the mapping $X_k : R^m_+ \to R^n$ has nonempty values on R^n_+ and

$$\langle y - y', F_k(y) - F_k(y')\rangle \geq \theta \left\| X_k(y) - X_k(y')\right\|^2 \quad \forall y, y' \in R^m_+; \qquad (9)$$

(ii) the mapping $F_k : R^m_+ \to R^m$ is locally Lipschitz continuous and locally co-coercive.

Proof. By assumption, X_k is well-defined and single-valued. Fix arbitrary points $y, y' \in R^m_+$ and set $x = X_k(y)$, $x' = X_k(y')$. Then, by definition, we have

$$\Phi(x, x') + \theta\langle x - x^{k-1}, x' - x\rangle + \langle y, H(x') - H(x)\rangle \geq 0,$$

$$\Phi(x', x) + \theta\langle x' - x^{k-1}, x - x'\rangle + \langle y', H(x) - H(x')\rangle \geq 0.$$

Adding these inequalities gives

$$\begin{aligned}\langle y - y', H(x') - H(x)\rangle &\geq -[\Phi(x, x') + \Phi(x', x)] + \theta \|x - x'\|^2 \\ &\geq \theta \|x - x'\|^2\end{aligned}$$

since Φ is monotone, i.e. (9) holds and assertion (i) is true. Moreover, it also follows that X_k maps bounded sets into bounded sets. In fact, letting $y' = 0$ in the above inequality gives

$$\theta \|x - x'\|^2 \leq \langle y, H(x') - H(x)\rangle \leq \sum_{i=1}^{n} y_i \langle g^i(x'), x' - x\rangle,$$

where $g^i(x')$ denotes a subgradient of H_i at x'. If we choose points y to be in a bounded set $\tilde{Y}$, then we have

$$\theta \|x - x'\|^2 \leq C\|x' - x\|, \quad \text{where} \quad C < \infty,$$

i.e. $\theta \|x - x'\| \leq C$ and the image set $X_k(\tilde{Y})$ is also bounded, hence, so is $F_k(\tilde{Y})$. The mapping $H : \tilde{X} \to R^m$ is then Lipschitz continuous on $X_k(\tilde{Y})$ with some modulus L_H, i.e.

$$\|H(x') - H(x)\| \leq L_H \|x - x'\| \quad \forall x, x' \in X_k(\tilde{Y}).$$

Applying this inequality in (9) gives

$$\langle y - y', F_k(y) - F_k(y')\rangle \geq \theta \|X_k(y) - X_k(y')\|^2 \geq (\theta / L_H^2) \|F_k(y) - F_k(y')\|^2.$$

Therefore, F_k is Lipschitz continuous and co-coercive on $\tilde{Y}$. It means that assertion (ii) is also true. □

The system (7) represents an extension of the saddle point optimality conditions for a nonlinearly constrained equilibrium problem. Such conditions are presented e.g. in [14]. From this observation it follows that we can define the following Lagrangean dual EP for SEP (7): Find a point $\tilde{v} \in Y$ such that

$$\Gamma(\tilde{v}, y) + \theta\langle \tilde{v} - y^{k-1}, y - \tilde{v}\rangle + \langle F_k(\tilde{v}), y - \tilde{v}\rangle \geq 0 \quad \forall y \in Y. \tag{10}$$

The relationships between EP (10) and SEP (7), which somewhat justify the above definition, can be stated as follows.

Proposition 1. *Let(A1)-(A4) hold. Then,*
(i) If (u^k, v^k) is a solution to SEP (7), then v^k solves EP (10);
(ii) If $\tilde{v}$ solves EP (10) and $\tilde{u} = X_k(\tilde{v})$, then $(u^k, v^k) = (\tilde{u}, \tilde{v})$ is a solution to SEP (7).

The proof follows directly from the definitions of both the problems and the mapping X_k. Thus, we can solve EP (10) instead of SEP (7). Note that it has always a unique solution since the cost mapping in (10) is strongly monotone due to Lemma 1 (ii) and the bifunction Γ is monotone. Thus, we can suggest various iterative algorithms to find a solution of EP (10).

In order to maintain the basic assumptions of this paper, we will apply the splitting type algorithm, which, starting from the initial point $v^{k,0} \in Y$ (e.g. we can set $v^{k,0} = y^{k-1}$), for each $k = 0, 1, \ldots$, computes the next iterate $v^{k,s+1} \in Y$ as the unique solution of the problem

$$\begin{aligned} \langle F_k(v^{k,s}) + \theta(v^{k,s} - y^{k-1}) + \lambda^{-1}(v^{k,s+1} - v^{k,s}), y - v^{k,s+1} \rangle \\ + \Gamma(v^{k,s+1}, y) \geq 0 \quad \forall y \in Y, \end{aligned} \tag{11}$$

where $\lambda > 0$ is a given number.

Theorem 2. *Let (A1)-(A4) hold. Then, for each k, there exists a number $\lambda'_k > 0$ such that the sequence $\{v^{k,s}\}$, defined in (11), where $\lambda \in (0, \lambda'_k)$, converges to a unique solution of EP (10) in a linear rate.*

Although there are several ways to prove convergence of splitting type methods (see e.g. [5, 16, 20]), they are either too complicated or require additional assumptions. For this reason, we give here another proof which is based on properties similar to those of the classical projection method for variational inequalities; see e.g. [19].

Fix a number $\lambda > 0$ and determine an extension of the proximal mapping for EPs as follows: $\tilde{v}$ is the value of the mapping P at a point $w \in R^m$ if it is a solution to the problem

$$\tilde{v} \in Y, \quad \langle \tilde{v} - w, v - \tilde{v} \rangle + \lambda \Gamma(\tilde{v}, v) \geq 0 \quad \forall v \in Y.$$

If (A1) and (A4) hold, then this EP has always a unique solution, i.e. the mapping $P : R^m \to R^m$ is well-defined and single-valued.

Lemma 2. *Let (A1) and (A4) hold. Then P is co-coercive with modulus 1 and nonexpansive.*

Proof. Choose $w, w' \in R^m$ and set $v = P(w)$, $v' = P(w')$. Thus,

$$\langle v - w, v' - v \rangle + \lambda \Gamma(v, v') \geq 0,$$

$$\langle v' - w', v - v' \rangle + \lambda \Gamma(v', v) \geq 0.$$

Adding these inequalities and using the monotonicity of Γ, we obtain

$$-\|v-v'\|^2+\langle w'-w,v'-v\rangle\geq 0,$$

i.e., P is co-coercive with modulus 1. Moreover, $\|w'-w\|\geq\|v-v'\|$, i.e., P is nonexpansive. □

Suppose that $Q: Y\to R^m$ is a continuous mapping. Let us consider the problem of finding a point $y^*\in Y$ such that

$$\langle Q(y^*),y-y^*\rangle+\Gamma(y^*,y)\geq 0\quad\forall y\in Y. \tag{12}$$

We denote by Y^* the solutions set of this EP.

Lemma 3. *Let (A1) and (A4) hold. Then, $y^*\in Y^*$ iff $y^*=P[y^*-\lambda Q(y^*)]$.*

Proof. If $y^*=P[y^*-\lambda Q(y^*)]$, we have

$$\langle y^*-[y^*-\lambda Q(y^*)],v-y^*\rangle+\lambda\Gamma(y^*,v)\geq 0\quad\forall v\in Y,$$

i.e. y^* solves EP (12). Conversely, let $y^*\in Y^*$, but $y^*\neq\tilde{y}=P[y^*-\lambda Q(y^*)]$. Then, by definition,

$$\langle\tilde{y}-[y^*-\lambda Q(y^*)],y^*-\tilde{y}\rangle+\lambda\Gamma(\tilde{y},y^*)\geq 0.$$

Since Γ is monotone, $\Gamma(\tilde{y},y^*)\leq-\Gamma(y^*,\tilde{y})$. It follows that

$$\lambda\langle Q(y^*),\tilde{y}-y^*\rangle+\lambda\Gamma(y^*,\tilde{y})\leq-\|y^*-\tilde{y}\|^2<0,$$

i.e. $y^*\notin Y^*$, a contradiction. □

The properties above allow us to derive the following convergence result. For the sake of clarity, we repeat all the assumptions here.

Proposition 2. *Suppose Y is a nonempty, convex and closed subset of R^m, $Q: Y\to R^m$ is strongly monotone with modulus τ and locally Lipschitz continuous, $\Gamma: Y\times Y\to R$ is a monotone continuous equilibrium bifunction such that $\Gamma(y,\cdot)$ is convex for each $y\in Y$. Then there exists a number $\lambda'>0$ such that any sequence $\{v^s\}$, defined by the rule*

$$v^{s+1}=P[v^s-\lambda Q(v^s)]\quad\text{for } k=0,1,\ldots, \tag{13}$$

where $\lambda\in(0,\lambda')$, converges to a unique solution of EP (12) in a linear rate.

Proof. Using Lemmas 2 and 3, we have

$$\begin{aligned}\left\|v^{s+1}-v^*\right\|&=\|P[v^s-\lambda Q(v^s)]-P[v^*-\lambda Q(v^*)]\|\\&\leq\|[v^s-v^*]-\lambda[Q(v^s)-Q(v^*)]\|,\end{aligned}$$

where v^* denotes the unique solution to EP (12). Fix L as the Lipschitz constant for Q on the set $\tilde{Y}=Y\cap\{v\mid\|v-v^*\|\leq\|v^0-v^*\|\}$. Clearly,

$v^0 \in \tilde{Y}$. Suppose that $v^s \in \tilde{Y}$. Then, taking into account the assumptions of this proposition, we obtain

$$\begin{aligned}\left\|v^{s+1} - v^*\right\|^2 &\leq \|v^s - v^*\|^2 - 2\lambda\langle v^s - v^*, Q(v^s) - Q(v^*)\rangle \\ &\quad +\lambda^2 \|Q(v^s) - Q(v^*)\|^2 \\ &\leq (1 - 2\lambda\tau + \lambda^2 L^2) \|v^s - v^*\|^2 \\ &= (1 - \lambda(2\tau - \lambda L^2)) \|v^s - v^*\|^2 = \nu^2 \|v^s - v^*\|^2,\end{aligned}$$

where $\nu \in (0,1)$ if $\lambda' = 2\tau/L^2$, and the result follows. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. First we observe that the assumptions of Proposition 2 are satisfied for Y, Γ, and the mapping Q, defined by

$$Q(y) = F_k(y) + \theta(y - y^{k-1}).$$

Namely, due to Lemma 1, Q is then strongly monotone with modulus $\tau = \theta$ and locally Lipschitz continuous. Next, EP (10) then coincides with (12), and the process (11) corresponds to (13). From Proposition 2 it follows that the assertion of the theorem is true. □

The splitting method above can be in principle replaced with the more usual projection method applied to the equivalent variational inequality formulation of EP (10). However, this approach requires additional differentiability conditions on Γ, more precisely, the mapping $G : Y \to R^m$, defined by

$$G(y) = \frac{\partial \Gamma(y, y')}{\partial y'}|_{y'=y},$$

has to be Lipschitz continuous for convergence.

Since SEP (7) is only an auxiliary problem, it has to be solved in a finite number of iterations within the estimate (6). For this reason, we make use of the gap function approach for ensuring the prescribed accuracy. In [14], the standard regularized gap function approach was used. Now we utilize a somewhat different approach suggested in [9], which is called nonlinear smoothing and based on transformation of the initial problem.

Given a point $z' = (x', y') \in \Omega = X \times Y$, we define the function

$$\mu_k(z') = \max_{z \in \Omega} \varphi_k(z', z) = \varphi_k(z', w(z')), \tag{14}$$

where $z = (x, y)$,

$$\varphi_k(z', z) = -\Psi(z', z) + 0.5\theta \left(\|z' - z^{k-1}\|^2 - \|z - z^{k-1}\|^2\right),$$

and Ψ is defined in (3). Since the function $\varphi_k(z', \cdot)$ is clearly continuous and strongly concave, the solution $w(z')$ of the inner problem in (14) is unique and always exists. The introduction of such a function is based on the following equivalence result, which can be derived directly from [10, Theorem 2.1.2].

Lemma 4. *Let (A1)-(A4) hold. Then SEP (7) (or equivalently, EP (8)) is equivalent to the problem: Find $w^k = (u^k, v^k) \in \Omega$ such that*

$$\Psi(w^k, z) + 0.5\theta \left(\|z - z^{k-1}\|^2 - \|w^k - z^{k-1}\|^2\right) \geq 0 \quad \forall z \in \Omega. \tag{15}$$

Thus, μ_k can be viewed as the primal gap function for EP (15), which is equivalent to EP (8). We obtain its basic properties directly from the definition.

Lemma 5. *Let (A1)-(A4) hold. Then,*
(i) $\mu_k(z') \geq 0$ for every $z' \in \Omega$;
(ii) $z' \in \Omega$ and $\mu_k(z') = 0 \Longleftrightarrow z'$ solves EP (8) $\Longleftrightarrow z' = w(z')$;
(iii) μ_k is continuous on Ω.

Besides, we need the following error bound.

Lemma 6. *Let (A1)-(A4) hold. Then,*

$$\mu_k(z') \geq 0.5\theta \left\|z' - w^k\right\|^2 \quad \forall z' \in \Omega, \tag{16}$$

where w^k denotes the unique solution to EP (8)(or (15)).

Proof. Take an arbitrary point $z' \in \Omega$ and a number $\alpha \in (0, 1)$. Then, for $z^\alpha = \alpha z' + (1 - \alpha) w^k$, we have

$$\Psi(w^k, z^\alpha) + 0.5\theta \left(\|z^\alpha - z^{k-1}\|^2 - \|w^k - z^{k-1}\|^2\right) \geq 0,$$

hence

$$\Psi(w^k, z') + 0.5\theta \left(\|z' - z^{k-1}\|^2 - \|w^k - z^{k-1}\|^2\right) \geq 0.5(1 - \alpha)\theta \|z' - w^k\|^2.$$

Taking the limit $\alpha \to 0$ gives

$$\Psi(w^k, z') + 0.5\theta \left(\|z' - z^{k-1}\|^2 - \|w^k - z^{k-1}\|^2\right) \geq 0.5\theta \|z' - w^k\|^2.$$

It was shown in Theorem 1 that Ψ is monotone, i.e. $\Psi(w^k, z') \leq -\Psi(z', w^k)$. Combining both the above inequalities, we obtain

$$\begin{aligned} \mu_k(z') &\geq -\Psi(z', w^k) + 0.5\theta \left(\|z' - z^{k-1}\|^2 - \|w^k - z^{k-1}\|^2\right) \\ &\geq 0.5\theta \left\|z' - w^k\right\|^2, \end{aligned}$$

i.e. (16) holds. □

We are now ready to describe the PPM with the dual auxiliary procedure, which converges to a solution of SEP (2) without any additional assumptions.

(DAPM) Choose a point $z^0 = (x^0, y^0) \in X \times Y$, a number $\theta > 0$, and a non-negative sequence $\{\delta_k\}$ such that

$$\sum_{k=0}^{\infty} \delta_k < \infty. \tag{17}$$

At the kth iteration, $k = 1, 2, \ldots$, we have a point $z^{k-1} = (x^{k-1}, y^{k-1}) \in X \times Y$, set $v^{k,0} = y^{k-1}$ and construct an iteration sequence $\{v^{k,s}\}$ as follows: the next iterate $v^{k,s} \in Y$ solves the problem

$$\langle F_k(v^{k,s-1}) + \theta(v^{k,s-1} - y^{k-1}) + \lambda_k^{-1}(v^{k,s} - v^{k,s-1}), y - v^{k,s}\rangle$$
$$+ f(v^{k,s}, y) \geq 0 \quad \forall y \in Y,$$

where $\lambda_k > 0$, until

$$\mu_k(z^{k,s}) \leq \delta_k^2, \tag{18}$$

where $z^{k,s} = (u^{k,s}, v^{k,s})$, $u^{k,s} = X_k(v^{k,s})$. Then set $x^k = u^{k,s}$, $y^k = v^{k,s}$. The kth iteration has been completed.

Thus, DAPM has a two-level structure. The upper level carries out the PPM process and the lower level corresponds to the dual process (11). We will call each lower level step (i.e. increasing s) an inner iteration.

Theorem 3. *Let (A1)-(A4) hold and SEP (2) be solvable. If DAPM generates a sequence $\{z^k\}$ with $z^k = (x^k, y^k)$, then there exists a sequence $\{\lambda_k'\}$ with $\lambda_k' \geq \lambda' > 0$ for $k = 0, 1, \ldots$, such that, for every $\lambda_k \in (0, \lambda_k')$, the following assertions hold:*
(i) for each k, the number of inner iterations is finite;
(ii) $\{z^k\}$ converges to a point $z^* = (x^*, y^*)$ which solves SEP (2).

Proof. For each fixed k, there exists a number $\lambda_k' > 0$ such that, for every $\lambda_k \in (0, \lambda_k')$, the sequence $\{v^{k,s}\}$ converges to the unique solution $\tilde{v}$ of EP (10) because of Theorem 2. On account of Proposition 1 (ii) and Lemma 1, the sequence $\{(u^{k,s}, v^{k,s})\}$ then converges to the unique solution w^k of EP (8). Applying Lemmas 4 and 5, we conclude that $\mu_k(z^{k,s}) \to 0$ as $s \to +\infty$, and assertion (i) is true. We have to show that conditions (5) and (6) hold. From (18) and (16) it follows that

$$0.5\theta \left\| z^k - w^k \right\|^2 \leq \mu_k(z^k) \leq \delta_k^2, \tag{19}$$

where $w^k = (u^k, v^k)$ is the unique solution to EP (8) (or SEP (7)). Hence, if we set

$$\varepsilon_k = \delta_k / \sqrt{0.5\theta},$$

the so defined sequence $\{\varepsilon_k\}$ will satisfy (5) due to (17). So, condition (6) holds and assertion (ii) follows now from Theorem 1. Next, by (17) and (19), $\lim_{k\to\infty} w^k = z^*$. From the proof of Proposition 2 we obtain $\lambda_k' = 2\tau_k / L_k^2$, where τ_k is the modulus of strong monotonicity of the mapping Q_k, defined by

$$Q_k(y) = F_k(y) + \theta(y - y^{k-1}),$$

and L_k is the Lipschitz constant for Q_k on the set

$$\tilde{Y}_k = Y \cap \{v \mid \|v - v^k\| \leq \|y^{k-1} - v^k\|\}.$$

From the definitions we see that $\tau_k \geq \theta$ and that $\|y^{k-1} - v^k\| \to 0$ as $k \to \infty$, i.e. $L_k \leq L < +\infty$. Therefore, there exists a positive lower bound λ' for the sequence $\{\lambda'_k\}$, i.e. the theorem is true. □

If the constraint mapping $H : \tilde{X} \to R^m$ is Lipschitz continuous with constant L_H, then we can specify the value of λ'. In fact, from the proofs of Lemma 1 (ii) and Theorem 3 we see that

$$\lambda' \geq 2\theta/(\theta + L_H^2/\theta)^2 = 2\theta^3/(\theta^2 + L_H^2)^2.$$

However, Theorem 3 shows that DAPM does not require this condition for convergence.

Thus, DAPM essentially exploits the specific structure of the system (2) and it seems simpler than possible variants of PPM involving the primal-dual methods. Note that the basic assumptions of this paper do not include strong (strict) monotonicity or differentiability of any function, hence the approach suggested can be applied for rather broad classes of problems.

4 Examples of Applications

In this section, we give several examples of problems which can be formulated as SEP (2). Hence, they can be solved by the method described.

4.1 Saddle Point Problems

Let us consider the problem of finding a pair of points $(x^*, y^*) \in X \times Y$ such that

$$M(x^*, y) \leq M(x^*, y^*) \leq M(x, y^*) \quad \forall x \in X, \forall y \in Y, \tag{20}$$

where the sets X and Y satisfy (A1),

$$M(x, y) = f(x) - \varphi(y) + \langle y, H(x)\rangle,$$

$f : X \to R$ and $\varphi : Y \to R$ are convex continuous functions, and $H : X \to R^m$ satisfies (A3). If we set $\varphi \equiv 0$ and $Y = R^m_+$, then (20) is nothing but the well-known Lagrangean saddle point optimality condition for the convex optimization problem:

$$\text{minimize} \quad f(x) \quad \text{over the set} \quad \{x \in X \mid H_i(x) \leq 0 \quad i = 1, \ldots, m\}.$$

Next, the other particular case of problem (20) with affine H and quadratic f and φ is known as the extended linear-quadratic program and it is intensively investigated in connection with its numerous applications in multistage

stochastic programming and discrete-time optimal control; see e.g. [18] and references therein. At the same time, (20) can be equivalently rewritten as follows:

$$\begin{cases} f(x) - f(x^*) + \langle y^*, H(x) - H(x^*)\rangle \geq 0 & \forall x \in X, \\ \varphi(y) - \varphi(y^*) - \langle H(x^*), y - y^*\rangle \geq 0 & \forall y \in Y, \end{cases}$$

i.e. it is a particular case of SEP (2), with

$$\Phi(x', x) = f(x) - f(x') \quad \text{and} \quad \Gamma(y', y) = \varphi(y) - \varphi(y').$$

Therefore, all the assumptions (A1)-(A4) hold and we can apply the corresponding variants of DAPM for finding a solution.

4.2 Spatial Price Equilibrium Models

There are many various formulations of spatial price equilibrium models which describe movements of goods among spatially distributed markets. We give one of the most popular models presented in [7].

The model is determined on a transportation network with the set of nodes N and the set of arcs A. For each node i, y_i denotes the price of a homogeneous commodity and $E_i(y)$ denotes the excess demand at this node where $y = (y_i)_{i\in N}$. For each arc $a \in A$, f_a denotes the flow and $c_a(f)$ denotes the transportation cost for shipping the commodity for this arc, where $f = (f_a)_{a\in A}$. Next, we denote by W the set of all origin-destination pairs in the net, then P_w denotes the set of paths joining pair w and $P = \bigcup_{w\in W} P_w$ denotes the set of all the paths. For each path $p \in P$, x_p denotes the flow and $C_p(x)$ denotes the transportation cost for this path, where $x = (x_p)_{p\in P}$. Set $C(x) = (C_p(x))_{p\in P}$ and $c(f) = (c_a(f))_{a\in A}$ then, clearly,

$$f = Dx \quad \text{and} \quad C(x) = D^T c(f), \tag{21}$$

where D is the arc-path incidence matrix, i.e. $D = (d_{ap})$,

$$d_{ap} = \begin{cases} 1 \text{ if path } p \text{ involves arc a,} \\ 0 \qquad \text{otherwise.} \end{cases}$$

A flow-price pattern (f^*, y^*) is said to be an equilibrium if it satisfies the following conditions:

$$\begin{gathered} \sum_{w=(k,i)\in W} \sum_{p\in P_w} x_p^* - \sum_{w=(i,j)\in W} \sum_{p\in P_w} x_p^* - E_i(y^*) \geq 0, \quad y_i^* \geq 0, \\ y_i^* \left[\sum_{w=(k,i)\in W} \sum_{p\in P_w} x_p^* - \sum_{w=(i,j)\in W} \sum_{p\in P_w} x_p^* - E_i(y^*) \right] = 0 \quad \forall i \in N; \end{gathered} \tag{22}$$

and

$$\begin{gathered} y_i^* - y_j^* + C_p(x^*) \geq 0, \quad x_p^* \geq 0, \\ x_p^* \left[y_i^* - y_j^* - C_p(x^*) \right] = 0 \quad \forall p \in P_w, \quad \forall w = (i,j) \in W. \end{gathered} \tag{23}$$

Conditions (22) represent equilibrium between input-output flows and prices at each market, whereas conditions (23) represent equilibrium between export flows and profits of shipping for each pair of origin-destination markets. Since these conditions are obviously complementarity problems, they can be equivalently rewritten as follows: Find $x^* \geq 0$ and $y^* \geq 0$ such that

$$\sum_{i \in N} \left[\sum_{w=(k,i) \in W} \sum_{p \in P_w} x_p^* - \sum_{w=(i,j) \in W} \sum_{p \in P_w} x_p^* - E_i(y^*) \right] \times (y_i - y_i^*) \geq 0 \quad \forall y_i \geq 0, i \in N;$$

$$\sum_{w \in W} \sum_{p \in P_w} \left[y_i^* - y_j^* + C_p(x^*) \right] (x_p - x_p^*) \geq 0 \quad \forall x_p \geq 0 \quad p \in P_w, w \in W.$$

This system is a particular case of SEP (2), moreover, the assumptions (A1) and (A3) are satisfied. If the negative excess demand $-E$ and the transportation cost C are monotone continuous mappings, then (A2) and (A4) also hold. At the same time, due to (21), $C(x)$ may not be strictly (strongly) monotone even if so is $c(f)$. Thus, the above system of equilibrium problems can be also solved by the proximal approach.

4.3 General Economic Equilibrium Models

There are many different economic equilibrium models which are formulated as systems of variational inequalities or equilibrium problems. Some of them were presented in [12, 13]. We describe here another economic equilibrium model, which is an extension of the known Cassel-Wald model; see e.g. [15] and, also, [12].

The model describes an economic system which deals in l commodities, n technologies of production, and m pure factors of production. In what follows, c_k denotes the price of the kth commodity, b_i denotes the total inventory of the ith factor, and a_{ij} denotes the inventory of the ith factor which is required for the unit level of the jth technology, so that $c = (c_1, \ldots, c_l)^T$, $b = (b_1, \ldots, b_m)^T$, $A = (a_{ij})_{m \times n}$. Next, x_j denotes the activity level of the jth technology, z_k denotes the output of the kth commodity, so that $x = (x_1, \ldots, x_n)^T$ and $z = (z_1, \ldots, z_l)^T$. We suppose that the relation between x and z is given by the single-valued output mapping $F : R_+^n \to R_+^l$, i.e. $z = F(x)$; prices are dependent of outputs, i.e. $c = c(z)$, and that inventories are dependent of shadow prices of factors $y = (y_1, \ldots, y_m)^T$, namely, $b \in B(y)$, where $B : R_+^m \to 2^{R_+^m}$. Thus, a set of resources may correspond to a single price vector of pure factors and the mapping B is multivalued in general.

We say that the pair $(x^*, y^*) \in R_+^n \times R_+^m$ represents an equilibrium solution in the model if the following inequalities are satisfied:

$$\langle c\left[F(x^*)\right], F(x^*) - F(x) \rangle + \langle y^*, Ax - Ax^* \rangle \geq 0 \quad \forall x \in R_+^n, \qquad (24)$$

$$\exists b^* \in B(y^*), \quad \langle b^* - Ax^*, y - y^* \rangle \geq 0 \quad \forall y \in R^m_+. \tag{25}$$

If we set

$$\Phi(x', x) = \langle c\,[F(x')]\,, F(x') - F(x) \rangle, \tag{26}$$

$$\Gamma(y', y) = \sup_{b \in B(y')} \langle b, y - y' \rangle, X = R^n_+, Y = R^m_+, \tag{27}$$

then SEP (2) coincides with (24), (25). Note that (25) is equivalent to the multivalued complementarity problem:

$$y^* \geq 0,\ \exists b^* \in B(y^*),\ b^* - Ax^* \geq 0 \text{ and } \langle b^* - Ax^*, y^* \rangle = 0,$$

which represents the usual equilibrium conditions between shadow prices and inventories limitations for all pure factors. Next, if F is differentiable, then (24) can be in principle replaced with the following complementarity problem:

$$x^* \geq 0, \quad A^T y^* - \tilde{c}(x^*) \geq 0, \quad \langle A^T y^* - \tilde{c}(x^*), x^* \rangle = 0,$$

where $\tilde{c}(x^*) = [\nabla F(x^*)]^T\, c\,[F(x^*)]$, ∇F is the Jacobian of F. This problem also represent the usual equilibrium conditions between activity levels and cost differences for all technologies. Next, in the general case, the assumptions (A1) and (A3) are clearly satisfied. Suppose that c and F are continuous mappings, $-c$ is monotone, and F has concave components. Then (A2) holds for the bifunction Φ given in (26). Note that strict (strong) monotonicity of $-c$ does not imply the same property for Φ in general.

Also, we can obtain the monotonicity of Γ in (27) from the same property of B. Again, it means that DAPM can be applied to find an equilibrium pair in this model.

Acknowledgements. The author is grateful to referees for their comments which improved essentially the exposition of the paper.

References

1. C. Baiocchi and A. Capelo. Variational and Quasivariational Inequalities. Applications to Free Boundary Problems. John Wiley and Sons, New York, 1984.
2. S.C. Billups and M.C. Ferris. QPCOMP: A quadratic programming based solver for mixed complementarity problems. Math. Programming, 76: 533–562, 1997.
3. E. Blum and W. Oettli. From optimization and variational inequalities to equilibrium problems. The Mathematics Student, 63: 127–149, 1994.
4. N. El Farouq. Pseudomonotone variational inequalities: Convergence of proximal methods. J. Optim. Theory Appl., 109:311–326, 2001.
5. D. Gabay. Application of the method of multipliers to variational inequalities. In: M. Fortin and R. Glowinski (eds), Augmented Lagrangian Methods: Application to the Numerical Solution of Boundary-Value Problems. North-Holland, Amsterdam, 299–331, 1983.

6. F. Giannessi (ed). Vector Variational Inequalities and Vector Equilibria. Mathematical Theories. Kluwer Academic Publishers, Dordrecht-Boston-London, 2000.
7. P.T. Harker and J.S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. Math. Programming, 48:161–220, 1990.
8. I.V. Konnov. A combined method for variational inequalities with monotone operators. Comp. Math. and Math. Physics, 39:1051–1056, 1999.
9. I.V. Konnov. Approximate methods for primal-dual mixed variational inequalities. Russian Math. (Iz. VUZ), 44:55–66, n.12, 2000.
10. I.V. Konnov. Combined Relaxation Methods for Variational Inequalities. Springer-Verlag, Berlin, 2001.
11. I.V. Konnov. Combined relaxation method for monotone equilibrium problems. J. Optim. Theory Appl., 111:327–340, 2001.
12. I.V. Konnov. Dual approach to one class of mixed variational inequalities. Comp. Math. and Math. Physics, 42:1276–1288, 2002.
13. I.V. Konnov. The splitting method with linear searches for primal-dual variational inequalities. Comp. Math. and Math. Physics, 43:494–507, 2003.
14. I.V. Konnov. Application of the proximal point method to nonmonotone equilibrium problems. J. Optim. Theory Appl., 119: 317–333, 2003.
15. H.W. Kuhn. On a theorem of Wald. In: H.W.Kuhn and A.W. Tucker(eds), Linear Inequalities and Related Topics. Annals of Mathem. Studies 38, Princeton University Press, Princeton, 265–273, 1956.
16. M.A. Noor. General algorithm for variational inequalities. Math. Japonica, 38:47–53, 1993.
17. M. Patriksson. Nonlinear Programming and Variational Inequality Problems: A Unified Approach. Kluwer Academic Publishers, Dordrecht, 1999.
18. R.T. Rockafellar and R.J.B. Wets. Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time. SIAM J. Control Optim., 28:810–822, 1990.
19. M. Sibony. Méthodes itératives pour les équations et inéquations aux dérivées partielles non lin'eares de type monotone. Calcolo, 7:65–183, 1970.
20. P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim., 29:119–138, 1991.

On Stability of Multistage Stochastic Decision Problems

Alexander Mänz[1] and Silvia Vogel[2]

[1] ASPECTA Lebensversicherung AG, Germany. AMaenz@aspecta.com
[2] Technische Universitäet Ilmenau, Germany. Silvia.Vogel@tu-ilmenau.de

Summary. The paper considers a general multistage stochastic decision problem which contains Markovian decision processes and multistage stochastic programming problems as special cases. The objective functions, the constraint sets and the probability measures are approximated. Making use of the Bellman Principle, (semi) convergence statements for the optimal value functions and the optimal decisions at each stage are derived. The considerations rely on stability assertions for parametric programming problems which are extended and adapted to the multistage case. Furthermore, new sufficient conditions for the convergence of objective functions which are integrals with respect to decision-dependent probability measures are presented. The paper generalizes results by Langen(1981) with respect to the convergence notions, the integrability conditions and the continuity assumptions.

1 Introduction

Many decision processes go in several steps: The decision maker, who wants to minimize a certain cost functional, chooses an action, obtains further information, reacts to this new aspects, again obtains new information, and so on, up to a finite horizon m. Costs arise at each step or at the end of the decision process only and they can depend on all states and actions observed so far. Often it is assumed that the information which becomes available between the actions can be modelled as random variable, which will be called state and whose distribution is known in advance. Normally, these random variables are not independent and, moreover, their distributions are influenced by foregoing actions.

Markovian decision processes and multistage stochastic programming problems are well-investigated models for decision processes of that kind. Despite several differences they have a similar structure (see e.g. [3]). One important common feature is that the decision maker tries to optimize the expected cost functional.

In the following we will assume that the random total costs, given a sequence of decisions $(x_1, x_2, \ldots, x_m)$ and a sequence of random states

$(S_1, \ldots, S_{m+1})$, have the following form:

$$F(S_1, \ldots, S_{m+1}, x_1, \ldots, x_m) := \sum_{k=1}^{m} c_k(S_1, \ldots, S_{k+1}, x_1, \ldots, x_k),$$

i.e. we have a sum of stage costs. The terminal costs can be included in $c_m(S_1, \ldots, S_{m+1}, x_1, \ldots, x_m)$. The aim then consists in finding a sequence $(\vartheta_1, \ldots, \vartheta_m)$ of non-anticipative deterministic decision functions that yields minimal expected total costs where the expectation is taken with respect to the common distribution of $S_1, \ldots, S_{m+1}$. Usually there are also constraints for the decisions.

Under some natural assumptions on the set of admissible decisions, which will be specified in the next section, the Bellman Principle is applicable, which enables the decision maker to determine the optimal sequence of decision functions (at least theoretically) in a 'backward procedure'. According to the Bellman Principle, at stage k, one has to solve an optimization problem of the following form

$$\min_{x_k \in D_k(\overline{s}_k, \overline{x}_{k-1})} \int_{\mathbb{S}} c_k(\overline{s}_k, s_{k+1}, \overline{x}_{k-1}, x_k) + \Phi_{k+1}(\overline{s}_k, s_{k+1}, \overline{x}_{k-1}, x_k) dP_{k+1|\overline{s}_k, \overline{x}_k}(s_{k+1})$$

where s_k means a realization of the random variable S_k. $\overline{s}_k = (s_1, \ldots, s_k)$ describes the so-called state-history and $\overline{x}_{k-1} = (x_1, \ldots, x_{k-1})$ the decision history, which are known when the decision x_k has to be chosen. $\Phi_{k+1}(\overline{s}_k, s_{k+1}, \overline{x}_{k-1}, x_k)$ denotes the minimal expected future costs. $P_{k+1|\overline{s}_k, \overline{x}_k}$ is the probability distribution of S_{k+1}, given $\overline{s}_k$ and $\overline{x}_k$. Here $D_k(\overline{s}_k, \overline{x}_{k-1})$ describes the set of admissible decisions, which will also be called 'constraint set'.

Unfortunately, there is often a lack of information about the true probability measures and they have to be approximated. Furthermore, the optimal 'future costs' are usually determined with a certain approximation error only. Hence there is a need for stability statements that clarify under what conditions the optimal values and optimal decision functions of the approximate problems come close to the corresponding quantities for the true problem.

Stability for multistage problems has been dealt with in an L^p-setting for stochastic programming problems with linear or linear-quadratic objective functions ([4], [19]) or via a stage-wise approach, which mainly relies on backward recursion (cf. [10] for Markovian decision processes and [6] for multistage stochastic programming problems).

We shall also use the stage-wise approach and derive qualitative stability results. A general model will be considered, which includes Markovian decision processes and multistage stochastic programming as well. In contrast to most stochastic programming models, we will allow for probability measures that depend on foregoing decisions. Approximations of the state space as considered by Langen [10], however, will not be considered, because approximations of this kind could be widely covered by an appropriate choice of the probability

measures. Apart from this exception, we shall give weak sufficient stability conditions which generalize the results by Langen [10] with respect to the convergence notions, the integrability conditions and continuity assumptions. Allowing for discontinuous integrands opens e.g. the possibility to deal with probabilistic objective functions and/or constraints.

The considerations rely on qualitative stability results for one-stage stochastic programs where the probability measure does not depend on the decision ([8], [9]) and extend them to the multistage case and decision-dependent probability measures.

The paper is organized as follows. In Section 2 we provide the mathematical model. Section 3 deals with general parameter-dependent one-stage optimization problems. In Section 4 the special form of the objective functions as parameter-dependent integrals is taken into account. Section 5 combines the results to the multistage setting.

2 Mathematical Model

We base the considerations on the following model: A stage consists of an observation of a state and an action which follows that observation. This agreement is in accordance with the point of view in Markovian decision models. In multistage stochastic programming problems a stage usually starts with an action, thus our model has to be specialized to apply to this case. In order to investigate a model which is as general as possible, we decided to consider stage costs c_k which may depend also on s_{k+1}, c.f. [15].

In what follows, m denotes the number of stages, the so-called horizon, and by N_m we mean the set $\{1, \ldots, m\}$.

We base our considerations on the investigations in [5] and [15]. The states or observations s_k at stage k are assumed to be elements of a standard Borel space $\mathbb{S}$, i.e. a non-empty Borel subset of a complete, separable, metric space, provided with the system of Borel subsets (to simplify presentation, here and in the following, we omit an additional symbol for the system of Borel subsets). The actions or decisions x_k are taken from a standard Borel space $\mathbb{A}$. The sets of possible actions can be constrained by certain conditions which can depend on the history so far. These conditions are described by means of multifunctions D_k. In order to explain these multifunctions we will use the following abbreviations: Let $\mathbb{H}_{1,k} := \mathbb{S}^k$, $k \in N_{m+1}$, and $\mathbb{H}_{2,k} := \mathbb{A}^k$, $k \in N_m$.

Now $D_k : \mathbb{H}_{1,k} \times \mathbb{H}_{2,k-1} \to 2^{\mathbb{A}}$, $k \in N_m \setminus \{1\}$, and $D_1 : \mathbb{H}_{1,1} \to 2^{\mathbb{A}}$ are multifunctions which determine for histories $(\overline{s}_k, \overline{x}_{k-1})$ and s_1 the constraint sets or sets of admissible actions $D_k(\overline{s}_k, \overline{x}_{k-1})$ and $D_1(s_1)$, respectively. We assume that all multifunctions D_k are closed-valued.

The probability measures $P_{k+1|\cdot,\cdot} : \mathbb{H}_{1,k} \times \mathbb{H}_{2,k} \to \mathcal{P}(\mathbb{S})$, $k \in N_m \setminus \{1\}$, describe how the state history $\overline{s}_k \in \mathbb{H}_{1,k}$ and the decision history $\overline{x}_k \in \mathbb{H}_{2,k}$ influence the probability distribution of the observation in stage $k+1$. $\mathcal{P}(\mathbb{S})$

means the set of all probability measures on the σ-field of Borel sets $\mathcal{B}(\mathbb{S})$ of $\mathbb{S}$. $P_1 \in \mathcal{P}(\mathbb{S})$ is the distribution of the first state.

The aim now consists in finding an optimal strategy (or policy, plan), i.e. a sequence of decision rules which tells the decision maker at each stage how to decide, given the foregoing observations and actions. Thus we define a strategy as a sequence $\vartheta = (\delta_k)_{k=1,\dots,m}$ of decision functions $\delta_1 : \mathbb{H}_{1,1} \to \mathbb{A}$ and $\delta_k : \mathbb{H}_{1,k} \times \mathbb{H}_{2,k-1} \to \mathbb{A}$. A sequence $(\overline{s}_k)_{k \in N_{m+1}}$ of observation histories and a strategy ϑ then define recursively a sequence of actions $(x_k(\overline{s}_k, \vartheta))_{k \in N_m}$ and decision histories $(\overline{x}_k(\overline{s}_k, \vartheta))_{k \in N_m}$ via

$$\overline{x}_1(s_1, \vartheta) = x_1(\overline{s}_1, \vartheta) := \delta_1(s_1),$$

$$x_k(\overline{s}_k, \vartheta) := \delta_k(\overline{s}_k, \overline{x}_{k-1}(\overline{s}_{k-1}, \vartheta)),\ \overline{x}_k(\overline{s}_k, \vartheta) := (\overline{x}_{k-1}(\overline{s}_{k-1}, \vartheta), x_k(\overline{s}_k, \vartheta)).$$

Thus probability measures $P_{k+1|\overline{s}_k, \vartheta}$ on $\mathcal{B}(\mathbb{S})$ can be defined by

$$P_{k+1|\overline{s}_k, \vartheta}(B) := P_{k+1|\overline{s}_k, \overline{x}_k(\overline{s}_k, \vartheta)}(B).$$

We assume that δ_k, $k = 1, \dots, m$, are Borel-measurable functions of their arguments. In order to guarantee this property for an optimal strategy we suppose that the cost functions $c_k : \mathbb{H}_{1,k+1} \times \mathbb{H}_{2,k} \to \mathbb{R} \cup \{+\infty\}$ are measurable with respect to the product sigma field of all arguments and that the graphs of the constraint multifunctions are measurable. Furthermore, we suppose that for each $B \in \mathcal{B}(\mathbb{S})$ the functions $(\overline{s}_k, \overline{x}_k) \to P_{k+1|\overline{s}_k, \overline{x}_k}(B)$ are Borel-measurable.

Then we can base our considerations on the measurable space $[\Omega_T, \Sigma]$ with

$$\Omega_T = \mathbb{S}^{m+1},\ \Sigma = \bigotimes_{i=1}^{m+1} \mathcal{B}(\mathbb{S}) \quad \text{and} \quad S_k(\omega) = s_k \text{ for } \omega = \overline{s}_{m+1} \in \Omega.$$

Using the abbreviation $\overline{S}_i := (S_1, \dots, S_i)$, a probability measure P_ϑ on $[\Omega_T, \Sigma]$ is defined by $P_\vartheta(S_1 \in B) := P_1(B)$, and

$$P_\vartheta(S_k \in B | \overline{S}_{k-1} = \overline{s}_{k-1}) := P_{k|\overline{s}_{k-1}, \vartheta}(B),\ \ B \in \mathcal{B}(\mathbb{S}),\ k \geq 2.$$

We will call a strategy $\vartheta = (\delta_k)_{k=1,\dots,m}$ *admissible*, if $\delta_1(s_1) \in D_1(s_1)$ for all $s_1 \in \text{supp}P_1$, and, for $k \in N_m \setminus \{1\}$,

$$\delta_k(\overline{s}_k, \overline{x}_{k-1}(\overline{s}_{k-1}, \vartheta)) \in D_k(\overline{s}_k, \overline{x}_{k-1}(\overline{s}_{k-1}, \vartheta)) \text{ for all } \overline{s}_k \in \text{supp}P_{k|\overline{s}_{k-1}, \vartheta}$$

where supp denotes the support of a probability measure. The set of admissible strategies will be denoted by Θ.

We exclude induced constraints, i.e. we assume that $D_k(\overline{s}_k, \overline{x}_{k-1}(\overline{s}_{k-1}, \vartheta))$ is nonempty for all admissible ϑ, all $k \in N_m \setminus \{1\}$, and all $\overline{s}_k \in \text{supp}P_{k|\overline{s}_{k-1}, \vartheta}$.

Now, for a given strategy ϑ, the random total costs can be written in the form

$$F_\vartheta(\omega) := \sum_{k=1}^{m} c_k(\overline{S}_{k+1}(\omega), \overline{x}_k(\overline{S}_k(\omega), \vartheta)).$$

The task for the decision maker consists in finding a strategy $\vartheta^* \in \Theta$ such that

$$\min_{\vartheta \in \Theta} \mathbb{E}_\vartheta F_\vartheta = \mathbb{E}_{\vartheta^*} F_{\vartheta^*}$$

where $\mathbb{E}_\vartheta$ denotes the expectation with respect to P_ϑ. We assume that there is at least one strategy ϑ such that $\mathbb{E}_\vartheta F_\vartheta < \infty$.

Given a history $(\overline{s}_m, \overline{x}_{m-1}) \in \mathbb{H}_{1,m} \times \mathbb{H}_{2,m-1}$, an optimal decision x_m^* can be obtained by

$$\inf_{x \in D_m(\overline{s}_m, \overline{x}_{m-1})} \int_{\mathbb{S}} c_m(\overline{s}_m, s, \overline{x}_{m-1}, x) dP_{m+1|\overline{s}_m, \overline{x}_{m-1}, x}(s)$$
$$= \int_{\mathbb{S}} c_m(\overline{s}_m, s, \overline{x}_{m-1}, x_m^*) dP_{m+1|\overline{s}_m, \overline{x}_{m-1}, x_m^*}(s) =: \Phi_m(\overline{s}_m, \overline{x}_{m-1}).$$

Furthermore, for $k = m-1, \ldots, 1$, x_k^* is obtained by

$$\inf_{x \in D_k(\overline{s}_k, \overline{x}_{k-1})} \int_{\mathbb{S}} c_k(\overline{s}_k, s, \overline{x}_{k-1}, x) + \Phi_{k+1}(\overline{s}_k, s, \overline{x}_{k-1}, x) dP_{k+1|\overline{s}_k, \overline{x}_{k-1}, x}(s)$$
$$= \int_{\mathbb{S}} c_k(\overline{s}_k, s, \overline{x}_{k-1}, x_k^*) + \Phi_{k+1}(\overline{s}_k, s, \overline{x}_{k-1}, x_k^*) dP_{k+1|\overline{s}_k, \overline{x}_{k-1}, x_k^*}(s)$$
$$=: \Phi_k(\overline{s}_k, \overline{x}_{k-1}).$$

In order to avoid permanent distinction between the cases $k = 1$ and $k > 1$, here and in the following we assume that dependence on a parameter x_{k-1} for $k = 1$ is ignored.

The above equations open the possibility to carry over results from one-stage optimization problems to the multi-stage case. Note that with the agreement $\Phi_{m+1}(\overline{s}_{m+1}, \overline{x}_m) := 0 \ \forall (\overline{s}_{m+1}, \overline{x}_m) \in \mathbb{H}_{1,m+1} \times \mathbb{H}_{2,m}$ there is a uniform structure for $k = m, \ldots, 1$.

It should be mentioned that Markovian decision processes as investigated by Langen [10] fit into this framework with the following agreements: $c_k(\overline{s}_{k+1}, \overline{x}_k) = \beta(s_1, x_1, s_2) \cdot \ldots \cdot \beta(s_{k-1}, x_{k-1}, s_k) \cdot r(s_k, x_k)$ where β denotes the bounded discount factor and r the reward function. Furthermore, $D_k(\overline{s}_k, \overline{x}_{k-1}) = \tilde{D}(s_k)$ and $P_{k+1|\overline{s}_k, \overline{x}_k}(B) = q(s_k, x_k, B)$ for a transition function q.

The well-investigated two-stage stochastic programming problems are obtained via $m = 2$, $c_1(s_1, s_2, x_1,) = \tilde{c}_1(x_1)$, $c_2(s_1, s_2, s_3, x_1, x_2) = \tilde{c}_2(x_1, s_2, x_2)$, $P_{2|s_1, x_1} = \tilde{P}_2$, $P_{3|\overline{s}_2, \overline{x}_2}$ arbitrary, $D_1(s_1) = \tilde{D}_1$, $D_2(s_1, x_1, s_2) = \tilde{D}_2(x_1, s_2)$.

Now we assume that each of the determining components $(D_k)_{k \in N_m}$, $(P_{k+1|\cdot,\cdot})_{k \in N_{m+1}}$, and $(c_k)_{k \in N_m}$ of our original model is approximated by a sequence in a suitable sense. Consequently, we have to investigate approximate models

$$(DM^{(n)}) \qquad (D_k^{(n)})_{k \in N_m},\ (P_{k|\cdot,\cdot}^{(n)})_{k \in N_{m+1}},\ (c_k^{(n)})_{k \in N_m}.$$

In the following, the original model will be indicated by the superscript '(0)':

$$(DM^{(0)}) \qquad (D_k^{(0)})_{k \in N_m},\ (P_{k|\cdot,\cdot}^{(0)})_{k \in N_{m+1}},\ (c_k^{(0)})_{k \in N_m}.$$

For all problems $(DM^{(n)})$, $n \in N_0 := \{0, 1, \dots,\}$, we impose the same assumptions as for the original problem. Hence we can proceed as indicated above and solve an optimization problem at each stage.

We will use the following abbreviations for $n \in N_0$:

$$\begin{aligned}
\Phi_{m+1}^{(n)}(\overline{s}_{m+1}, \overline{x}_m) &:= 0, \text{ and, for } k \in N_m, \\
\varphi_k^{(n)}(\overline{s}_{k+1}, \overline{x}_k) &:= c_k^{(n)}(\overline{s}_{k+1}, \overline{x}_k) + \Phi_{k+1}^{(n)}(\overline{s}_{k+1}, \overline{x}_k), \\
f_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}, x_k) &:= \int_{\mathbb{S}} \varphi_k^{(n)}(\overline{s}_k, s, \overline{x}_{k-1}, x_k) dP_{k+1|\overline{s}_k, \overline{x}_{k-1}, x_k}^{(n)}(s), \\
\Phi_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}) &:= \inf_{x_k \in D_k^{(n)}(\overline{s}_k, \overline{x}_{k-1})} f_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}, x_k).
\end{aligned}$$

Furthermore, we introduce - for the original and the approximate problems - the so-called solution sets for each stage $k \in N_m$, which contain the optimal decisions:

$$W_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}) := \left\{ x_k \in D_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}) \,:\, f_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}, x_k) = \Phi_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}) \right\}.$$

Our aim consists in deriving conditions which ensure that the approximate problems yield strategies $\vartheta^{(n)}$ such that $\lim_{n\to\infty} \mathbb{E}_{\vartheta^{(n)}} F_{\vartheta^{(n)}} = \mathbb{E}_{\vartheta^*} F_{\vartheta^*}$.

Sufficient for this equality are, for instance, the following two conditions (where $H_{1,k}$ and D_k are suitable sets which will be specified in Section 5):

(a) for each $k \in \{2, \dots, m\}$, all $\overline{s}_k \in H_{1,k}$, all $\overline{x}_{k-1}^{(0)} \in D_{k-1}$, all sequences $(\overline{s}_k^{(n)}, \overline{x}_{k-1}^{(n)})_{n\in N} \to (\overline{s}_k^{(0)}, \overline{x}_{k-1}^{(0)})$, one has $\lim_{n\to\infty} \Phi_k^{(n)}(\overline{s}_k^{(n)}, \overline{x}_{k-1}^{(n)}) = \Phi_k^{(0)}(\overline{s}_k^{(0)}, \overline{x}_{k-1}^{(0)})$,

(b) for P_1-almost all s, $\lim_{n\to\infty} \Phi_1^{(n)}(s) = \Phi_1^{(0)}(s)$.

Furthermore, it will be shown that the conditions which guarantee these equations also yield $K\text{–}\limsup_{n\to\infty} W_k^{(n)}(\overline{s}_k^{(0)}, \overline{x}_{k-1}^{(0)}) \subset W_k^{(0)}(\overline{s}_k^{(0)}, \overline{x}_{k-1}^{(0)})$ for all $\overline{s}_k \in H_{1,k}$ and all $\overline{x}_{k-1}^{(0)} \in D_{k-1}$, and $K\text{–}\limsup_{n\to\infty} W_1^{(n)}(s) \subset W_1^{(0)}(s)$ for P_1-almost all s.

Moreover, we will provide conditions under which even 'pointwise' convergence with respect to the state histories (together with continuous convergence with respect to the decisions) yields the desired statement.

The so-called outer Kuratowski-Painlevé-Limes $K\text{–}\limsup_{n\to\infty}$ is defined in the following section. Aiming at approximating the whole solution set of the original problem would require rather strong conditions and is in fact more than one really needs.

3 Stability of Parametric One-Stage Problems

We consider the optimization problem which occurs at a fixed stage k. In comparison to one-stage problems, in the multistage-stage setting there is

mainly one new aspect that has to be coped with, namely dependance on the parameter 'history', which can occur in the constraint sets, the integrands and the probability measures. Because for stability investigations the states and the actions have to be handled in a different way, we will distinguish the state and the decision history in the functions and multifunctions under consideration.

We shall investigate optimization problems of the following form

$$(P^{(n)}(\overline{s},\overline{x})) \qquad \min_{x\in D^{(n)}(\overline{s},\overline{x})} f^{(n)}(\overline{s},\overline{x},x),$$

with $n \in N_0$. Here $\overline{s}$ denotes an element of a standard Borel space $\mathbb{H}_1$ and $\overline{x}$ denotes an element of a standard Borel space $\mathbb{H}_2$. The optimal value functions will be denoted by $\Phi^{(n)}$ and the solution set multifunctions by $W^{(n)}$.

The optimization problems which occur at stage $k \in N_m$ in the 'backward procedure' have form $(P^{(n)}(\overline{s},\overline{x}))$. $\overline{s}$ may be regarded as 'state history' and $\overline{x}$ as 'decision history', which is available when the new decision has to be chosen. $\Phi^{(n)}(\overline{s},\overline{x})$ means the optimal 'rest costs' and $W^{(n)}(\overline{s},\overline{x})$ denotes the set of optimal decisions given the history $(\overline{s},\overline{x})$. As the optimal value function at stage k is a main part of the objective function for stage $k-1$, we aim at deriving stability assertions for the optimal value functions which are known to be desirable for the objective functions. Continuous convergence of the objective functions of a sequence of optimizations problems has proved to be an appropriate convergence notion for stability considerations. If the constraint sets remain fixed, the continuous convergence condition can be weakened. In one-stage problems often epi-convergence is imposed. However, the sum of two epi-convergent sequences is in general not epi-convergent. Hence we adapt a condition which was considered by Langen (for maximization problems) and in [10] called upper-semi-continuous convergence. We shall call this kind of convergence lower semicontinuous pointwise convergence. For the constraint sets and the solution sets we need the concept of Kuratowski-Painlevé convergence.

In [1] and [13] nowadays classical stability results for parametric optimization problems are compiled. Here we have to deal with three kinds of parameters. Firstly, the upper (approximation) index $^{(n)}$ can be interpreted as parameter. Therefore semicontinuous behavior in [1], [13] appears here in the form of semi-approximations. Furthermore, there are the parameters $\overline{s}$ and $\overline{x}$. They play a different role in stability considerations, see below. Hence we have to modify the classical results to apply to our special parametric sequences.

We start recalling the definition of Kuratowski-Painlevé-convergence:

Definition 1. *Let $(M_n)_{n\in\mathbb{N}}$ be a sequence of nonempty sets in $\mathbb{A}$. Then the Limes superior (in the Kuratowski-Painlevé sense) or 'outer limit' $K\text{-}\limsup_{n\to\infty} M_n$ and the Limes inferior (in the Kuratowski-Painlevé sense) or 'inner limit' $K\text{-}\liminf_{n\to\infty} M_n$ are defined by*

$$K\text{–}\limsup_{n\to\infty} M_n := \left\{ x \in \mathbb{A} \mid \begin{array}{l} \exists (x_n)_{n\in\mathbb{N}} \to x \text{ such that} \\ \forall n \in \mathbb{N}\ \exists m > n :\ x_m \in M_m \end{array} \right\},$$
$$K\text{–}\liminf_{n\to\infty} M_n := \left\{ x \in \mathbb{A} \mid \begin{array}{l} \exists (x_n)_{n\in\mathbb{N}} \to x \text{ such that} \\ \exists n_0 \in \mathbb{N}\ \forall n > n_0 :\ x_n \in M_n \end{array} \right\}.$$

If both limits coincide, the Kuratowski-Painlevé -Limes $K\text{–}\lim_{n\to\infty}$ *exists:*

$$K\text{–}\lim_{n\to\infty} M_n := K\text{–}\limsup_{n\to\infty} M_n = K\text{–}\liminf_{n\to\infty} M_n.$$

We have to extend these notions to multifunctions $\{C^{(n)},\ n \in N_0\}$ which map into the Borel sets of $\mathbb{A}$ and are defined on the cross product of standard Borel spaces $\tilde{\mathbb{H}}_1 \times \tilde{\mathbb{H}}_2$. $\tilde{\mathbb{H}}_1$ and $\tilde{\mathbb{H}}_2$ may be different from $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively. We introduce the new notions, because we do not need semicontinuous behavior with respect to all history parameters. Semicontinuous behavior with respect to the actions is assumed in the stability statements for the optimization problems and cannot be dispensed with. Semicontinuity assumptions with respect to the states are convenient for the derivation of sufficient conditions in Section 4. They can, however, often be replaced with pointwise convergence, compare, e.g. Theorem 4. Semicontinuous behavior with respect to s_1 is never needed. Thus an element of $\tilde{\mathbb{H}}_1$ can be understood as a whole state history for the stage under consideration or the first state only. Then the elements of $\tilde{\mathbb{H}}_2$ are in the first case the action histories and in the second case the state histories except s_1 and the complete action histories. Thus we always have $\tilde{\mathbb{H}}_1 \times \tilde{\mathbb{H}}_2 = \mathbb{H}_1 \times \mathbb{H}_2$.

In the following G_i denotes a Borel subset of $\tilde{\mathbb{H}}_i$, $i = 1, 2$.

Definition 2. *Let* $\{C^{(n)},\ n \in N_0\}$ *be a family of multifunctions* $C^{(n)} : \tilde{\mathbb{H}}_1 \times \tilde{\mathbb{H}}_2 \to 2^{\mathbb{A}}$. *The sequence* $(C^{(n)})_{n\in N}$ *is said to be*

(i) an inner semi-approximation to $C^{(0)}$ *on* G_2 *given* G_1 *(abbreviated* $C^{(n)} \xrightarrow[G_2|G_1]{K-i} C^{(0)}$*) if*

$$\forall s \in G_1\ \forall y \in G_2\ \forall (y_n)_{n\in\mathbb{N}} \to y : K\text{–}\limsup_{n\to\infty} C^{(n)}(s, y_n) \subset C^{(0)}(s, y),$$

(ii) an outer semi-approximation to $C^{(0)}$ *on* G_2 *given* G_1 *(abbreviated* $C^{(n)} \xrightarrow[G_2|G_1]{K-o} C^{(0)}$*) if*

$$\forall s \in G_1\ \forall y \in G_2\ \forall (y_n)_{n\in\mathbb{N}} \to y : K\text{–}\liminf_{n\to\infty} C^{(n)}(s, y_n) \supset C^{(0)}(s, y).$$

(iii) convergent in the Kuratowski-Painlevé sense to $C^{(0)}$ *on* G_2 *given* G_1 *(abbreviated* $C^{(n)} \xrightarrow[G_2|G_1]{K} C^{(0)}$*) if*

$$C^{(n)} \xrightarrow[G_2|G_1]{K-i} C^{(0)} \quad \text{and} \quad C^{(n)} \xrightarrow[G_2|G_1]{K-o} C^{(0)}.$$

Now we introduce the convergence notions for sequences of functions we shall deal with. $\tilde{\mathbb{A}}$ denotes a standard Borel space. In the following $\tilde{\mathbb{A}}$ will usually be interpreted as $\tilde{\mathbb{H}}_2 \times \mathbb{A}$.

Definition 3. *Let* $\{f^{(n)},\ n \in N_0\}$ *be a family of functions* $f^{(n)} : \tilde{\mathbb{H}}_1 \times \tilde{\mathbb{A}} \to \overline{\mathbb{R}}$ *and* C *a Borel subset of* $\tilde{\mathbb{A}}$. *The sequence* $(f^{(n)})_{n\in N}$ *is said to be a*

(i) lower semicontinuous approximation to $f^{(0)}$ *on* C *given* G_1 *(abbreviated* $f^{(n)} \xrightarrow[C|G_1]{l} f^{(0)}$*) if*

$$\forall s \in G_1\ \forall y \in C\ \forall (y_n)_{n\in\mathbb{N}} \to y : \liminf_{n\to\infty} f^{(n)}(s, y_n) \geq f^{(0)}(s, y),$$

(ii) upper semicontinuous approximation to $f^{(0)}$ *on* C *given* G_1 *(abbreviated* $f^{(n)} \xrightarrow[C|G_1]{u} f^{(0)}$*), if*

$$-f^{(n)} \xrightarrow[C|G_1]{l} -f^{(0)},$$

(iii) continuously convergent to $f^{(0)}$ *on* C *given* G_1 *(abbreviated* $f^{(n)} \xrightarrow[C|G_1]{c} f^{(0)}$*) if*

$$f^{(n)} \xrightarrow[C|G_1]{l} f^{(0)} \quad \text{and} \quad f^{(n)} \xrightarrow[C|G_1]{u} f^{(0)},$$

(iv) lower semicontinuously pointwise convergent to $f^{(0)}$ *on* C *given* G_1 *(abbreviated* $f^{(n)} \xrightarrow[C|G_1]{l-p} f^{(0)}$*) if*

$$f^{(n)} \xrightarrow[C|G_1]{l} f^{(0)} \quad \text{and} \quad \forall s \in G_1\ \forall y \in C : \lim_{n\to\infty} f^{(n)}(s, y) = f^{(0)}(s, y).$$

In order to employ results from parametric programming in our setting, the following lemmas are helpful. $\mathbb{A}_1$ denotes an auxiliary metric space.

Recall that a multifunction $\hat{C} : \tilde{\mathbb{H}}_2 \times \mathbb{A}_1 \to 2^{\mathbb{A}}$ is *closed* at a point (y_0, λ_0), if for all pairs of sequences $(y_n, \lambda_n)_{n\in N}$ and $(x_n)_{n\in N}$ with the properties $(y_n, \lambda_n) \to (y_0, \lambda_0)$, $x_n \in \hat{C}(y_n, \lambda_n)$ and $x_n \to x_0$ the property $x_0 \in \hat{C}(y_0, \lambda_0)$ follows. A multifunction $\hat{C} : \tilde{\mathbb{H}}_2 \times \mathbb{A}_1 \to 2^{\mathbb{A}}$ is *lower semicontinuous (l.s.c.) in the sense of Berge* at a point (y_0, λ_0), if for each open set Q satisfying $Q \cap \hat{C}(y_0, \lambda_0) \neq \emptyset$ there exists a neighborhood $U(y_0, \lambda_0)$ of (y_0, λ_0) such that for all $(y, \lambda) \in U(y_0, \lambda_0)$ the set $\hat{C}(y, \lambda) \cap Q$ is non-empty.

Lemma 1. *Let a family* $\Lambda := \{\lambda_n,\ n \in N_0\}$ *of elements of* $\mathbb{A}_1$ *with* $\lambda_n \to \lambda_0$ *and a multifunction* $\hat{C} : \tilde{\mathbb{H}}_1 \times \tilde{\mathbb{H}}_2 \times \Lambda \to 2^{\mathbb{A}}$ *be given. Suppose that* $C^{(n)}(s, y) := \hat{C}(s, y, \lambda_n)$, $n \in N_0$, $\lambda_n \in \Lambda$. *Furthermore, assume that for all* $s \in G_1$ *the multifunction* $\hat{C}(s, \cdot, \lambda_0)$ *is closed-valued. Then*

(i) $C^{(n)} \xrightarrow[G_2|G_1]{K-i} C^{(0)} \iff \forall s \in G_1,\ \hat{C}(s, \cdot, \cdot)$ *is closed on* $G_2 \times \{\lambda_0\}$,

(ii) $C^{(n)} \xrightarrow[G_2|G_1]{K-o} C^{(0)} \iff \forall s \in G_1,\ \hat{C}(s, \cdot, \cdot)$ *is l.s.c. in the sense of Berge on* $G_2 \times \{\lambda_0\}$.

Lemma 2. *Let a family* $\Lambda := \{\lambda_n,\ n \in N_0\}$ *of elements of* $\mathbb{A}_1$ *with* $\lambda_n \to \lambda_0$, *a function* $\hat{f} : \tilde{\mathbb{H}}_1 \times \tilde{\mathbb{A}} \times \Lambda \to \overline{\mathbb{R}}$ *and a Borel subset* $C \subset \tilde{\mathbb{A}}$ *be given. Suppose that* $f^{(n)}(s,y) := \hat{f}(s,y,\lambda_n)$, $n \in N_0$, $\lambda_n \in \Lambda$. *Furthermore, assume that for all* $s \in G_1$ *the function* $\hat{f}(s,\cdot,\lambda_0)$ *is l.s.c. Then,* $f^{(n)} \xrightarrow[C|G_1]{l} f^{(0)} \iff \forall s \in G_1$, $\hat{f}(s,\cdot,\cdot)$ *is l.s.c. on* $C \times \{\lambda_0\}$.

Combining these assertions, corresponding statements can be derived for continuous convergence and lower semicontinuous pointwise convergence. The proofs of the lemmas are straightforward and will be omitted. Note that the closed-valuedness and lower semicontinuity, respectively, are needed for the '$\Rightarrow$'-direction of the proofs only.

In order to formulate the stability results for our setting, we use the following assumptions. Let $C(G_1,G_2) := \{(\overline{y},x) :\ x \in D^{(0)}(\overline{s},\overline{y}),\ \overline{s} \in G_1,\ \overline{y} \in G_2\}$.

(A1) For all $\overline{s} \in G_1$, the function $f^{(0)}(\overline{s},\cdot,\cdot)$ is u.s.c. on $C(G_1,G_2)$ and $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{u} f^{(0)}$.

(A2) For all $\overline{s} \in G_1$ and $\overline{y} \in G_2$, there exists $x^{(0)}(\overline{s},\overline{y}) \in W^{(0)}(\overline{s},\overline{y})$ such that $f^{(0)}(\overline{s},\cdot,\cdot)$ is u.s.c. at $(\overline{y},x^{(0)})$ and $f^{(n)} \xrightarrow[\{(\overline{y},x^{(0)}(\overline{s},\overline{y}))\}|\{\overline{s}\}]{u} f^{(0)}$.

Now, for instance, the following statements can be proved making use of results in [1, Chapter 4] and [13].

Theorem 1. *(i) Let (A1) or (A2) hold and assume that* $D^{(n)} \xrightarrow[G_2|G_1]{K-o} D^{(0)}$. *Then* $\Phi^{(0)}(\overline{s},\cdot)$ *is u.s.c. on* G_2 *and* $\Phi^{(n)} \xrightarrow[G_2|G_1]{u} \Phi^{(0)}$.

(ii) Let $f^{(0)}(\overline{s},\cdot,\cdot)$ *be l.s.c. on* $C(G_1,G_2)$ *for all* $\overline{s} \in G_1$ *and assume that* $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{l} f^{(0)}$, $D^{(n)} \xrightarrow[G_2|G_1]{K-i} D^{(0)}$. *Furthermore, suppose that for all* $\overline{s} \in G_1$ *and all* $\overline{y} \in G_2$ *there is a compact set* K *such that for all sequences* $(y_n)_{n\in N} \to \overline{y}$ *there is an* n_0 *with* $D^{(n)}(\overline{s},y_n) \subset K\ \forall n \geq n_0$. *Then,* $\Phi^{(0)}(\overline{s},\cdot)$ *is l.s.c. on* G_2 *and* $\Phi^{(n)} \xrightarrow[G_2|G_1]{l} \Phi^{(0)}$.

(iii) Let $\Phi^{(n)} \xrightarrow[G_2|G_1]{u} \Phi^{(0)}$. *Furthermore, assume that, for all* $\overline{s} \in G_1$, $f^{(0)}(\overline{s},\cdot,\cdot)$ *is l.s.c. on* $C(G_1,G_2)$, $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{l} f^{(0)}$, *and* $D^{(n)} \xrightarrow[G_2|G_1]{K-i} D^{(0)}$. *Then,* $W^{(n)} \xrightarrow[G_2|G_1]{K-i} W^{(0)}$.

If the constraint set does not vary with n, the continuity and continuous convergence conditions can be weakened.

Theorem 2. *Suppose that for all* $\overline{s} \in G_1$ *and all* $\overline{y} \in G_2$ *there is a nonempty compact set* $D(\overline{s},\overline{y})$ *with* $D^{(n)}(\overline{s},\overline{y}) = D(\overline{s},\overline{y})\ \forall n \in N_0$. *Furthermore, assume that for all* $\overline{s} \in G_1$ *the function* $f^{(0)}(\overline{s},\cdot,\cdot)$ *is l.s.c. on* $C(G_1,G_2)$, *and* $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{l-p} f^{(0)}$. *Then* $\Phi^{(0)}(\overline{s},\cdot)$ *is l.s.c. on* G_2, $\Phi^{(n)} \xrightarrow[G_2|G_1]{l-p} \Phi^{(0)}$ *and, for all* $\overline{s} \in G_1$ *and all* $\overline{y} \in G_2$, *the inclusion* $K\text{–}\limsup_{n\to\infty} W^{(n)}(\overline{s},\overline{y}) \subset W^{(0)}(\overline{s},\overline{y})$ *holds.*

Proof. Taking Theorem 1(ii) into account, we still have to show that for all $\overline{s} \in G_1$ and all $\overline{y} \in G_2$ $\limsup_{n\to\infty} \Phi^{(n)}(\overline{s},\overline{y}) \leq \Phi^{(0)}(\overline{s},\overline{y})$ and $K\text{–}\limsup_{n\to\infty} W^{(n)}(\overline{s},\overline{y}) \subset W^{(0)}(\overline{s},\overline{y})$ hold. Let $\overline{s} \in G_1$ and $\overline{y} \in G_2$ be fixed. $f^{(0)}(\overline{s},\overline{y},\cdot)$ being l.s.c. and $D(\overline{s},\overline{y})$ being compact, there is an $x \in D(\overline{s},\overline{x})$ such that $\Phi^{(0)}(\overline{s},\overline{y}) = f^{(0)}(\overline{s},\overline{y},x)$. Consequently,

$$\limsup_{n\to\infty} \Phi^{(n)}(\overline{s},\overline{y}) \leq \limsup_{n\to\infty} f^{(n)}(\overline{s},\overline{y},x) \leq f^{(0)}(\overline{s},\overline{y},x) = \Phi^{(0)}(\overline{s},\overline{y}).$$

Now, assume that there is a sequence $(x_{n_k})_{k\in N}$ with $x_{n_k} \in W^{(n_k)}(\overline{s},\overline{y})$ and $x_{n_k} \to x_0 \notin W^{(0)}(\overline{s},\overline{y})$. $x_{n_k} \in W^{(n_k)}(\overline{s},\overline{y})$ implies $x_0 \in D(\overline{s},\overline{y})$. Otherwise there is an $x \in W^{(0)}(\overline{s},\overline{y})$, consequently, $f^{(0)}(\overline{s},\overline{y},x) < f^{(0)}(\overline{s},\overline{y},x_0)$. Thus, because of $\lim_{n\to\infty} \Phi^{(n)}(\overline{s},\overline{y}) = \Phi^{(0)}(\overline{s},\overline{y})$, we have

$$\lim_{k\to\infty} f^{(n_k)}(\overline{s},\overline{y},x_{n_k}) = f^{(0)}(\overline{s},\overline{y},x) < f^{(0)}(\overline{s},\overline{y},x_0)$$

in contradiction to $\liminf_{n\to\infty} f^{(n)}(\overline{s},\overline{y},x_n) \geq f^{(0)}(\overline{s},\overline{y},x_0)$. □

Combining and specializing the above results, e.g. Theorem 2.8 in [10] can be derived (if approximations of the state space are not taken into account).

For multifunctions $D^{(n)}$, which are described by inequality constraints, sufficient conditions are available (see e.g. [1, 17]). Semicontinuous convergence of the constraint functions plays a central role in these statements too.

4 Sufficient Conditions for Continuous Convergence and Epi-Convergence

In this section we investigate lower semicontinuous convergence for functions which are integrals. The results can then be employed to obtain sufficient conditions for either continuous convergence or (together with assertions on pointwise convergence) for lower semicontinuous pointwise convergence and hence also for epi-convergence. Corollary 3.4 in [10] and further results that rely on Corollary 3.4 give sufficient conditions assuming weak convergence of the probability measures, continuous convergence or upper-semi-continuous convergence of the integrands and uniform (with respect to the decision and the history) boundedness of the integrands. We will - among other generalizations - particularly weaken the uniform boundedness condition and the convergence condition with respect to the states.

For $i_n = n$ and $i_n = 0$, we shall investigate functions $f^{(i_n)} : \mathbb{H}_1 \times \mathbb{H}_2 \times \mathbb{A} \to \overline{\mathbb{R}}$ of the following form:

$$f^{(i_n)}(\overline{s},\overline{x},x) = \int_{\mathbb{S}} \varphi^{(i_n)}(\overline{s},s,\overline{x},x) dP^{(i_n)}_{\overline{s},\overline{x},x}(s)$$

where $P^{(i_n)}_{\overline{s},\overline{x},x}$, $n \in N_0$, are probability measures on $\mathcal{B}(\mathbb{S})$ and $\varphi^{(i_n)} : \mathbb{H}_1 \times \mathbb{S} \times \mathbb{H}_2 \times \mathbb{A} \to \overline{\mathbb{R}}$, $n \in N_0$, are integrands which are supposed to be measurable with respect to the product-σ-algebra of the arguments and integrable with respect to the probability measures under consideration. The 'parameter' i_n has been introduced in order to reduce the effort for the notation and the proof of the results, because usually the same considerations lead either (for $i_n = n$) to semi-approximation properties of $(f^{(n)})_{n\in N}$ or (for $i_n = 0$) to semicontinuity of $f^{(0)}$.

Sufficient conditions for semicontinuous convergence of sequences of functions which are integrals with respect to a probability measure that does not depend on the decision are given in [9]. We will extend the results of [9] to the parameter-dependent case. Two approaches are suggested: The first one (so-called direct approach, which was suggested by P. Lachout), assumes weak convergence of the probability measures, a lower semi-approximation property for the integrands and lower equi-integrability defined below. It can be employed to generalize Theorem 3.3 in [10]. The second approach (so called pointwise approach [18] or scalarization [7]) reduces convergence considerations for sequences of functions to convergence of sequences of real values. It is especially favorable in a random setting, but works in our case as well. It may be regarded as a bridge to results of asymptotic statistics and limit theorems of probability theory. Furthermore, it does not assume that the integrands 'behave semicontinuously' with respect to the state history.

The direct approach uses the following definition [9]:

Definition 4. *Let a sequence $(\hat{\varphi}^{(n)})_{n\in\mathbb{N}}$ of Borel-measurable functions $\hat{\varphi}^{(n)} : \mathbb{S} \to \overline{\mathbb{R}}$ and a sequence $(P^{(n)})_{n\in\mathbb{N}}$ of probability measures on $\mathcal{B}(\mathbb{S})$ be given. The family $\{(\hat{\varphi}^{(n)}, P^{(n)}), n \in \mathbb{N}\}$ is called lower equi-integrable, if there exists a $k \in \mathbb{N}$ such that*

$$\lim_{\Delta\to\infty} \inf_{n\geq k} \int_{\mathbb{S}} \hat{\varphi}^{(n)}(s)\chi_{\{\hat{\varphi}^{(n)}(s)<-\Delta\}} dP^{(n)}(s) = 0. \tag{1}$$

Let $G_1 \subset \mathbb{H}_1$ and $G_2 \subset \mathbb{H}_2$ be given.

Theorem 3. *Assume that for all $\overline{s} \in G_1$, $\overline{x}^{(0)} \in G_2$, $x^{(0)} \in D^{(0)}(\overline{s}, \overline{x}^{(0)})$ and all sequences $(\overline{x}^{(n)}, x^{(n)})_{n\in N} \to (\overline{x}^{(0)}, x^{(0)})$ the following assumptions are satisfied for $i_n = n$ and $i_n = 0$:*

(i) $P^{(i_n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}} \xrightarrow{w} P^{(0)}_{\overline{s},\overline{x}^{(0)},x^{(0)}}$,

(ii) $\liminf_{n\in\mathbb{N}} \varphi^{(i_n)}(\overline{s}, s^{(n)}, \overline{x}^{(n)}, x^{(n)}) \geq \varphi^{(0)}(\overline{s}, s, \overline{x}^{(0)}, x^{(0)})$ *for $P^{(0)}_{\overline{s},\overline{x}^{(0)},x^{(0)}}$-almost all s and all sequences $(s^{(n)})_{n\in N} \to s$,*

(iii) the functions $\varphi^{(n)}(\overline{s}, \cdot, \overline{x}^{(n)}, x^{(n)})$ are $P^{(n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}}$-integrable for all $n \in \mathbb{N}_0$ and the family $\{(\varphi^{(i_n)}(\overline{s}, \cdot, \overline{x}^{(n)}, x^{(n)}), P^{(i_n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}}), n \in \mathbb{N}\}$ is lower equi-integrable.

Then, for all $\overline{s} \in G_1$, *the function* $f^{(0)}(\overline{s}, \cdot, \cdot)$ *is l.s.c. on* $C(G_1, G_2)$ *and* $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{l} f^{(0)}$.

Proof. We follow the proof of Theorem 3.1 in [9]. Although Theorem 3.1 is formulated for functions which are defined on $\mathbb{R}^p \times \mathbb{R}^m$ only, it holds for functions which are defined on cross-products of metric spaces. Let $P_n := P^{(i_n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}}$; $P_0 := P^{(0)}_{\overline{s},\overline{x}^{(0)},x^{(0)}}$; $\xi_n := 0$ and $\varphi_n(\overline{x}, x, s) := \varphi^{(i_n)}(\overline{s}, s, \overline{x}, x)$. Then application of Theorem 3.1 to the case $i_n = 0$ yields the lower semicontinuity result for $f^{(0)}$, and application to $i_n = n$ the assertion concerning $f^{(n)}$. □

The pointwise approach can also be applied to parameter-dependent probability measures. The following result is in the spirit of Theorem 3.2 (i) in [9]. We need the following auxiliary quantities: for $\overline{s} \in G_1 \subset \mathbb{H}_1$, a family $\{(\overline{x}^{(n)}, x^{(n)}), n \in N_0\}$, $\varepsilon > 0$ and $i_n = 0$ and $i_n = n$, respectively, we define

$$Z^{(i_n)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(n)}, x^{(n)}) := \int_{\mathbb{S}} \inf_{(\overline{y},y)\in U_\varepsilon(\overline{x}^{(0)},x^{(0)})} \varphi^{(i_n)}(\overline{s}, s, \overline{y}, y) dP^{(i_n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}}(s)$$

where $U_\varepsilon(\overline{x}^{(0)}, x^{(0)})$ denotes a closed ball of radius ε and center $(\overline{x}^{(0)}, x^{(0)})$.

Theorem 4. *Let, for given sets* $G_1 \subset \mathbb{H}_1$ *and* $G_2 \subset \mathbb{H}_2$, *the following assumptions be satisfied for each* $\overline{s} \in G_1$, *each* $(\overline{x}^{(0)}, x^{(0)}) \in C(G_1, G_2)$, *all sequences* $(\overline{x}^{(n)}, x^{(n)})_{n\in N} \to (\overline{x}^{(0)}, x^{(0)})$, $i_n = n$ *and* $i_n = 0$:

(i) $\varphi^{(0)}(\overline{s}, s, \cdot, \cdot)$ *is l.s.c. at* $(\overline{x}^{(0)}, x^{(0)})$ *for* $P^{(0)}_{\overline{s},\overline{x}^{(0)},x^{(0)}}$*-almost all* s.

(ii) There is an $\bar{\varepsilon} > 0$ *such that* $Z^{(0)}_{\bar{\varepsilon}}(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(0)}, x^{(0)}) > -\infty$ *and* $Z^{(i_n)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(n)}, x^{(n)})$, $Z^{(0)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(0)}, x^{(0)})$ *exist for each* $0 < \varepsilon < \bar{\varepsilon}$ *and each* $n \in N$.

(iii) $\forall \varepsilon \in (0, \bar{\varepsilon})$,

$$\liminf_{n\to\infty} Z^{(i_n)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(n)}, x^{(n)}) \geq Z^{(0)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(0)}, x^{(0)}).$$

Then, for all $\overline{s} \in G_1$, *the function* $f^{(0)}(\overline{s}, \cdot, \cdot)$ *is l.s.c. on* $C(G_1, G_2)$ *and* $f^{(n)} \xrightarrow[C(G_1,G_2)|G_1]{l} f^{(0)}$.

Proof. Let $\overline{s} \in G_1$ and $(\overline{x}^{(0)}, x^{(0)}) \in C(G_1, G_2)$ be fixed. According to the monotone convergence lemma we have

$$\sup_{\varepsilon>0} Z^{(0)}_\varepsilon(\overline{s}, \overline{x}^{(0)}, x^{(0)}, \overline{x}^{(0)}, x^{(0)}) = f^{(0)}(\overline{s}, \overline{x}^{(0)}, x^{(0)}).$$

Furthermore, for each $0 < \varepsilon < \bar{\varepsilon}$ and each $(\overline{x}^{(n)}, x^{(n)})_{n\in N} \to (\overline{x}^{(0)}, x^{(0)})$, the relation

$$\liminf_{n\to\infty} f^{(i_n)}(\overline{s}, \overline{x}^{(n)}, x^{(n)})$$

$$\geq \liminf_{n\to\infty} \int \inf_{(\bar{y},y)\in U_\varepsilon(\overline{x}^{(0)},x^{(0)})} \varphi^{(i_n)}(\overline{s},s,\bar{y},y)dP^{(i_n)}_{\overline{s},\overline{x}^{(n)},x^{(n)}}(s)$$

$$\geq Z^{(0)}_\varepsilon(\overline{s},\overline{x}^{(0)},x^{(0)},\overline{x}^{(0)},x^{(0)})$$

holds. □

Assumption (iii) can be supplemented by several sufficient conditions. If the probability measure does not depend on the decision, considerations in [9] can be employed. In the general case, e.g. one of the following approaches may be used: if there exists a dominating measure for all probability measures, (iii) is implied by suitable semicontinuity assumptions (with respect to the actions) of the Radon-Nikodym-derivatives. Furthermore, laws of large numbers for triangular arrays are often helpful. Eventually, there is the possibility to proceed via weak convergence of probability measures. Then, however, semicontinuity with respect to the states is needed.

The above theorems generalize Langen's Theorems 3.3 and 3.5. The boundedness condition is weakened considerably. Moreover, the semicontinuity assumption with respect to the states can either be omitted or at least restricted to almost all state histories. Thus probabilities among the objective and/or constraint functions can be taken into account. For the treatment of probabilities see e.g. [9].

5 Stability of Multistage Problems

We come back to the m-stage problem. We can combine Theorem 1 or Theorem 2 with Theorem 3 or Theorem 4 in order to derive stability statements for the multistage case. We will, for example, demonstrate how Theorem 4 together with Theorem 1 can be employed (making use of Theorem 2 the continuity and approximation assumptions with respect to foregoing actions could be further weakened).

In order to make clear in what points semicontinuous behavior with respect to the states is really needed, we introduce the following sets:

Let Θ^* be the set of all optimal strategies $\vartheta^* = (\delta^*_k)_{k\in N_m}$ for the original problem with $\delta^*_k(\overline{s}_k,\overline{x}_{k-1}) = x^*_k(\overline{s}_k,\overline{x}_{k-1})$. Each ϑ^* induces a probability measure P_{ϑ^*} on Σ.

Consider a standard Borel space $H_1(\Theta^*) \subset \mathbb{S}^{m+1}$ with $P_{\vartheta^*}(H_1(\Theta^*)) = 1$ $\forall\, \vartheta^* \in \Theta^*$, and define

$$H_{1,k} := \{\overline{s}_k : (\overline{s}_k, s_{k+1},\dots,s_{m+1}) \in H_1(\Theta^*)\},\ k = 1,\dots,m,$$

$$D_1 := \{x_1 \in D^{(0)}_1(\overline{s}_1) : \overline{s}_1 \in H_{1,1}\},$$

$$D_k := \{(\overline{x}_{k-1},x_k) : \overline{x}_{k-1} \in D_{k-1}, x_k \in D^{(0)}_k(\overline{s}_k,\overline{x}_{k-1}), \overline{s}_k \in H_{1,k}\},$$
$$k = 2,\dots,m.$$

Eventually, let, for $k \in N_m$ and $i_n = 0$ and $i_n = n$, respectively,

$$Z_{k,\varepsilon}^{(i_n)}(\overline{s}_k, \overline{x}_{k-1}^{(0)}, x_k^{(0)}, \overline{x}_{k-1}^{(n)}, x_k^{(n)}) := \int_{\mathbb{S}} \sup_{(\overline{y},y)\in U_\varepsilon(\overline{x}_{k-1}^{(0)}, x_k^{(0)})} |\varphi_k^{(i_n)}(\overline{s}_k, s, \overline{y}, y)| dP^{(i_n)}_{k|\overline{s}_k, \overline{x}_{k-1}^{(n)}, x_k^{(n)}}(s).$$

Theorem 5. *Let the following assumptions be satisfied for each $k \in N_m$, all $\overline{s}_k \in H_{1,k}$, all $\overline{x}_k^{(0)} \in D_k$, all sequences $(\overline{x}_k^{(n)})_{n\in N} \to \overline{x}_k^{(0)}$, $i_n = n$ and $i_n = 0$:*

(i) $\lim_{n\to\infty} |c_k^{(i_n)}(\overline{s}_k, s^{(0)}, \overline{x}_k^{(n)}) - c_k^{(0)}(\overline{s}_k, s^{(0)}, \overline{x}_k^{(0)})| = 0$ *for* $P_{k+1|\overline{s}_k, \overline{x}_k^{(0)}}$*-almost all* $s^{(0)}$.

(ii) $D_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}^{(n)}) \xrightarrow{K} D_k^{(0)}(\overline{s}_k, \overline{x}_{k-1}^{(0)})$.

(iii) $\exists$ *compact* K $\exists n_0$ $\forall n \geq n_0$ *such that* $D_k^{(n)}(\overline{s}_k, \overline{x}_{k-1}^{(n)}) \subset K$.

(iv) There is an $\bar{\varepsilon} > 0$ *such that* $Z_{k,\bar{\varepsilon}}^{(0)}(\overline{s}_k, \overline{x}_{k-1}^{(0)}, x_k^{(0)}, \overline{x}_{k-1}^{(0)}, x_k^{(0)}) < \infty$ *and* $Z_{k,\varepsilon}^{(n)}(\overline{s}_k, \overline{x}_{k-1}^{(0)}, x_k^{(0)}, \overline{x}_{k-1}^{(n)}, x_k^{(n)})$ *exist for each* $\varepsilon \in (0, \bar{\varepsilon})$ *and each* $n \in N$ *and* $\forall\, 0 < \varepsilon < \bar{\varepsilon}$,

$$\lim_{n\to\infty} Z_{k,\varepsilon}^{(i_n)}(\overline{s}_k, \overline{x}_{k-1}^{(0)}, x_k^{(0)}, \overline{x}_{k-1}^{(n)}, x_k^{(n)}) = Z_{k,\varepsilon}^{(0)}(\overline{s}_k, \overline{x}_{k-1}^{(0)}, x_k^{(0)}, \overline{x}_{k-1}^{(0)}, x_k^{(0)}).$$

Then, for $k = 2, \ldots, m$, *the function* $\Phi_k^{(0)}(\overline{s}_k, \cdot)$ *is continuous on* D_{k-1}, $\Phi_k^{(n)} \xrightarrow[D_{k-1}|H_{1,k}]{c} \Phi_k^{(0)}$, $W_k^{(n)} \xrightarrow[D_{k-1}|H_{1,k}]{K-i} W_k^{(0)}$, $\lim_{n\to\infty} \Phi_1^{(n)}(s) = \Phi_1^{(0)}(s)$, *and* $W_1^{(n)}(s) \xrightarrow{K-i} W_1^{(0)}(s)$ *for* P_1*-almost all* s.

Proof. We proceed by backward induction. Because of $\Phi_{m+1}^{(i_n)}(\overline{s}_{m+1}, \overline{x}_m) = 0$ for all $(\overline{s}_{m+1}, \overline{x}_m) \in \mathbb{H}_{1,k+1} \times \mathbb{H}_{2,k}$ and all $n \in N$, we have $\varphi_m^{(i_n)} = c_m^{(i_n)}$. Applying Theorem 4 to $c_m^{(i_n)}$ and $-c_m^{(i_n)}$, $G_1 = H_{1,m}$, $G_2 = D_{m-1}$ and $C(G_1, G_2) = D_m$, we obtain the continuity of $f_m^{(0)}(\overline{s}_m, \cdot)$ on D_m and $f_m^{(i_n)} \xrightarrow[D_m|H_{1,m}]{c} f_m^{(0)}$. This, together with the assumptions (ii) and (iii) gives, by Theorem 1, in case $i_n = 0$ the continuity of $\Phi_m^{(0)}(\overline{s}_m, \cdot)$ on D_{m-1} and, for $i_n = n$, $\Phi_m^{(n)} \xrightarrow[D_{m-1}|H_{1,m}]{c} \Phi_m^{(0)}$, and $W_m^{(n)} \xrightarrow[D_{m-1}|H_{1,m}]{K-i} W_m^{(0)}$. For the stage k we can proceed in the same way. The continuity assumptions for $\varphi_k^{(0)}$ are satisfied because of (i) and the continuity of $\Phi_{k+1}^{(0)}$. Integrability is assumed in (iv). □

Acknowledgements. The authors are grateful to the referees for helpful remarks and suggestions.

References

1. B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer. Nonlinear Parametric Optimization. Akademie-Verlag Berlin, 1982.
2. D.P. Bertsekas and S.E. Shreve. Stochastic Optimal Control: The Discrete Time Case. Academic Press, New York, 1978.

3. J. Dupačová and K. Sladký. Comparison of Multistage Stochastic Programs with Recourse and Stochastic Dynamic Programs with Discrete Time. Z. Angew. Math. Mech. 82: 11-12: 753-765, 2002.
4. O. Fiedler and W. Römisch. Stability of multistage stochastic programming. Ann. Oper. Res. 56: 79-93, 1995.
5. K. Hinderer. Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter. Lecture Notes in Oper. Res. and Math. Syst., 33, 1970.
6. V. Kaňková. A Remark on Analysis of Multistage Stochastic Programs: Markov Dependence. Z. Angew. Math. Mech. 82, 11-12: 781-793, 2002.
7. L.A. Korf and R.J.-B. Wets. Random lsc functions: an ergodic theorem. Math. Oper. Res. 26: 421-445, 2001.
8. S. Vogel and P. Lachout. On continuous convergence and epi-convergence of random functions - Theory and relations. Kybernetika 39: 75-98, 2003.
9. S. Vogel and P. Lachout. On continuous convergence and epi-convergence of random functions - Sufficient conditions and applications. Kybernetika 39: 99-118, 2003.
10. H.-J. Langen. Convergence of Dynamic Programming Models. Math. Oper. Res. 6: 493-512, 1981.
11. A. Mänz. Stabilität mehrstufiger stochastischer Optimierungsprobleme. Diploma Thesis, TU Ilmenau, 2003.
12. P.H. Müller and V. Nollau (ed.). Steuerung stochastischer Prozesse. Akademie-Verlag, Berlin, 1984.
13. S.M. Robinson. Local epi-continuity and local optimization. Math. Programming 37: 208 - 222, 1987.
14. R.T. Rockafeller and R. J.-B. Wets. Variational Analysis. Springer, 1998.
15. M. Schäl. Conditions for Optimality in Dynamic Programming and for the Limit of n-Stage Optimal Policies to Be Optimal. Z. Wahrscheinlichkeitstheorie verw. Gebiete 32: 179-196, 1975.
16. S. Vogel. Stochastische Stabilitätskonzepte. Habilitation, TU Ilmenau, 1991.
17. S. Vogel. A stochastic approach to stability in stochastic programming. J. Comput. and Appl. Mathematics, Series Appl. Analysis and Stochastics 56: 65-96, 1994.
18. S. Vogel.On stability in stochastic programming - Sufficient conditions for continuous convergence and epi-convergence. Preprint TU Ilmenau, 1994.
19. J. Wang. Stability of multistage stochastic programming. Ann. Oper. Res. 56: 313-322, 1995.

Nonholonomic Optimization

C. Udrişte[1], O. Dogaru[1], M. Ferrara[2], and I. Ţevy[1]

[1] University Politehnica of Bucharest, Department of Mathematics, Splaiul Independenţei 313, 060042 Bucharest, Romania. udriste@mathem.pub.ro
[2] University of Messina, Faculty of Economics, Via dei Verdi 75, 98122 Messina, Italy. mferrara@unime.it

Summary. In this paper one generalizes various types of constrained extremum, keeping the Lagrange or Kuhn-Tucker multipliers rule. The context which supports this development is the nonholonomic optimization theory which requires a holonomic or nonholonomic objective function subject to nonholonomic or holonomic constraints. We refined such a problem using two new ideas: the replacement of the point or velocity constraints by a curve selector, and the geometrical interpretation of the Lagrange and Kuhn-Tucker parameters. The classical optimization theory is recovered as a particular case of extremum constrained by a curve selector.

1 Extremum Constrained by a Curve Selector

This section contains an improved version of curve selector theory, developed for the first time in [15]. Let $D \subset \mathbb{R}^n$ be an open set and $f : D \to \mathbb{R}$ be a real function. In order to establish that $x_0 \in D$ is a free extremum point, or an extremum point constrained by holonomic/nonholonomic equalities or inequalities, it is enough to consider the values of the function f on some parametrized curves passing through x_0 [8]-[11]. More precisely, if $\alpha : I \to D$, $\alpha(t_0) = x_0$, is such a curve, then we must take into consideration the values $f(\alpha(t))$ for t in a neighborhood of t_0. Recently [14], these results were refined, in the sense that it is sufficient to consider the values $f(\alpha(t))$ for $t \in [t_0, t_0 + \varepsilon)$. In this context we discovered the notions of curve selector and of extremum constrained by a curve selector, that permit a unification of all types of extremum and even a generalization.

For each $x \in D$, we shall denote by Γ_x a family of parametrized curves passing through the point x at a given moment. This family of curves will be specified from case to case.

Definition 1. *Let $\mathcal{P}(\Gamma_x)$ be the power set of Γ_x. Any function*

$$\Upsilon : D \to \cup_{x \in D} \mathcal{P}(\Gamma_x), \quad \Upsilon(x) \subset \Gamma_x$$

is called curve selector on D. The elements of $\Upsilon(x)$ are called admissible curves at the point x.

Definition 2. *Let $f : D \to \mathbb{R}$ be a given function and Υ be a curve selector on D. If $f(\alpha(t)) \geq f(x_0)$, $\forall t \in [t_0, t_0 + \varepsilon)$, $\forall \alpha \in \Upsilon(x_0)$, $x_0 = \alpha(t_0)$, then $x_0 \in D$ is called minimum point of f constrained by the selector Υ.*

Let us suppose that Γ_x is either the set of regular C^1 curves at x or the set of C^2 curves having x either as a regular point or as singular point of order 2. Under this hypothesis, in [15] was shown that a free extremum problem or a constrained extremum problem (with equalities or inequalities) can be considered as an extremum problem constrained by a curve selector.

Let

$$\omega^a(x) = \sum_{j=1}^{n} \omega_j^a(x) dx^j, \quad a = \overline{1,p}, \quad p < n$$

be C^1 Pfaff forms. These Pfaff forms can be used to create equality constraints (Pfaff equations) or inequality constraints (for example, Pfaff inequalities) on velocities.

In [1]-[15] was studied the following type of extremum.

Definition 3. *The point $x_0 \in D$ is called minimum point of a function f constrained by the Pfaff system $\omega^a = 0$, $a = \overline{1,p}$, if for each integral curve $\alpha : I \to D$ of this system, with $\alpha(t_0) = x_0$, it follows*

$$f(\alpha(t_0)) \leq f(\alpha(t)), \ \forall t \in (t_0 - \varepsilon, t_0 + \varepsilon).$$

The Pfaff system (ω^a) generates the partial selectors

$$\Upsilon^a(x) = \{\alpha \in \Gamma_x | \alpha \text{ is an integral curve of the Pfaff equation } \omega^a(x) = 0\},$$

which produce the general selector (associated to the Pfaff system)

$$\Upsilon(x) = \cap_{a=1}^{p} \Upsilon^a(x).$$

Theorem 1. *$x_0 \in D$ is a minimum point of a function $f : D \to \mathbb{R}$ constrained by the selector Υ if and only if $x_0 \in D$ is a minimum point for f constrained by the Pfaff system $\omega^a(x) = 0$, $a = \overline{1,p}$.*

Also, we remark that we can replace the selector Υ with those determined by the partial selectors

$$\Upsilon^a(x_0) = \{\alpha \in \Gamma_{x_0} | \omega^a(\alpha(t)) = 0, \ \forall t \in [t_0, t_0 + \varepsilon), \ \alpha(t_0) = x_0\}.$$

The primitive of each Pfaff form $\omega^a(x)$ defines the partial selectors

$$\Upsilon^a(x_0) = \left\{\alpha \in \Gamma_{x_0} | \int_{t_0}^{t} \langle \omega(\alpha(u)), \alpha'(u) \rangle du \geq 0, \ \forall t \in [t_0, t_0 + \varepsilon)\right\},$$

where $\alpha(t_0) = x_0$. From this point of view, the selector associated to all Pfaff forms is

$$\Upsilon(x_0) = \cap_{a=1}^{p} \Upsilon^a(x_0).$$

Definition 4. *The point $x_0 \in D$ is called a minimum point of a function $f : D \to \mathbb{R}$ constrained by $\omega^a \geq 0$, $a = \overline{1,p}$ if x_0 is a minimum point of f constrained by the selector Υ.*

This type of extremum was studied in [6], [10].

Remark 1. In the case of classical constraints, the system of C^1 functions $g^a : D \to \mathbb{R}$, $a = \overline{1,p}$, induces two types of constraints: point constraints defined by $g^a(x) = 0$ or $g^a(x) \geq 0$ and velocity constraints defined by the subspace $dg^a(x) = 0$, $a = \overline{1,p}$ of the tangent space T_xD. The points described by these constraints split in two types: interior points ($g^a(x) > 0$, $a = \overline{1,p}$) and boundary points ($g^a(x) \geq 0$, $a = \overline{1,p}$ and $\exists a$ with $g^a(x) = 0$), with the usual topological significance.

The constraints on points and on velocities, which are correlated by the functions g^a, induce a selector of curves, one contribution coming from the quality of a point to be interior or boundary point.

In the case of extremum points with Pfaff constraints, it appears only the velocity constraints defined by the subspace $\sum_{i=1}^{n} \omega_i^a(x)dx^i = 0$, $a = \overline{1,p}$ of the tangent space $T_x(D)$. Here each point is considered like boundary point, and it is susceptible of being extremum point.

These remarks permit to introduce a more general type of extremum in which the constraints on points and the constraints on velocities are not necessarily correlated, but the Lagrange multipliers rule still survives.

2 Extremum with Point and/or Velocity Constraints

Let $\omega(x) = \sum_{j=1}^{n} \omega_j(x)dx^j$ be a C^0 Pfaff form on D. Let S and bS be two arbitrary disjoint subsets in D. The points of S will be called "*interior points*" and the points of bS will be called "*boundary points*". In this context, we say that the set $M = S \cup bS$ represents the constraints of *inequality type*. The pair (ω, M) determines an *inequality curve selector*:

$$\Upsilon(x_0) = \begin{cases} \Gamma_{x_0} \text{ if } x_0 \in S \\ \left\{\alpha \in \Gamma_{x_0} \middle| \int_{t_0}^{t} \langle \omega(\alpha(u)), \alpha'(u)\rangle du \geq 0,\ \forall t \in [t_0, t_0 + \varepsilon)\right\} \text{ if } x_0 \in bS \\ \emptyset \text{ if } x_0 \in D \setminus M. \end{cases}$$

If $T \subset D$ is an arbitrary subset, then the pair (ω, T) defines an *equality curve selector*

$$\Upsilon_0(x_0) = \begin{cases} \{\alpha \in \Gamma_{x_0} | \langle \omega(\alpha(t)), \alpha'(t)\rangle = 0,\ \forall t \in [t_0, t_0 + \varepsilon]\} & \text{if } x_0 \in T \\ \phi, & \text{if } x_0 \notin T. \end{cases}$$

In this sense, we say that the set T represents the constraints of *equality type*.

Remark 2. . Any equality selector can be expressed by inequality selectors. More precisely, for an arbitrary subset $T \subset D$, let us consider the inequality selectors Υ_+ and Υ_- defined by $(\omega, S \cup bS)$ respectively $(-\omega, S \cup bS)$, where $S = \emptyset$ and $bS = T$. Then $\Upsilon_0(x) = \Upsilon_+(x) \cap \Upsilon_-(x), \forall x \in D$. Also, the inequality selector Υ_0 can be deactivated considering $T = D$.

Having in mind the previous idea, we can introduce inequality selectors Υ^a defined by the pairs $(\omega^a, S_a \cup bS_a)$, $a = \overline{1,p}$ and equality selectors Υ_0^i defined by the pairs (η^i, T_i), $i = \overline{1,m}$. Then we built

$$N = \cap_{a=1}^{p}(S_a \cup bS_a), \quad bS = \{x \in N | \exists a = \overline{1,p}, x \in bS_a\},$$

$$S = N \setminus bS, \quad T = \cap_{i=1}^{m} T_i, \quad M = N \cap T.$$

In case of absence of equality constraints, we have $M = N$ and in case of absence of inequality constraint we have $M = T$.

If ω means the Pfaff form system (ω^a) and η means the Pfaff form system (η^i), then the triple (ω, η, M) defines the curve selector

$$\Upsilon(x) = (\cap_{a=1}^{p} \Upsilon^a(x)) \cap \left(\cap_{i=1}^{m} \Upsilon_0^i(x)\right), \quad \forall x \in D.$$

Definition 5. *Let $f : D \to \mathbb{R}$ be a real function. We say that $x_0 \in M$ is a minimum point of f constrained by (ω, η, M) if x_0 is a minimum point constrained by the selector Υ. We say that ω and η represents the velocity constraints, and M represents the point constraints. The triple (ω, η, M) is called system of point/velocity constraints or system of constraints.*

Convenient selection of the objects ω, η, M leads to all the types of extremum mentioned in the previous paragraph.

- *Case of free extremum*: $S_a = D, bS_a = \emptyset, \omega^a =$ arbitrary (without equation constraints).
- *Case of classical equality constraints*: $T_i = \{x \in D | g^i(x) = 0\}$, $\eta^i = dg^i$ (without inequation constraints).
- *Case of classical inequality constraints*: $S_a = \{x \in D | g^a(x) > 0\}$, $bS_a = \{x \in D | g^a(x) = 0\}$, $\omega^a = dg^a$ (without equation constraints).
- *Case of Pfaff equality constraints*: $T_i = D$, $\eta^i = \omega^i$ (without inequation constraints).
- *Case of Pfaff inequality constraints*: $S_a = \emptyset$, $bS_a = D$ (without equation constraints).

The previous remark shows that any type of extremum can be considered as extremum constrained by inequation constraints.

Definition 6. *Let (ω, η, M) be a system of point/velocity constraints. Let $x_0 \in M$ and $B(x_0) = \{a | x_0 \in bS_a\} \subset \{1, \ldots, p\}$. The system (ω, η) is called regular at x_0 if rank $(\omega^a(x_0), \eta^i(x_0)) = m +$ card $B(x_0)$, where $a \in B(x_0), i = \overline{1,m}$.*

Theorem 2. *Let $f : D \to \mathbb{R}$ be a C^1 function and $x_0 \in M$ such that the system (ω, η) is regular at x_0. Suppose x_0 is a minimum point of f constrained by (ω, η, M). Then there exist $\lambda_a \geq 0$, $a = \overline{1,p}$ and $\mu_i \in R$, $i = \overline{1,m}$, such that*

$$df(x_0) = \sum_{a=1}^{p} \lambda_a \omega^a(x_0) + \sum_{i=1}^{m} \mu_i \eta^i(x_0).$$

Moreover, if $\lambda_a > 0$, then $x_0 \in bS_a$.

Proof. Let $v \in \mathbb{R}^n$, $v \neq 0$, such that $\langle \omega^a(x_0), v\rangle \geq 0, \forall a \in B(x_0)$ and $\langle \eta^i(x_0), v\rangle = 0$, $\forall i = \overline{1,m}$. Let

$$J(x_0) = \{a \in B(x_0) | \langle \omega^a(x_0), v\rangle = 0\}.$$

From the regularity condition, it follows the existence of an integral curve of the Pfaff system $\omega^a = 0, \eta^i = 0$, $a \in J(x_0)$ and $i = \overline{1,m}$ with $\alpha(t_0) = x_0$ and $\alpha'(t_0) = v$. Hence,

1. $a \in J(x_0)$, implies $\int_{t_0}^{t} \langle \omega^a(\alpha(u)), \alpha'(u)\rangle du = 0$,
2. $a \in B(x_0) \setminus J(x_0)$ implies $\int_{t_0}^{t} \langle \omega^a(\alpha(u)), \alpha'(u)\rangle du \geq 0$, $\forall t \in [t_0, t_0 + \varepsilon)$,
3. $i = \overline{1,m}$ implies $\langle \eta^i(\alpha(t)), \alpha'(t)\rangle = 0$, $\forall t$.

Consequently $\alpha \in \Upsilon(x_0)$. Since x_0 is a minimum point of f constrained by the selector Υ, it follows $f(\alpha(t_0)) \leq f(\alpha(t))$, $\forall t \in [t_0, t_0 + \varepsilon)$. Hence $\langle \mathrm{grad} f(x_0), v\rangle \geq 0$. Finally, we obtain

$$df(x_0) = \sum_{a \in B(x_0)} \lambda_a \omega^a(x_0) + \sum_{i=1}^{m} \mu_i \eta^i(x_0),$$

with $\lambda_0 \geq 0$. For $a \notin B(x_0)$, we consider $\lambda_a = 0$. □

The multipliers λ and μ from the Theorem 2.3 are unique.

The regularity in Definition 3 can be replace by a more general condition.

Definition 7. *We say that (ω, η, M) satisfies the Kuhn-Tucker regularity condition at $x_0 \in M$ if from $x_0 \in bS \cap T$ it follows that for any vector $v \neq 0$ with $\langle \omega^a(x_0), v\rangle \geq 0$, $\forall a \in B(x_0) = \{a | x_0 \in bS_a\}$ and $\langle \eta^i(x_0), v\rangle = 0$, $\forall i = \overline{1,m}$, it exists a parametrized curve $\alpha \in \Upsilon_0$, $\alpha(t_0) = x_0$ such that $\alpha'(t_0) = v$.*

Theorem 3. *Let $f : D \to \mathbb{R}$ be a C^1 function. If the constraints triple (ω, η, M) satisfies the Kuhn-Tucker regularity condition at $x_0 \in M$ and x_0 is a minimum point of f constrained by (ω, η, M), then there exist $\lambda_a \geq 0$ and $\mu_i \in \mathbb{R}$ such that*

$$df(x_0) = \sum_{a=1}^{p} \lambda_a \omega^a(x_0) + \sum_{i=1}^{m} \mu_i \eta^i(x_0).$$

Moreover, if $\lambda_a > 0$ implies $x_0 \in bS_a$.

The proof is contained in the proof of Theorem 2. The Kuhn-Tucker multipliers λ and μ from the Theorem 3 are not unique; see Example 2 in next section.

Suppose now that Γ_x represents the family of all parametrized C^2 curves which passes through x and which are regular at x.

Theorem 4. *Let us consider the constraints (ω, η, M) with ω, η of class C^2. Let $f : D \to \mathbb{R}$ be of class C^2 and $x_0 \in M$. Suppose that*
i) there exist $\lambda_a \geq 0$, $a = \overline{1,p}$ and $\mu_i \in \mathbb{R}$ such that

$$df(x_0) = \sum_{a=1}^{p} \lambda_a \omega^a(x_0) + \sum_{i=1}^{m} \mu_i \eta^i(x_0),$$

and, if $\lambda_a > 0$, then $x_0 \in bS_a$;
ii) the restriction of the quadratic form

$$d^2 f(x_0) - \frac{1}{2} \sum_{a=1}^{p} \lambda_a \sum_{i,j=1}^{n} \left(\frac{\partial \omega_i^a}{\partial x^j} + \frac{\partial \omega_j^a}{\partial x^i} \right) (x_0) dx^i dx^j -$$

$$- \frac{1}{2} \sum_{k=1}^{m} \mu_k \sum_{i,j=1}^{n} \left(\frac{\partial \eta_i^k}{\partial x^j} + \frac{\partial \eta_j^k}{\partial x^i} \right) (x_0) dx^i dx^j$$

to the velocity subspace

$$\begin{cases} \sum_{j=1}^{n} \omega_j^a(x_0) dx^j = 0, \ a \in J^1(x_0) = \{a \in B(x_0) | \lambda_a > 0\} \\ \sum_{j=1}^{n} \eta_j^i(x_0) dx^j = 0, \ i = \overline{1,m} \end{cases} \tag{1}$$

is positive definite.

Then, x_0 is a minimum point of f constrained by (ω, η, M).

Proof. Let $\alpha \in \Upsilon(x_0)$ with $\alpha(t_0) = x_0$. Hence

$$\int_{t_0}^{t} \langle \omega^a(\alpha(u)), \alpha'(u) \rangle du \geq 0, \ \forall a \in [t_0, t_0 + \varepsilon), \ \forall a \in B(x_0)$$

and

$$\langle \eta^i(\alpha(t), \alpha'(t) \rangle = 0, \ \forall t \in [t_0, t_0 + \varepsilon), \quad \forall i = \overline{1,m}. \tag{2}$$

Case 1. If there exists $a \in J'(x_0)$, with $\langle \omega^a(x_0), \alpha'(t_0) \rangle > 0$, then, taking account the relations in i) and (2), it follows

$$df(x_0)(\alpha'(t_0)) = \sum_{a=1}^{p} \lambda_a \langle \omega^a(x_0), \alpha'(x_0) \rangle > 0.$$

Using the Taylor formulas

$$f(x) - f(x_0) = df(x_0)(x - x_0) + \mathcal{O}(\|x - x_0\|),$$

$$\alpha(t) - \alpha(t_0) = \alpha'(t_0)(t - t_0) + \beta(t) \cdot (t - t_0),$$

with $\lim_{t \to t_0} \beta(t) = 0$, we obtain

$$f(\alpha(t)) - f(\alpha(t_0)) = (t - t_0)df(x_0)(\alpha'(t_0)) + (t - t_0)df(x_0)(\beta(t))$$

$$+ \mathcal{O}(\|\alpha(t) - \alpha(t_0)\|) = (t - t_0)df(x_0)(\alpha'(t_0)) + \mathcal{O}(t - t_0) \geq 0, \ \forall t \in [t_0, t_0 + \varepsilon).$$

Case 2. Suppose $\langle \omega^a(x_0), \ \alpha'(t_0) \rangle = 0, \forall a \in J'(x_0)$. Hence $\alpha'(t_0)$ belong to the subspace (6). The composed function

$$\varphi(t) = f(\alpha(t)) - \sum_{a=1}^{p} \lambda_a \int_{t_0}^{t} \langle \omega^a(\alpha(u)), \alpha'(u) \rangle du$$

has the derivative

$$\varphi'(t) = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(\alpha(t)) \frac{dx^i}{dt} - \sum_{a=1}^{p} \lambda_a \sum_{i=1}^{n} \omega_i^a(\alpha(t)) \frac{dx^i}{dt},$$

and hence

$$\varphi'(t_0) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x^i}(x_0) - \sum_{a=1}^{p} \lambda_a \omega_i^a(x_0) \right) \frac{dx^i}{dt}$$

$$= (df(x_0) - \sum_{a=1}^{p} \lambda_a \omega^a(x_0))(\alpha'(t_0)) \overset{(2)}{=} 0.$$

Also

$$\varphi''(t) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x^i \partial x^j}(\alpha(t)) \frac{dx^i}{dt} \frac{dx^j}{dt} + \sum_{i=1}^{n} \frac{\partial f}{dx^i}(\alpha(t)) \frac{d^2 x^i}{dt^2}$$

$$-\frac{1}{2} \sum_{a=1}^{p} \lambda_a \sum_{i,j=1}^{n} \left(\frac{\partial \omega_i^a}{\partial x^j} + \frac{\partial \omega_j^a}{\partial x^i} \right) (\alpha(t)) \frac{dx^i}{dt} \frac{dx^j}{dt} - \sum_{a=1}^{p} \lambda_a \sum_{i=1}^{n} \omega_i^a(\alpha(t)) \frac{d^2 x^i}{dt^2}.$$

Then

$$\varphi''(t_0) = d^2 f(x_0) - \frac{1}{2} \sum_{a=1}^{p} \lambda_a \sum_{i,j=1}^{n} \left(\frac{\partial \omega_i^a}{\partial x^j} + \frac{\partial \omega_j^a}{\partial x^i} \right) (x_0) \frac{dx^i}{dt}(t_0) \frac{dx^j}{dt}(t_0)$$

$$+ \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x^i}(x_0) - \sum_{a=1}^{p} \lambda_a \omega_i^a(x_0) \right) \frac{d^2 x^i}{dt^2}(t_0).$$

Taking into account the relation of i), it follows that the coefficient of $\frac{d^2x^i}{dt^2}(t_0)$ is $\sum_{k=1}^{n} \mu_k \eta_i^k(x_0)$. On the other hand, from (2) we obtain

$$\sum_{i=1}^{n} \eta_i^k(\alpha(t)) \frac{dx^i}{dt} = 0, \quad t \in [t_0, t_0 + \varepsilon), \quad k = \overline{1,m}.$$

Taking again the derivative, and making $t = t_0$, we obtain

$$\sum_{i=1}^{n} \eta_i^k(\alpha(t_0)) \frac{d^2x^i}{dt^2}(t_0) = -\frac{1}{2} \sum_{i,j=1}^{n} \left(\frac{\partial \eta_i^k}{\partial x_j} + \frac{\partial \eta_i^k}{\partial x^i} \right)(x_0) \frac{dx^i}{dt}(t_0) \frac{dx^j}{dt}(t_0).$$

Finally,

$$\varphi''(t_0) = d^2 f(x_0) - \frac{1}{2} \sum_{a=1}^{p} \lambda_a \left(\frac{\partial \omega_i^a}{\partial x^j} + \frac{\partial \omega_j^a}{\partial x^i} \right)(x_0) \frac{dx^i}{dt}(t_0) \frac{dx^j}{dt}(t_0) -$$

$$-\frac{1}{2} \sum_{k=1}^{n} \mu_k \left(\frac{\partial \eta_i^k}{\partial x^j} + \frac{\partial \eta_j^k}{\partial x^i} \right)(x_0) \frac{dx^i}{dt}(t_0) \frac{dx^j}{dt}(t_0) > 0.$$

Hence

$$\varphi(t) - \varphi(t_0) = \frac{1}{2} \varphi''(t_0)(t - t_0)^2 + \mathcal{O}((t - t_0)^2)$$

or $f(\alpha(t)) \geq f(x_0)$, $\forall t \in (t_0 - \varepsilon, t_0 + \varepsilon)$. □

Theorem 5. *Let $f : D \to \mathbb{R}$ be a C^2 function and (ω, η, M) a system of restrictions on D with ω, η of class C^1. Let $x_0 \in M$ be a point at which it is satisfied the regularity condition upon rank or, more generally, the Kuhn-Tucker regularity condition. Suppose x_0 is a minimum point of f constrained by (ω, η, M). Then the restriction of the quadratic form Ω in the Theorem 2. 6 to the subspace (6) is positive semidefinite.*

Proof. From Theorem 2 or Theorem 3 it follows that the condition i) of Theorem 2 is satisfied. By absurdum, we suppose the existence of a nonzero vector v in the subspace (*) such that $\Omega(v, v) < 0$. From the regularity condition, it follows the existence of a parametrized curve $\alpha \in \Upsilon(x_0)$ such that $\alpha(t_0) = x_0$ and $\alpha'(t_0) = v$. Considering the function φ from the previous proof, and following the same computation, we find $f(\alpha(t)) \leq f(x_0)$, $\forall t \in (-t_0 - \varepsilon, t_0 + \varepsilon)$, which contradicts the hypothesis that x_0 is a minimum point constrained by (ω, η, M). □

In a classical extremum problem with constraints defined by equations and inequations, the point constraints and the velocity constraints are correlated. On the other hand, the previous Theorems shows that the finding of constrained extremum points is based essentially only on velocity constraints. This permits ourselves that in case of a classical extremum problem to renounce to the point constraints, replacing the initial problem with a family of extremum problems having the same velocity constraints.

3 Illustrating Examples

Many equilibrium problems from economics and important applied problems from diverse engineering fields can be profitably formulated via curve selector theory. That is why, here we give details using significant examples.

Example 1. The Pfaff form $\omega = dx - zdy$ verifies the rank regularity condition in the Definition 3, and implicitly the Kuhn-Tucker regularity conditions. Let us find effectively the admissible curves asked by these last regularity conditions. Let $v = (v_1, v_2, v_3) \in \mathbb{R}^3 \setminus \{(0,0,0)\}$, with $\langle \omega, v \rangle \geq 0$, i.e., $v_1 - z_0 v_2 \geq 0$. Then for the curve $\alpha : \mathbb{R} \to \mathbb{R}^3$, $\alpha(t) = (x(t), y(t), z(t))$, $x(t) = x_0 + v_1 t + z_0 t^2, y(t) = y_0 + v_2 t + t^2, z(t) = z_0$, we have

$$\langle \omega(\alpha(t)), \alpha'(t) \rangle = v_1 - z_0 v_2 \geq 0,$$

$$\int_0^t \langle \omega(\alpha(t)), \alpha'(t) \rangle dt = (v_1 - z_0 v_2)t \geq 0, \quad \forall t \geq 0,$$

i.e., $\alpha \in \Upsilon(x_0, y_0, z_0)$.

The semispaces defined by the condition $\langle \omega, v \rangle \geq 0$ in the tangent space to R^3, at the point (x_0, y_0, z_0), are of the form $E \times \mathbb{R}$, where E is a semiplane of the plane (v_1, v_2).

Let us apply the Theorem 2 to the objective function $f(x, y, z) = (x^2 + y^2 + z^2)/2$ constrained by (ω, M), where $M = S \cup bS$ is an arbitrary point constraint set. The system $df = \lambda\omega$ is $x = \lambda$, $y = -\lambda z$, $z = 0$. It follows that the set of susceptible points for extremum is a part of the axis Ox defined by $(\lambda, 0, 0)$ with $\lambda \geq 0$. For $\lambda = 0$ one obtains the free extremum $(0, 0, 0)$ which is also a global minimum point. For $\lambda \neq 0$ the quadratic form mentioned at the step (ii) of the Theorem 2 can be written

$$\Omega = dx^2 + dy^2 + dz^2 + \lambda dydz,$$

and on the subspace $\omega(x_0, y_0, z_0) = 0$, i.e., $dx = 0$, it becomes

$$\Omega_0 = dy^2 + \lambda dydz + dz^2.$$

For $\lambda \in [0, 2)$, the quadratic form Ω_0 is positive definite and hence the point $(\lambda, 0, 0)$ is a minimum point if and only if it belongs to bS. For $\lambda > 2$, the quadratic form Ω_0 is not definite. According the Theorem 5 the point $(\lambda, 0, 0)$ cannot be extremum point. What happens for $\lambda = 2$?

Example 2. Let $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x, y, z) = x + y + z$ and the inequality constraints $(\omega^1, S_1 \cup bS_1)$, $(\omega^2, S_2 \cup bS_2)$, where

$$\omega^1 = -2xdx - 2ydy - 2zdz, \quad \omega^2 = -2xdx - 2ydy + dz,$$

and $S_1 \cup bS_1$, $S_2 \cup bS_2$ are arbitrary point constraints. For any point $P(x, y, z)$ with $z \neq -1/2$, we have rank $(\omega^1, \omega^2) = 2$. In such points, Theorem 2 is

applicable. Let us show that at points $P(x_0, y_0, -1/2)$ it is satisfied the Kuhn-Tucker regularity condition, and consequently Theorem 3 is applicable. Let $g^1(x,y,z) = -x^2 - y^2 - z^2 + C_1$ with $C_1 = x_0^2 + y_0^2 + 1/4$ and $g^2(x,y,z) = -x^2 - y^2 + z + C_2$ with $C_2 = x_0^2 + y_0^2 + 1/2$. It is obvious that $\omega^1 = dg^1$, $\omega^2 = dg^2$, $g^1(P) = 0$, $g^2(P) = 0$, $\omega^1(P) = \omega^2(P)$. Also, we remark that the set $\{(x,y,z) \in \mathbb{R}^3 | g^1(x,y,z) \geq 0\}$ is contained $\{(x,y,z) \in \mathbb{R}^3 | g^2(x,y,z) \geq 0\}$. Let $v \in \mathbb{R}^3 \setminus \{(0,0,0)\}$ such that $\langle \omega^1(P), v\rangle = \langle \omega^2(P), v\rangle \geq 0$. We can find a parametrized C^2 curve $\alpha : I \to \mathbb{R}^3$ such that $\alpha(0) = 0$, $\alpha'(0) = v$ and $g^1(\alpha(t)) \geq 0$, $\forall t \in I$. It follows $g^2(\alpha(t)) \geq 0$, $\forall t \in I$ and hence

$$\int_0^t \langle \omega^1(u), \alpha'(u)\rangle du \geq 0, \quad \int_0^t \langle \omega^2(u), \alpha'(u)\rangle du \geq 0, \quad \forall t \in I.$$

So, $\alpha \in \Upsilon(P)$. Let $M = (S_1 \cup bS_1) \cap (S_2 \cup bS_2)$, $bS = M \cap (bS_1 \cup bS_2)$ and $S = M \setminus bS$. From Theorems 2 and 3, we decide that the minimum constrained points of f are between the points of M which verify the relation $df(x) = \lambda_1 \omega^1(x) + \lambda_2 \omega^2(x)$, with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$. Consequently

$$1 + 2x(\lambda_1 + \lambda_2) = 0,\ 1 + 2y(\lambda_1 + \lambda_2) = 0,\ 1 + 2\lambda_1 z - \lambda_2 = 0$$

for $\lambda_1, \lambda_2 \in [0, \infty)$. Let $P(x,y,z)$ be a solution of this system with $P \in M$. Let Ω be the quadratic form of Theorem 2. In our case $\Omega = 2(\lambda_1 + \lambda_2)dx^2 + 2(\lambda_1 + \lambda_2)dy^2 + 2\lambda_1 dz^2$. Let us denote by V the subspace described in the Theorem 2 which represents the velocity constraints.

Case 1. Suppose $\lambda_1 + \lambda_2 \neq 1$. It follows $\lambda_1 + \lambda_2 > 0$ and $\lambda_1 > 0$, since, in the contrary situation, the system do not have solutions. We obtain

$$P\left(-\frac{1}{2(\lambda_1 + \lambda_2)}, -\frac{1}{2(\lambda_1 + \lambda_2)}, \frac{\lambda_2 - 1}{2\lambda_1}\right).$$

The quadratic form $\Omega(P)$ is positive definite. The point P is a minimum point if and only if $P \in bS_1$.

Case 2. Suppose $\lambda_1 + \lambda_2 = 1$. If $\lambda_1 = 0$, we find $P(-1/2, -1/2, \lambda)$ with $\lambda \in \mathbb{R}$ and $\lambda_2 = 1$. In this case the subspace V is defined by the equation $dz = dx + dy$. The quadratic form $\Omega | V$ is positive definite, so P is a minimum point, if and only if $P \in bS_2$. If $\lambda_1 > 0$, we obtain $P(-1/2, -1/2, -1/2)$. In this case the quadratic form Ω is positive definite, and consequently P is a minimum point if and only if $P \in bS_1$. The previous theorems show that we have not other minimum points.

Since ω^1, ω^2 are exact Pfaff forms, the classic case with inequality constraints can be considered as a particular case in which the point constraints and the velocity constraints are correlated. Let $g^1(x,y,z) = -x^2 - y^2 - z^2$, $g^2(x,y,z) = -x^2 - y^2 + z$. From the previous results it appears that $P(-1/2, -1/2, \lambda)$, $\lambda \in \mathbb{R}$ is a minimum point constrained by $g^1(x,y,z) \geq C_1$, $g^2(x,y,z) \geq C_2$ if and only if $g^1(P) \geq C_1$ and $g^2(P) \geq C_2$.

Example 3. Let $f(x,y,z) = x^2 + y^2 - z$ subject to the constraints (ω, M), where $\omega = xdy - zdz$, and the set $M = S \cup bS$ is arbitrary. One observes that,

at points of the form $Q(0, y_0, 0)$, the rank regularity condition of Definition 3 is not satisfied. Let us show that in these points, Kuhn-Tucker regularity condition is not even satisfied. Let $v = (v_1, v_2, v_3)$ be a nonzero vector satisfying $(\omega(Q), v) \geq 0$. Since $\omega(Q) = 0$, the vector v is arbitrary. Let us show that there exists v such that any C^2 parametrized curve α tangent to v at Q is not an admissible curve. Since $\alpha(0) = Q$, $\alpha'(0) = v$, we can write $x = v_1 t + t^2 a(t)$, $y = y_0 + v_2 t + t^2 b(t)$, $z = v_3 t + t^2 c(t)$, where a, b, c are continuous functions in a neighborhood of the origin. It follows

$$I(t) = \int_0^t \langle \omega(\alpha(u)), \alpha'(u) \rangle du = \frac{t^2}{2}(v_1 v_2 - v_3^2) + t^3 \phi(t),$$

with $\phi(t) \to 0$ as $t \to 0$. If $(v_1 v_2 - v_3^2) < 0$, then $I(t) \leq 0$, $\forall t \in (-\varepsilon, \varepsilon)$. Applying the Lagrange multipliers rule, we obtain the system

$$x = 0,\ 2y - \lambda y = 0,\ -1 + \lambda z = 0.$$

For $\lambda \neq 0$, we have the solutions $P(0, 0, 1/\lambda)$. The quadratic form of Theorem 2 is

$$\Omega = 2dx^2 + 2dy^2 + \lambda dx^2 - \lambda dx dy.$$

The restriction of Ω to the subspace of velocity constraints is $2dx^2 + 2dy^2 - \lambda dx dy$, having the determinant $4 - \lambda^2/4$. Hence, for any $\lambda \in (0, 4)$, the point $P(0, 0, 1/\lambda)$ is a minimum point if and only if $P \in bS$. From Theorem 5 it follows that $P \in M$, with $\lambda \in (-\infty, 0) \cup (4, \infty)$, cannot be a minimum point. For $\lambda = 4$, Theorem 2 cannot be applied. By a direct evaluation we shall show that $P(0, 0, 1/4)$ cannot be a minimum point. For that we use the integral curve α defined by $x(t) = z(t)z'(t)$, $y(t) = t$, $z(t) = 1/4 + 2t^2 + at^3$, $\alpha(0) = P$. Hence $\alpha \in \Upsilon(P)$, i.e., it is an admissible curve. We obtain $f(\alpha(t)) - f(P) = at^3/2 + t^3\phi(t)$, with $\phi(t) \to 0$ as $t \to 0$. For $a < 0$, we have $f(\alpha(t)) \leq f(P)$, $\forall t \in [0, \varepsilon)$, i.e., P cannot be a minimum point for f constrained by (ω, M). Since the Kuhn-Tucker regularity conditions are not satisfied at points $Q(0, y_0, 0)$, the previous theorems are not applicable. Let us show that these points cannot be minimum points. For that we use the integral curve

$$\alpha:\ x(t) = a^2 t,\ y(t) = y_0 + t,\ z(t) = at, \quad \alpha(0) = Q.$$

It follows

$$f(\alpha(t)) - f(Q) = t^2(a^4 + 1) + t(2y_0 - a).$$

For $2y_0 < a$, we find $f(\alpha(t)) \leq f(Q)$, $\forall t \in [0, \varepsilon)$, i.e., Q is not a minimum point of f constrained by (ω, M). Consequently, the only points that can be minimum points are those mentioned above.

4 Totally Geodesic Submanifold Described by Lagrange or Kuhn-Tucker Parameters

Let $f : D \to \mathbb{R}$ be a C^3 function and (ω, M) be a system of inequality constraints, where the system ω is reduced to a Pfaff form. The system

$$\frac{\partial f}{\partial x^i}(x) = \lambda \omega_i(x), i = 1, ..., n$$

in the unknown (x, λ), describes the constrained critical points in the problem, that is, the *catastrophe set.* Generally, via the implicit function theorem, the solution is a curve $C : x = x(\lambda),\ \lambda \in (0, \infty)$. Let us show that C is a decreasing curve in a neighborhood of a point x_0, if the matrix of elements

$$g_{ij}(x) = \frac{\partial^2 f}{\partial x^i \partial x^j}(x) - \frac{\lambda}{2}(\frac{\partial \omega_i}{\partial x^j} + \frac{\partial \omega_j}{\partial x^i})(x),\ i, j = 1, ..., n$$

is positive definite (Riemannian metric around the point x_0). For that we take the derivative along C, and we obtain

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) - \frac{\lambda}{2}(\frac{\partial \omega_i}{\partial x^j} + \frac{\partial \omega_j}{\partial x^i})(x) - \frac{\lambda}{2}(\frac{\partial \omega_i}{\partial x^j} - \frac{\partial \omega_j}{\partial x^i})(x)\right)\frac{dx^j}{d\lambda} = \frac{1}{\lambda}\frac{\partial f}{\partial x^i}(x),$$

(here, and throughout this section, it is used the Einstein convention of summation), where the matrix of elements

$$a_{ij}(x) = \frac{\partial^2 f}{\partial x^i \partial x^j}(x) - \frac{\lambda}{2}(\frac{\partial \omega_i}{\partial x^j} + \frac{\partial \omega_j}{\partial x^i})(x) - \frac{\lambda}{2}(\frac{\partial \omega_i}{\partial x^j} - \frac{\partial \omega_j}{\partial x^i})(x)$$

is not symmetric but still positive definite around the point x_0.

Using the reparametrization $\frac{du}{d\lambda} = -\frac{1}{\lambda}$, i.e., $\lambda = e^{-u},\ u \in (-\infty, \infty)$, and the inverse $(a^{ij}(x))$ of the matrix $(a_{ij}(x))$, the previous system can be written as

$$\frac{dx^j}{du} = -a^{ij}(x)\frac{\partial f}{\partial x^i}(x).$$

Consequently, the curve $C : x = x(u),\ u \in (-\infty, \infty)$ is a "minus gradient like line" around the minimum point x_0. In other words, the parameter λ indicates a rate of decreasing.

Taking again the derivative along a solution, the first ODEs system

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) - e^{-u}\frac{\partial \omega_i}{\partial x^j}(x)\right)\frac{dx^j}{du} = -\frac{\partial f}{\partial x^i}(x)$$

is prolonged to the second order ODEs system

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) - e^{-u}\frac{\partial \omega_i}{\partial x^j}(x)\right)\frac{d^2 x^j}{du^2} + \left(\frac{\partial^3 f}{\partial x^i \partial x^j \partial x^k}(x) - e^{-u}\frac{\partial \omega_i}{\partial x^j \partial x^k}(x)\right)\frac{dx^j}{du}\frac{dx^k}{du}$$

$$= -\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) + e^{-u}\frac{\partial \omega_i}{\partial x^j}(x)\right)\frac{dx^j}{du}.$$

Since

$$\left(\frac{\partial^3 f}{\partial x^i \partial x^j \partial x^k}(x) - e^{-u}\frac{\partial \omega_i}{\partial x^j \partial x^k}(x)\right) = 2g_{ijk}$$

are the Christoffel symbols produced by the tensor g_{ij}, it follows that the curve $C : x = x(u),\ u \in (-\infty, \infty)$ is a reparametrized geodesic of the Otsuki connection [16] $(g^{ki}a_{ij}, G^k_{ij})$, around the point x_0.

Theorem 6. *Suppose $g_{ij}(x)$ is a Riemannian metric around the point x_0. One has:*

i) If f is a C^2 function with the constrained minimum point x_0, then the curve $C : x = x(u),\ u \in (-\infty, \infty)$ is a minus gradient line.

ii) If f is a C^3 function with the constrained minimum point x_0, then the curve $C : x = x(u),\ u \in (-\infty, \infty)$ is a reparametrized geodesic of the Otsuki connection $(g^{ki}a_{ij}, G^k_{ij})$.

Let $f : D \to \mathbb{R}$ be a C^3 function and (ω, M) be a system of p inequality constraints. The system

$$\frac{\partial f}{\partial x^i}(x) = \lambda_a \omega_i^a(x), i = 1, ..., n; a = 1, ..., p$$

in the unknown (x, λ), describes the constrained critical points in the problem. Generally, via the implicit function theorem, the solution is a p-dimensional submanifold $N : x = x(\lambda),\ \lambda = (\lambda_1, ..., \lambda_p),\ \lambda_a \in (0, \infty)$.

Let us show that this is a totally geodesic submanifold with respect to an Otsuki connection. For that, taking the derivative with respect to λ_b we find

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) - \lambda_a \frac{\partial \omega_i^a}{\partial x^j}(x)\right)\frac{\partial x^j}{\partial \lambda_b} = \omega_i^b(x).$$

Automatically, it appear a Riemannian metric similar to $g_{ij}(x)$ and a matrix similar to $(a_{ij}(x))$. The transvection with $-\lambda = (-\lambda_1, ..., -\lambda_p)$ shows that this vector gives a descent direction.

Taking again the derivative with respect to λ_c, we find

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(x) - \lambda_a \frac{\partial \omega_i^a}{\partial x^j}(x)\right)\frac{\partial^2 x^j}{\partial \lambda_b \partial \lambda_c} =$$

$$-\left(\frac{\partial^3 f}{\partial x^i \partial x^j \partial x^k}(x) - \lambda_a \frac{\partial^2 \omega_{ik}^a}{\partial x^j}(x)\right)\frac{\partial x^j}{\partial \lambda_b}\frac{\partial x^k}{\partial \lambda_c} + \frac{\partial \omega_i^c}{\partial x^j}\frac{\partial x^j}{\partial \lambda_b} + \frac{\partial \omega_i^b}{\partial x^j}\frac{\partial x^j}{\partial \lambda_c}.$$

Having in mind the previous explanations, we find

Theorem 7. *i) If f is a C^2 function with the constrained minimum point x_0, then the vector $-\lambda = (-\lambda_1, ..., -\lambda_p)$ gives a descent direction.*
ii) If f is a C^3 function with the constrained minimum point x_0, then the submanifold $N : x = x(\lambda)$, $\lambda = (\lambda_1, ..., \lambda_p)$, $\lambda_a \in (0, \infty)$ is a reparametrized totally geodesic submanifold of the Otsuki connection $(g^{ki}a_{ij}, G^k_{ij})$ (cf. [16]).

Open problem. Study the *singularity set* (consisting of singular critical points of the catastrophe submanifold), and its projection in the set of parameters (*bifurcation set*).

References

1. C. Udrişte and O. Dogaru. Mathematical Programming Problems with Nonholonomic Constraints. Seminarul de Mecanică, Univ. of Timişoara, Fac. de Ştiinţe ale Naturii, vol. 14, 1988.
2. C. Udrişte and O. Dogaru. Extrema with nonholonomic constraints. Buletinul Institutului Politehnic Bucureşti, Seria Energetică, 50:3-8, 1988.
3. C. Udrişte and O. Dogaru. Extreme condiţionate pe orbite. Sci. Bull., 51:3-9, 1991.
4. C. Udrişte and O. Dogaru. Convex nonholonomic hypersurfaces. In: G. Rassias (ed), Math. Heritage of C.F. Gauss, World Scientific, pp. 769-784, 1991.
5. C. Udrişte, O. Dogaru and I. Ţevy. Sufficient conditions for extremum on differentiable manifolds. Sci. Bull. P.I.B., Elect. Eng., 53(3-4):341-344, 1991.
6. C. Udrişte, O. Dogaru and I. Ţevy. Extremum points associated with Pfaff forms. Tensor, N.S., 54:115-121, 1993.
7. C. Udrişte, O. Dogaru and I. Ţevy. Open problems in extrema theory. Sci. Bull. P.U.B., Series A, 55(3-4):273-277, 1993.
8. O. Dogaru, I. Ţevy and C. Udrişte. Extrema constrained by a family of curves and local extrema. JOTA,97(3):605-621, 1998.
9. C. Udrişte, O. Dogaru and I. Ţevy. Extrema constrained by a Pfaff system. In: T. Gill, K. Liu, and E. Trell (Eds.), Fundamental Open Problems in Science at the End of Millenium, vol. I-III, Hadronic Press, Inc., Palm Harbor, FL, pp. 559-573, 1999.
10. O. Dogaru and I. Ţevy. Extrema constrained by a family of curves. Proc. Workshop on Global Analysis, Differential Geometry and Lie Algebras, 1996, Gr. Tsagas (Ed.), Geometry Balkan Press, pp. 185-195, 1999.
11. O. Dogaru and V. Dogaru. Extrema constrained by C^k curves. Balkan J. Geometry and Its Applications, 4(1):45-52, 1999.
12. C. Udrişte. Geometric Dynamics. Kluwer Academic Publishers, 2000.
13. C. Udrişte, O. Dogaru and I. Ţevy. Extrema with Nonholonomic Constraints. Monographs and Textbooks 4, Geometry Balkan Press, 2002.
14. C. Udrişte, O. Dogaru, M. Ferrara, I. Ţevy. Pfaff inequalities and semi-curves in optimum problems. In: G.P. Crespi, A. Guerragio, E. Miglierina, M. Rocca (Eds.), Recent Advances in Optimization, DATANOVA, pp.191-202, 2003.
15. C. Udrişte, O. Dogaru, M. Ferrara, I. Ţevy. Extrema with constraints on points and/or velocities. Balkan J. Geometry and Its Applications, 8(1):115-123, 2003.
16. T. Otsuki. General connections. Math. J. Okayama Univ. 32:227-242, 1990.

A Note on Error Estimates for some Interior Penalty Methods *

A. F. Izmailov[1] and M. V. Solodov[2]

[1] Moscow State University, Faculty of Computational Mathematics and Cybernetics, Department of Operations Research, Leninskiye Gori, GSP-2, 119992 Moscow, Russia. izmaf@ccas.ru

[2] Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, RJ 22460-320, Brazil. solodov@impa.br

Summary. We consider the interior penalty methods based on the logarithmic and inverse barriers. Under the Mangasarian-Fromovitz constraint qualification and appropriate growth conditions on the objective function, we derive computable estimates for the distance from the subproblem solution to the solution of the original problem. Some of those estimates are shown to be sharp.

1 Introduction and Preliminaries

We consider the optimization problem

$$\begin{array}{l} \text{minimize} \quad f(x) \\ \text{subject to } x \in D = \{x \in \mathbb{R}^n \mid G(x) \leq 0\}, \end{array} \tag{1}$$

where the set

$$D^0 = \{x \in \mathbb{R}^n \mid G(x) < 0\}$$

is assumed to be nonempty. Under the stated assumption, one of the classical schemes [3] for solving problem (1) is the *interior penalty (or barrier) method.* It consists in replacing (1) by a sequence of (in some sense, unconstrained) subproblems of the form

$$\begin{array}{l} \text{minimize} \quad \varphi_\sigma(x) \\ \text{subject to } x \in D^0, \end{array} \tag{2}$$

where $\sigma > 0$ is the penalty (barrier) parameter, and

* Research of the first author is supported by Russian Foundation for Basic Research Grant 04-01-00341 and RF President's Grant NS-1815.2003.1 for the support of leading scientific schools. The second author is supported in part by CNPq Grants 300734/95-6 and 471780/2003-0, by PRONEX–Optimization, and by FAPERJ.

$$\varphi_\sigma : D^0 \to \mathbb{R}, \quad \varphi_\sigma(x) = f(x) + \sigma \sum_{i=1}^{m} b(-G_i(x)),$$

with $b : \mathbb{R}_{++} \to \mathbb{R}_{++}$ being the barrier function. The most popular are the logarithmic barrier

$$b(t) = -\ln t, \tag{3}$$

and the inverse barrier

$$b(t) = 1/t. \tag{4}$$

For basic convergence results of this type of methods, we refer the reader to [3, 10, 11]. Here, we only mention that to ensure convergence, the barrier parameter σ must be driven to zero.

For each $\sigma > 0$, let x_σ be a solution of (2). Let $\bar{x}$ be a solution of (1), and suppose that $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. We note that our analysis can easily treat the case when there is a given sequence $\{\sigma_k\} \subset \mathbb{R}_+$ such that $\sigma_k \to 0+$, and the corresponding sequence $\{x^k\} = \{x_{\sigma^k}\}$ is convergent to $\bar{x}$. The modifications to cover this case are straightforward. We assume that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable, and the derivatives of f and G are continuous at $\bar{x}$. For some results, the problem data will further be assumed twice differentiable.

In this paper, we are interested in estimates of the distance from x_σ to $\bar{x}$ via some computable quantity, that is, in *error bounds* of the form

$$\|x_\sigma - \bar{x}\| = O(r(x_\sigma, \sigma)), \tag{5}$$

where $r : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}_+$ is some (easily) computable function such that, at least, $r(x_\sigma, \sigma) \to 0$ as $\sigma \to 0+$ and $x_\sigma \to \bar{x}$. Moreover, it is desirable that the bound (5) should be sharp, i.e., not improvable under the given assumptions. In some cases, it is possible to eliminate the dependence on x_σ in the right-hand side of (5), and then (5) can be considered as a convergence rate estimate. But in any case, computable error bounds are very useful. In particular, they provide reliable stopping tests for the related algorithms.

Denote the set of Lagrange multipliers associated with $\bar{x}$ by

$$\mathcal{M} = \mathcal{M}(\bar{x}) = \left\{ \mu \in \mathbb{R}^m \;\middle|\; \frac{\partial L}{\partial x}(\bar{x}, \mu) = 0,\ \mu \geq 0,\ \langle \mu, G(\bar{x}) \rangle = 0 \right\},$$

where

$$L(x, \mu) = f(x) + \langle \mu, G(x) \rangle, \quad x \in \mathbb{R}^n,\ \mu \in \mathbb{R}^m,$$

is the Lagrangian of problem (1). Recall that the *linear independence constraint qualification* (LICQ) for problem (1) at $\bar{x}$ consists of saying that $G_i'(\bar{x})$, $i \in A$, are linearly independent, where $A = A(\bar{x}) = \{i = 1, \ldots, m \mid G_i(\bar{x}) = 0\}$ is the set of constraints active at $\bar{x}$. Under LICQ, $\mathcal{M}$ is necessarily a singleton. The weaker *Mangasarian–Fromovitz constraint qualification* (MFCQ) for problem (1) at $\bar{x}$ consists of saying that there exists $\bar{\xi} \in \mathbb{R}^n$ such that

$G'_A(\bar{x})\bar{\xi} < 0$. If this condition holds, then $\mathcal{M}$ is necessarily a nonempty polyhedral and compact set. We say that *strict complementarity* holds at $\bar{x}$ if there exists some $\bar{\mu} \in \mathcal{M}$ such that $\bar{\mu}_A > 0$.

Our goal is to obtain computable error bounds under assumptions that do not use strict complementarity and do not invoke any CQ-type conditions stronger than MFCQ (so that, in particular, $\mathcal{M}$ need not be a singleton).

We note that under appropriate assumptions, it can be possible to estimate the distance from an *arbitrary* $x \in \mathbb{R}^n$ to $\bar{x}$, independently of any specific algorithmic framework, i.e., regardless of how x was produced or chosen (sometimes these estimates are of a primal-dual nature; see below). We refer the reader to [12] for a survey of algorithm-independent error bounds and their applications. On the other hand, algorithm-based error bounds can sometimes be established under weaker or different assumptions than their algorithm-independent counterparts (e.g., [7], available at http://www.preprint.impa.br/Shadows/SERIE_A/2004/303.html).

As an algorithm-independent error bound relevant in our context, we mention the following result, based on [6, Lemma 2] and [4, Theorem 2]. Suppose that with some $\bar{\mu} \in \mathcal{M}$, the following second-order sufficient optimality condition holds:

$$\frac{\partial^2 L}{\partial x^2}(\bar{x}, \bar{\mu})[\xi, \xi] > 0 \quad \forall \xi \in C \setminus \{0\},$$

where

$$C = C(\bar{x}) = \{\xi \in \mathbb{R}^n \mid G'_A(\bar{x})\xi \leq 0, \ \langle f'(\bar{x}), \xi \rangle \leq 0\} \tag{6}$$

is the critical cone of problem (1) at $\bar{x}$. Then for $(x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m$ close enough to $(\bar{x}, \bar{\mu})$, it holds that

$$\|x - \bar{x}\| = O(r(x, \mu)),$$

where

$$r(x, \mu) = \left\| \left(\frac{\partial L}{\partial x}(x, \mu), \min\{\mu, -G(x)\} \right) \right\|,$$

with the minimum taken componentwise. Note that the quantity $r(x, \mu)$ is the *natural residual* of the *Karush-Kuhn-Tucker system*

$$\frac{\partial L}{\partial x}(x, \mu) = 0, \quad \mu \geq 0, \quad G(x) \leq 0, \quad \langle \mu, G(x) \rangle = 0,$$

which characterizes stationary points of problem (1) and the associated multipliers. This result does not rely on any CQ. Note, however, that if we are in some algorithmic framework, then for this result to be applicable, the given primal-dual sequence generated by the method has to converge to the specific $(\bar{x}, \bar{\mu})$ which satisfies the above condition. In our results, no assumptions about convergence of the dual part of the sequence are necessary, as long as the needed growth conditions (related to sufficient optimality conditions, see below) are satisfied. We refer the reader to [8] for other algorithm-independent

error bounds under assumptions which subsume some CQ-type conditions, as well as for a detailed discussion and comparisons of error bounds and regularity conditions for KKT systems.

The results presented in this paper are strongly inspired by [14] and [5], where the log-barrier method has been analyzed. The analysis in these works is based on the following assumptions: MFCQ, the strict complementarity condition, and the second-order sufficient optimality condition in the following strong form:

$$\frac{\partial^2 L}{\partial x^2}(\bar{x}, \mu)[\xi, \xi] > 0 \quad \forall \mu \in \mathcal{M}, \ \forall \xi \in C \setminus \{0\}. \tag{7}$$

In [14], the case of violation of strict complementarity (but with the other two assumptions satisfied) is discussed as well. The assertions to be stated below are weaker than those in [14] and [5], but our assumptions are different. We also assume MFCQ, but never strict complementarity. When we assume second-order sufficiency (actually, we assume a certain quadratic growth condition, but in the given setting it is equivalent to second-order sufficiency), it is in a form significantly weaker than (7), see the discussion below. On the other hand, for some results we assume convexity of the objective function and/or of the constraints.

Our analysis relies on the so-called growth conditions, which we discuss next. We say that the *linear growth* condition is satisfied at $\bar{x}$ if there exist $\gamma > 0$ and a neighborhood U of $\bar{x}$ such that

$$f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\| \quad \forall x \in D \cap U. \tag{8}$$

As is well known (see, e.g., [2, Lemma 3.24]), the linear growth is guaranteed by the *first-order sufficient condition* (FOSC)

$$C = \{0\}. \tag{9}$$

Moreover, the two conditions are equivalent provided MFCQ holds at $\bar{x}$.

We say that the *quadratic growth* condition is satisfied at $\bar{x}$ if there exist $\gamma > 0$ and a neighborhood U of $\bar{x}$ such that

$$f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\|^2 \quad \forall x \in D \cap U. \tag{10}$$

Obviously, quadratic growth is a weaker property than linear growth. If $\mathcal{M} \neq \emptyset$, then the following *second-order sufficient condition* (SOSC) becomes relevant:

$$\forall \xi \in C \setminus \{0\} \ \exists \mu \in \mathcal{M} \text{ such that } \frac{\partial^2 L}{\partial x^2}(\bar{x}, \mu)[\xi, \xi] > 0. \tag{11}$$

According to [2, Theorem 3.70], the latter condition is sufficient for the quadratic growth, and equivalent to it if MFCQ holds at $\bar{x}$.

We emphasize that if $\mathcal{M}$ is not a singleton, the second order sufficient optimality condition (11) is significantly weaker than (7).

The approach we use in this paper is quite similar to the one employed in sensitivity theory for deriving Lipschitz and Hölder stability of optimal solutions (see [2]). However, the context of barrier methods possesses a special feature which can (and should) be taken into account: the perturbed solutions remain feasible for the original problem.

We start with considering the case of log-barrier in Section 2. For this barrier, it is possible to obtain convergence rate estimates where the right-hand side in (5) does not depend on x_σ. This is not the case for the inverse barrier, considered in Section 3, where the right-hand side in (5) involves x_σ. Nevertheless, it still gives a computable estimate.

2 Error Estimates for the Log-Barrier Method

Throughout this section, φ_σ is defined with the logarithmic barrier (3).

For each $\sigma > 0$, denote

$$\mu_\sigma = -\sigma(1/G_1(x_\sigma), \ldots, 1/G_m(x_\sigma)) > 0. \tag{12}$$

By direct computation,

$$\langle \mu_\sigma, G(x_\sigma) \rangle = -m\sigma, \tag{13}$$

and by the first-order necessary optimality conditions for problem (2) at x_σ,

$$\frac{\partial L}{\partial x}(x_\sigma, \mu_\sigma) = \varphi'_\sigma(x_\sigma) = 0. \tag{14}$$

We start with the case when (1) is a convex minimization problem. The following result is well-known (as we were informed by a referee, it probably first appeared in [1]). We include its short proof, for the sake of completeness.

Proposition 1. *Let f and G_i, $i = 1, \ldots, m$, be convex. For $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3). Then it holds that*

$$f(x_\sigma) \leq \inf_{x \in D} f(x) + m\sigma. \tag{15}$$

Proof. Associated to (1) is its *Wolfe dual* problem (e.g., see [9])

$$\begin{aligned} &\text{maximize } L(x, \mu) \\ &\text{subject to } (x, \mu) \in \Delta = \{(x, \mu) \in \mathbb{R}^n \times \mathbb{R}^m \mid \tfrac{\partial L}{\partial x}(x, \mu) = 0,\ \mu \geq 0\}. \end{aligned}$$

By weak duality [9, Theorem 8.1.3], it holds that

$$\sup_{(x, \mu) \in \Delta} L(x, \mu) \leq \inf_{x \in D} f(x). \tag{16}$$

Observe that, by (12) and (14), $(x_\sigma, \mu_\sigma) \in \Delta$. By (13), we obtain that

$$L(x_\sigma, \mu_\sigma) = f(x_\sigma) - m\sigma.$$

The assertion now follows from (16). □

Note that in the above, solvability of problem (1) is not needed. But in our setting, estimate (15) gives

$$f(x_\sigma) \leq f(\bar{x}) + O(\sigma). \tag{17}$$

By (17), using the feasibility of x_σ in the original problem (1) and relation (8) (in the case of linear growth) or (10) (in the case of quadratic growth), we can immediately obtain convergence rate estimates, stated in Theorem 1 below. We note that estimate (19) of this theorem was obtained in [10, 11] assuming strong convexity of the Lagrangian for some fixed multiplier, which is stronger than the quadratic growth condition. Overall, we do not have a direct reference for Theorem 1, but we have no doubt that it is known.

Theorem 1. *Let f and G_i, $i = 1, \ldots, m$, be convex. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then the following assertions are valid:*

(i) If the linear growth condition is satisfied at $\bar{x}$, then

$$\|x_\sigma - \bar{x}\| = O(\sigma). \tag{18}$$

(ii) If the quadratic growth condition is satisfied at $\bar{x}$, then

$$\|x_\sigma - \bar{x}\| = O(\sigma^{1/2}). \tag{19}$$

The estimates obtained in Theorem 1 can be regarded as pure convergence rate estimates. Moreover, these estimates are sharp, even under LICQ, as demonstrated by the following simple examples.

Example 1. Let $n = m = 1$, $f(x) = x$, $G(x) = -x$ (linear functions). Clearly, $\bar{x} = 0$ is the unique solution of problem (1), and moreover, LICQ and FOSC (9) (hence, the linear growth condition) are satisfied at $\bar{x}$. It can be directly verified that for each $\sigma > 0$, the unique solution of subproblem (2) with the barrier function defined in (3) is given by $x_\sigma = \sigma$ and the estimate (18) is sharp. Moreover, $f(x_\sigma) = x_\sigma = \sigma$, and the estimate (17) is sharp as well.

Example 2. Let $n = m = 1$, $f(x) = x^2/2$, $G(x) = -x$ (convex functions). Note that the constraint in this example is the same as in Example 1. Evidently, $\bar{x} = 0$ is the unique solution of problem (1), and moreover, LICQ and SOSC (11) (hence, the quadratic growth condition, but not the linear growth condition!) are satisfied at $\bar{x}$. Note that the strict complementarity condition does not hold in this example. It can be directly verified that for each $\sigma > 0$, the unique solution of subproblem (2) with the barrier function defined in (3) is given by $x_\sigma = \sigma^{1/2}$ and the estimate (19) is sharp. Moreover, $f(x_\sigma) = x_\sigma^2/2 = \sigma/2$, and the estimate (17) is sharp as well.

In Proposition 1 and Theorem 1 we do not assume that MFCQ holds at $\bar{x}$, but MFCQ is implicitly subsumed in these results. As is well known, in the case of convex constraints, the condition $D^0 \neq \emptyset$ (called the *Slater CQ* [9]) is equivalent to MFCQ.

In order to proceed with the estimates for the nonconvex case, we need to assume explicitly that MFCQ holds at $\bar{x}$. In this case, it can be seen that the values μ_σ are bounded for all $\sigma > 0$ small enough. Indeed, suppose that there exists a sequence $\{\sigma_k\} \to 0+$ such that $\|\mu_{\sigma_k}\| \to \infty$. For each k, set $\bar{\mu}^k = \mu_{\sigma_k}/\|\mu_{\sigma_k}\|$. Then the sequence $\{\bar{\mu}^k\}$ has an accumulation point $\bar{\mu} \in \mathbb{R}^n \setminus \{0\}$. According to (12), for each $i \in \{1, \ldots, m\} \setminus A$, it evidently holds that $(\mu^k)_i = \sigma/(-G_i(x_{\sigma_k})) \to 0$. Therefore, for such i, $(\bar{\mu}^k)_i = (\mu^k)_i/\|\mu_{\sigma_k}\| \to 0$ as $\sigma \to 0+$, where we have also taken into account that $\|\mu_{\sigma_k}\| \to \infty$, by the assumption. Hence, $\bar{\mu}_i = 0$ for all $i \in \{1, \ldots, m\} \setminus A$. From (14), it follows that

$$\frac{1}{\|\mu_{\sigma_k}\|} f'(x_{\sigma_k}) + (G'(x_{\sigma_k}))^{\mathrm{T}} \bar{\mu}^k = 0,$$

and by passing onto the limit along an appropriate subsequence, we obtain

$$(G'_A(\bar{x}))^{\mathrm{T}} \bar{\mu}_A = 0, \quad \bar{\mu}_A \geq 0, \ \bar{\mu}_A \neq 0,$$

which contradicts (the dual form of) MFCQ.

Another useful observation is that each accumulation point of μ_σ (as $\sigma \to 0+$) belongs to $\mathcal{M}$; this follows from (12), (14).

We first consider the simpler case when the linear growth condition holds.

Theorem 2. *Assume that the linear growth condition and MFCQ hold at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then estimates (17) and (18) are valid.*

Proof. Using (13) and (14), and the above-mentioned fact that μ_σ is bounded as $\sigma \to 0+$, we obtain

$$\begin{aligned} f(x_\sigma) - f(\bar{x}) &= L(x_\sigma, \mu_\sigma) - f(\bar{x}) - \langle \mu_\sigma, G(x_\sigma) \rangle \\ &= L(x_\sigma, \mu_\sigma) - L(\bar{x}, \mu_\sigma) + \langle \mu_\sigma, G(\bar{x}) \rangle + m\sigma \\ &\leq -\left\langle \frac{\partial L}{\partial x}(x_\sigma, \mu_\sigma), x_\sigma - \bar{x} \right\rangle + m\sigma + o(\|x_\sigma - \bar{x}\|) \\ &= m\sigma + o(\|x_\sigma - \bar{x}\|). \end{aligned} \tag{20}$$

Estimate (18) follows from (8) and (20), while estimate (17) follows directly from (18) and (20). □

The estimates obtained in Theorem 2 are sharp, even under LICQ, as is demonstrated by Example 1. Moreover, Theorem 2 actually extends assertion (i) of Theorem 1 to the nonconvex case.

The case when the weaker quadratic growth condition is assumed instead of the linear growth condition, is more complex. We consider this case next.

Theorem 3. *Assume that the quadratic growth condition and MFCQ hold at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then it holds that*

$$f(x_\sigma) \leq f(\bar{x}) + O\left(\sigma\left(\sum_{i \in A} |\ln(\mu_\sigma)_i| + 1\right)\right), \tag{21}$$

$$\|x_\sigma - \bar{x}\| = O\left(\sigma^{1/2}\left(\sum_{i \in A} |\ln(\mu_\sigma)_i| + 1\right)^{1/2}\right). \tag{22}$$

In particular,

$$f(x_\sigma) \leq f(\bar{x}) + O(\sigma|\ln\sigma|), \tag{23}$$

$$\|x_\sigma - \bar{x}\| = O(\sigma^{1/2}|\ln\sigma|^{1/2}). \tag{24}$$

Proof. For each $\sigma \geq 0$, consider the set

$$D_\sigma = \{x \in \mathbb{R}^n \mid G_i(x) \leq -\sigma,\ i = 1, \ldots, m\}.$$

From MFCQ and Robinson's stability theorem [13], it follows that there exists $\tilde{x}_\sigma \in D_\sigma$ such that

$$\|\tilde{x}_\sigma - \bar{x}\| = O(\sigma). \tag{25}$$

Note that necessarily $\tilde{x}_\sigma \in D^0$ for each $\sigma > 0$. Using optimality of x_σ in problem (2), and (25), we obtain

$$\begin{aligned}
f(x_\sigma) &= \varphi_\sigma(x_\sigma) + \sigma\sum_{i=1}^{m} \ln(-G_i(x_\sigma)) \\
&\leq \varphi_\sigma(\tilde{x}_\sigma) + \sigma\sum_{i=1}^{m} \ln(-G_i(x_\sigma)) \\
&= f(\tilde{x}_\sigma) + \sigma\sum_{i=1}^{m} (\ln(-G_i(x_\sigma)) - \ln(-G_i(\tilde{x}_\sigma))) \\
&\leq f(\bar{x}) + \sigma\sum_{i=1}^{m} \ln(G_i(x_\sigma)/G_i(\tilde{x}_\sigma)) + O(\sigma) \\
&\leq f(\bar{x}) + \sigma\sum_{i \in A} \ln(-G_i(x_\sigma)/\sigma) + O(\sigma) \\
&= f(\bar{x}) - \sigma\sum_{i \in A}^{m} \ln(\mu_\sigma)_i + O(\sigma),
\end{aligned}$$

which implies estimate (21). Estimate (22) now follows from (10) and (21). The last two estimates in the assertion of the theorem are direct consequences of the first two, taking into account that for $i \in A$ and all $\sigma > 0$ sufficiently small, it holds that $\mu_\sigma = -\sigma/G_i(x_\sigma) \geq \sigma$, and hence, $|\ln(\mu_\sigma)_i| \leq |\ln\sigma|$. □

Let us discuss the estimates obtained in Theorem 3. Clearly, estimates (21) and (22) are shaper than estimates (23) and (24), respectively. But the latter can be regarded as pure convergence rate estimates and are easier to use. Also, note that the right-hand side of (24) is $o(\sigma^\nu)$ for any $\nu \in (0, 1/2)$. Nevertheless, estimate (24) is somewhat weaker than (19), of course.

Estimate (22) cannot be regarded as a computable error bound of the form (5), because it contains the set A which depends on unknown $\bar{x}$. On the other hand, (22) implies (5) with

$$r(x_\sigma, \sigma) = \sigma^{1/2} \left(\sum_{i=1}^{m} |\ln(\mu_\sigma)_i| + 1 \right)^{1/2}, \tag{26}$$

which is computable. Of course, such bound is in general weaker than (22), because $(\mu_\sigma)_i \to 0$ as $\sigma \to 0+$ for each $i \in \{1, \ldots, m\} \setminus A$; the latter follows from the above-mentioned fact that each accumulation point of μ_σ as $\sigma \to 0+$ belongs to $\mathcal{M}$. On the other hand, error bound (5) with $r(\cdot, \cdot)$ defined in (26) may be sharper than (24) (e.g., when $A = \{1, \ldots, m\}$).

In the rest of this section, we are concerned with the possibilities to improve the estimates in Theorem 3 under some additional assumptions. Recall that this can be done for convex optimization problems: in this case, "ideal" estimates (17) and (19) hold, according to Proposition 1 and the assertion (ii) of Theorem 1.

Assume that f and G are twice differentiable, and their second derivatives are continuous at $\bar{x}$. Then, by direct computation, we obtain that for each $\sigma > 0$, it holds that

$$\frac{\partial^2 L}{\partial x^2}(x_\sigma, \mu_\sigma)[\xi, \xi] + \sigma \sum_{i=1}^{m} \frac{\langle G_i'(x_\sigma), \xi \rangle^2}{(G_i(x_\sigma))^2} = \varphi_\sigma''(x_\sigma)[\xi, \xi] \geq 0 \quad \forall\, \xi \in \mathbb{R}^n, \tag{27}$$

where the second-order necessary optimality conditions for problem (2) at x_σ are taken into account.

We start with an auxiliary estimate, which involves $\bar{x}$.

Proposition 2. *Assume that MFCQ holds at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then it holds that*

$$f(x_\sigma) \leq f(\bar{x}) + \frac{\sigma}{2} \sum_{i \in A} \frac{\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle^2}{(G_i(x_\sigma))^2} + m\sigma + o(\|x_\sigma - \bar{x}\|^2). \tag{28}$$

Proof. Using (13), (14), and (27), and the above-mentioned fact that μ_σ is bounded as $\sigma \to 0+$, we obtain

$$
\begin{aligned}
f(x_\sigma) - f(\bar{x}) &= L(x_\sigma, \mu_\sigma) - f(\bar{x}) - \langle \mu_\sigma, G(x_\sigma) \rangle \\
&= L(x_\sigma, \mu_\sigma) - L(\bar{x}, \mu_\sigma) + \langle \mu_\sigma, G(\bar{x}) \rangle + m\sigma \\
&\leq -\left\langle \frac{\partial L}{\partial x}(x_\sigma, \mu_\sigma), x_\sigma - \bar{x} \right\rangle - \frac{1}{2}\frac{\partial^2 L}{\partial x^2}(x_\sigma, \mu_\sigma)[x_\sigma - \bar{x}, x_\sigma - \bar{x}] \\
&\quad + m\sigma + o(\|x_\sigma - \bar{x}\|^2) \\
&= -\frac{1}{2}\varphi''_\sigma(x_\sigma)[x_\sigma - \bar{x}, x_\sigma - \bar{x}] + \frac{\sigma}{2}\sum_{i=1}^{m} \frac{\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle^2}{(G_i(x_\sigma))^2} \\
&\quad + m\sigma + o(\|x_\sigma - \bar{x}\|^2) \\
&\leq \frac{\sigma}{2}\sum_{i \in A} \frac{\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle^2}{(G_i(x_\sigma))^2} + m\sigma + o(\|x_\sigma - \bar{x}\|^2). \quad \square
\end{aligned}
$$

Thus, the crucial question is the behavior of $\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle / G_i(x_\sigma)$ for $i \in A$. If all these quantities are bounded as $\sigma \to 0+$, then (28) implies "ideal" estimate (17), and hence (19) under the quadratic growth condition. Clearly, this question is concerned with geometry of the feasible set and of the trajectory $\sigma \to x_\sigma$. In particular, it can be shown that if there exist $i \in A$ and a sequence $\{\sigma_k\} \to 0+$ such that $\langle G_i'(x_{\sigma_k}), x_{\sigma_k} - \bar{x} \rangle / G_i(x_{\sigma_k}) \to \infty$ as $k \to \infty$, then

$$
G_i(x_{\sigma_k}) = o(\|x_{\sigma_k} - \bar{x}\|^2), \quad \langle G_i'(x_{\sigma_k}), x_{\sigma_k} - \bar{x} \rangle = O(\|x_{\sigma_k} - \bar{x}\|^2).
$$

It turns out that the ideal estimates hold when the constraints are convex.

Theorem 4. *Assume that the quadratic growth condition and MFCQ hold at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (3), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Let, in addition, G_i, $i \in A$, be convex. Then the estimates (17) and (19) are valid.*

Proof. According to the first-order convexity criterion for differentiable functions, for each $\sigma > 0$ it holds that

$$
-G_i(x_\sigma) = G_i(\bar{x}) - G_i(x_\sigma) \geq -\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle \quad \forall\, i \in A.
$$

Thus for $i \in A$, quantities $\langle G_i'(x_\sigma), x_\sigma - \bar{x} \rangle / G_i(x_\sigma)$ are bounded (above by one), and the needed estimates follow from Proposition 2. $\square$

Theorem 4 extends assertion (*ii*) of Theorem 1 to the case where the objective function need not be convex.

3 Error Estimates for the Inverse-Barrier Method

Throughout this section, φ_σ is defined with the inverse barrier (4). The development is similar to that of Section 2, except that we obtain (computable)

estimates in terms of σ and x_σ, rather than convergence rates depending on σ only.

For each $\sigma > 0$, denote

$$\mu_\sigma = \sigma(1/(G_1(x_\sigma))^2, \ldots, 1/(G_m(x_\sigma))^2) > 0. \tag{29}$$

By direct computation,

$$\langle \mu_\sigma, G(x_\sigma) \rangle = \sigma \sum_{i=1}^{m} 1/G_i(x_\sigma), \tag{30}$$

and by the first-order necessary optimality conditions for problem (2) at x_σ, we also have condition (14) (but with μ_σ defined in (29)).

If (1) is a convex minimization problem, we have the following.

Proposition 3. *Let f and G_i, $i = 1, \ldots, m$, be convex. For $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (4). Then it holds that*

$$f(x_\sigma) \leq \inf_{x \in D} f(x) - \sigma \sum_{i=1}^{m} 1/G_i(x_\sigma).$$

Proof. The assertion follows from observing that (x_σ, μ_σ) is a feasible point for the Wolfe dual of (1), and using the weak duality relation. □

In our setting, it therefore holds that

$$f(x_\sigma) \leq f(\bar{x}) + O\left(\sigma \sum_{i=1}^{m} (-1/G_i(x_\sigma))\right). \tag{31}$$

Recall that in the convex case, MFCQ is equivalent to our standing assumption that $D^0 \neq \emptyset$. By the same argument as in Section 2 (but using (29) instead of (12)), it can be shown that the values μ_σ are bounded for all $\sigma > 0$ small enough, and that each accumulation point of μ_σ (as $\sigma \to 0+$) belongs to $\mathcal{M}$. Then (30) implies that

$$\sigma \sum_{i=1}^{m} (-1/G_i(x_\sigma)) \to 0 \text{ as } \sigma \to 0+,$$

which shows that the estimate (31) is meaningful.

From Proposition 3, we immediately obtain the following result.

Theorem 5. *Let f and G_i, $i = 1, \ldots, m$, be convex. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (4), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then the following assertions are valid:*

(i) If the linear growth condition is satisfied at $\bar{x}$, then

$$\|x_\sigma - \bar{x}\| = O\left(\sigma \sum_{i=1}^{m} (-1/G_i(x_\sigma))\right). \tag{32}$$

(ii) If the quadratic growth condition is satisfied at $\bar{x}$, then

$$\|x_\sigma - \bar{x}\| = O\left(\sigma^{1/2} \sum_{i=1}^{m} (-1/G_i(x_\sigma))^{1/2}\right). \tag{33}$$

Examples 1 and 2 can be used in order to demonstrate that the estimates obtained in Theorem 5 are sharp; we omit the details.

We proceed with the estimates for the nonconvex case, explicitly assuming that MFCQ holds at $\bar{x}$.

Theorem 6. *Assume that the linear growth condition and MFCQ hold at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (4), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then estimates (31) and (32) are valid.*

Proof. The proof is similar to that of Theorem 2, but using (30) instead of (13). □

We note that a counterpart of Theorem 3 does not hold for the inverse barrier.

Assume now that f and G are twice differentiable, and their second derivatives are continuous at $\bar{x}$. By direct computation, we obtain that for each $\sigma > 0$, it holds that

$$\frac{\partial^2 L}{\partial x^2}(x_\sigma, \mu_\sigma)[\xi, \xi] - 2\sigma \sum_{i=1}^{m} \frac{\langle G_i'(x_\sigma), \xi\rangle^2}{(G_i(x_\sigma))^3} = \varphi_\sigma''(x_\sigma)[\xi, \xi] \geq 0 \quad \forall \xi \in \mathbb{R}^n, \tag{34}$$

where the second-order necessary optimality conditions for problem (2) at x_σ are taken into account.

Following the lines of the proofs of Proposition 2 and Theorem 4, we obtain the next two results.

Proposition 4. *Assume that MFCQ holds at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (4), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Then it holds that*

$$f(x_\sigma) \leq f(\bar{x}) + \sigma \sum_{i \in A} \frac{1}{-G_i(x_\sigma)} \left(\frac{\langle G_i'(x_\sigma), x_\sigma - \bar{x}\rangle^2}{(G_i(x_\sigma))^2} + 1\right) + o(\|x_\sigma - \bar{x}\|^2).$$

Theorem 7. *Assume that the quadratic growth condition and MFCQ hold at $\bar{x}$. For each $\sigma > 0$, let x_σ be a solution of problem (2) with the barrier function defined in (4), and let $x_\sigma \to \bar{x}$ as $\sigma \to 0+$. Let, in addition, G_i, $i \in A$, be convex. Then the estimates (31) and (33) are valid.*

4 Concluding Remarks

We presented computable error bound estimates and convergence rate results for some interior penalty methods. Our assumptions are essentially the Mangasarian-Fromovitz constraint qualification and the linear or quadratic growth condition (in this setting, the latter are equivalent to the first-order or second-order sufficient optimality conditions, respectively).

Some of the estimates are shown to be sharp. But at this time, it is an open question whether the estimates for the log-barrier method given under the Mangasarian-Fromovitz constraint qualification and the quadratic growth condition (Theorem 3) are sharp.

References

1. L. Bittner. Eine Verallgemeinerung des Verfahrens des logarithmischen Potentials von Frisch für nichtlineare Optimierungprobleme. In: A. Pekora (ed.), Colloquium on Applications of Mathematics to Economics, Budapest, 1963, Akademiai Kiado. Publishing House of the Hungarian Acad. of Sciences, 1965.
2. J.F. Bonnans and A. Shapiro. Perturbation Analysis of Optimization Problems. Springer-Verlag, New York, 2000.
3. A.V. Fiacco and G.P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Techniques. John Wiley & Sons, New York, 1968.
4. A. Fischer. Local behaviour of an iterative framework for generalized equations with nonisolated solutions. Math. Program. 94: 91–124, 2002.
5. C. Grossman, D. Klatte, and B. Kummer. Convergence of primal-dual solutions for the nonconvex log-barrier method without LICQ. Kybernetika 40: 571–584, 2004.
6. W.W. Hager and M.S. Gowda. Stability in the presence of degeneracy and error estimation. Math. Program. 85: 181–192, 1999.
7. A.F. Izmailov and M.V. Solodov. Computable primal error bounds based on the augmented Lagrangian and Lagrangian relaxation algorithms. Preprint A 2004/303, IMPA, Rio de Janeiro, 2004.
8. A.F. Izmailov and M.V. Solodov. Karush-Kuhn-Tucker systems: regularity conditions, error bounds and a class of Newton-type methods. Math. Program. 95: 631–650, 2003.
9. O.L. Mangasarian. Nonlinear Programming McGraw–Hill, New York, 1969
10. R. Mifflin. On the convergence of the logarithmic barrier function method. In: F.A. Lootsma (ed.), Numerical Methods for Nonlinear Optimization. Academic Press, London, 1972.
11. R. Mifflin. Convergence bounds for nonlinear programming algorithms. Math. Program. 8: 251–271, 1975.
12. J.S. Pang. Error bounds in mathematical programming. Math. Program. 79: 299–332, 1997.
13. S.M. Robinson. Stability theorems for systems of inequalities, Part II: differentiable nonlinear systems. SIAM J. Numer. Anal. 13: 497–513, 1976.
14. S.J. Wright and D. Orban. Properties of the log-barrier function on degenerate nonlinear programs. Math. Oper. Res. 27: 585–613, 2002.

Part II

Optimal Control and Calculus of Variations

L^1–Optimal Boundary Control of a String to Rest in Finite Time

Martin Gugat

Lehrstuhl 2 für Angewandte Mathematik, Martensstr. 3, 91058 Erlangen, Germany. gugat@am.uni-erlangen.de

Summary. In this paper, the problem to control a finite string to the zero state in finite time from a given initial state by controlling the state at the two boundary points is considered. The corresponding optimal control problem where the objective function is the L^1-norm of the controls is solved in the sense that the controls that are successful and minimize at the same time the objective function are determined as functions of the initial state.

1 Introduction

We consider a string of finite length that is governed by the wave equation. The string is controlled through the boundary values at both ends of the string (two–point Dirichlet control). The boundary control of the wave equation has been studied by many authors and results about exact controllability are well–known. The method of moments is an important tool to analyze this system (see e.g. [1, 7, 8, 10, 12] and the references therein). Also the controllability of the discretized problems and the relation between the optimal controls for the continuous and the discrete case have been the subject of recent investigations, see [14]. A related problem of one–point time optimal control has been solved in [11], where the control functions are assumed to have a second derivative whose norm is constrained. In [13], exact controllability is studied for a string with elastic fixing at one end.

In this paper, our main interest is to study the structure of the optimal controls and to give an explicit representation of the optimal controls in terms of the given initial data. This yields valuable test examples for numerical algorithms.

From a given initial state where the position and the integral of the velocity are given by a Lebesgue–integrable function the system is controlled to the zero state in a given finite time.

To guarantee that this control problem is solvable for all initial states, the control time has to be greater than or equal to the time that a wave needs

to travel from one end of the string to the other (the characteristic time). In Theorem 1 we give an exact controllability result where the initial states that can be steered to zero with boundary controls from the spaces L^p ($p \in [1, \infty]$) are characterized: These are the initial states where the initial position and the integral of the initial velocity are functions in the spaces L^p on the space interval.

The requirements that the target state is reached in the given terminal time do not determine a unique solution. So we can choose from the set of successful controls a point that minimizes our objective function which is the L^1–norm of the controls. In general, this optimization problem does not have a unique solution. In Theorem 2 the solutions are given explicitly in terms of the initial data.

In [2], [4] and [6], we have studied the related problem to steer the system from the zero state to a given terminal state in such a way that the L^p-norm ($p \in [2, \infty]$) of the control functions is minimized. In these papers, the method of moments and Fourier–series have been used in the proofs. In the present paper we use the method of characteristics for our proofs. Note that in contrast to the L^1–case, for $p \in (1, \infty)$ the corresponding optimal controls are uniquely determined.

This paper has the following structure: We define the optimal control problem and some important auxiliary variables, for example the characteristic time and the defect. Then the problem is transformed and reformulated in terms of the Riemann invariants. For this purpose, we use the d'Alembert solution of the wave equation. After the introduction of auxiliary functions as variables in the optimization problem, the exact controllability result Theorem 1 can be proved. Then the objective function is also written in terms of the auxiliary functions, which allows to reformulate the optimization problem such that it decouples to time–parametric finite dimensional problems that can be solved explicitly. (These auxiliary problems also do not have a unique solution.) This allows to solve the optimal control problem. In Theorem 2, the solutions of the L^1–optimal control problem are given in terms of the initial state. Finally we present some examples.

2 The Problem

Let $L^1(0,T)$ denote the space of Lebesgue–integrable functions on the interval $(0,T)$, and let

$$\|(u_1, u_2)\|_{1,(0,T)} = \int_0^T |u_1(t)| + |u_2(t)|\, dt.$$

Let the length $L > 0$, the time $T > 0$ and the wave velocity $c > 0$ be given. Let $y_0 \in L^1(0,L)$ and y_1 be given such that the function $x \mapsto \int_0^x y_1(s)\, ds$ is in $L^1(0,L)$.

We consider the problem

$$\mathcal{P}: \quad \text{minimize } \|(u_1, u_2)\|_{1,(0,T)} \text{ subject to } u_1, u_2 \in L^1(0,T) \text{ and} \tag{1}$$

$$y(x,0) = y_0(x),\ y_t(x,0) = y_1(x),\ x \in (0,L) \tag{2}$$

$$y(0,t) = u_1(t),\ y(L,t) = u_2(t),\ t \in (0,T) \tag{3}$$

$$y_{tt}(x,t) = c^2 y_{xx}(x,t),\ (x,t) \in (0,L) \times (0,T) \tag{4}$$

$$y(x,T) = 0,\ y_t(x,T) = 0,\ x \in (0,L). \tag{5}$$

3 Definition of the Characteristic Time

Define the characteristic time $t_0 = L/c$ that a characteristic curve needs to travel from one end of the string to the other. In the sequel we assume that

$$T \geq t_0.$$

For the solution of the problem, we need to know how often the characteristic time t_0 fits into the time interval $[0,T]$. Define the natural number

$$k = \max\{j \in I\!N : j\ t_0 \leq T\} \tag{6}$$

and the defect

$$\Delta = T - k\ t_0 \geq 0. \tag{7}$$

The definition of Δ implies the equation $T = k\ t_0 + \Delta$.

4 Transformation of the Problem

In order to come closer to a solution of Problem $\mathcal{P}$, we transform it to a form that we can solve. For this purpose, we write the solution of the wave equation in the form

$$y(x,t) = [\alpha(x+ct) + \beta(x-ct)]/2 \tag{8}$$

which means that we describe our solution in terms of the Riemann invariants or in other words, as the sum of travelling waves. For an introduction to linear hyperbolic systems see [9].

The *end conditions* (5) yield the equations

$$\alpha(x+cT) + \beta(x-cT) = 0,\ \alpha'(x+cT) - \beta'(x-cT) = 0,\ x \in (0,L) \tag{9}$$

where the derivatives are in the sense of distributions. This is equivalent to

$$\alpha(x) = -\beta(x-2cT),\ \alpha'(x) = \beta'(x-2cT),\ x \in (cT, cT+L). \tag{10}$$

Differentiation of the first equation in (10) yields

$$\alpha'(x) = -\beta'(x - 2cT),\ x \in (cT, cT + L)$$

hence we have $\alpha'(x) = -\alpha'(x)$ and thus

$$\alpha'(x) = 0,\ x \in (cT, cT + L);\ \ \beta'(x) = 0,\ x \in (-cT, -cT + L). \tag{11}$$

So the first equation in (10) implies that there exists a real constant r such that

$$\alpha(x) = r,\ x \in (cT, cT + L);\ \ \beta(x) = -r,\ x \in (-cT, -cT + L). \tag{12}$$

We have shown that if (8) satisfies the end conditions (5), then (12) holds. The reverse statement is obviously true.

The *initial conditions* (2) yield the equations

$$y_0(x) = (1/2)\,[\alpha(x) + \beta(x)],\ y_1(x) = (c/2)\,[\alpha'(x) - \beta'(x)],\ x \in (0, L). \tag{13}$$

Hence we have

$$y_0(x) + (1/c)\int_0^x y_1(s)\,ds = \alpha(x) - k_1,\ x \in (0, L) \tag{14}$$

$$y_0(x) - (1/c)\int_0^x y_1(s)\,ds = \beta(x) + k_1,\ \ x \in (0, L) \tag{15}$$

for a real constant k_1 that we can choose as zero, which implies

$$\alpha(x) = y_0(x) + (1/c)\int_0^x y_1(s)\,ds,\ x \in (0, L), \tag{16}$$

$$\beta(x) = y_0(x) - (1/c)\int_0^x y_1(s)\,ds,\ x \in (0, L). \tag{17}$$

We have shown that if (8) satisfies the initial conditions (2), then (16), (17) hold. The converse also holds: If α, β satisfy (16), (17), the initial conditions (2) are valid for y given by (8).

5 Exact Controllability

The considerations in the last section imply the following exact controllability result:

Theorem 1. *Let $T \geq L/c$ and $p \in [1, \infty]$ be given. The initial boundary–value problem (2)–(4) has a travelling waves solution in the sense (8) that satisfies the end conditions (5) with u_1, $u_2 \in L^p(0, T)$, if and only if the initial states y_0, y_1 satisfy the following conditions: $y_0 \in L^p(0, L)$ and $Y_1 \in L^p(0, L)$, where $Y_1(x) = \int_0^x y_1(s)\,ds$, that is $y_1 \in W^{-1,p}(0, L)$.*

This implies that Problem $\mathcal{P}$ is solvable if and only if y_0 and Y_1 are in $L^1(0, L)$.

Proof of one direction. Assume that y_0 and $Y_1 \in L^p(0, L)$. Define

$$u_1(t) = y(0,t) = [\alpha(ct) + \beta(-ct)]/2$$

$$u_2(t) = y(L,t) = [\alpha(L+ct) + \beta(L-ct)]/2$$

where the functions $\alpha \in L^p(0, L+ct)$, $\beta \in L^p(-cT, L)$ are chosen such that (12) and (16), (17) hold, for example with $r = 0$ and $\alpha(x) = 0$ for $x \in (L, cT)$ and $\beta(x) = 0$ for $x \in (L - cT, 0)$. Then the solution y given by (8) satisfies the initial conditions (2) and the end conditions (5). Moreover, u_1 and u_2 are in $L^p(0,T)$. The proof of the converse is given in the next section. □

Remark 1: For the case $p \in [2, \infty]$, Theorem 1 is already proved in [6] using Fourier series. Note however, that in [6] the initial state is the zero state which is controlled in the time T to the target state (y_0, y_1).

6 Definition of Auxiliary Functions and Completion of the Proof of Theorem 1

For $j \in \{0, 1, ..., k\}$ and $t \in (0, t_0)$ define the functions

$$\alpha_j(t) = \alpha(ct + jL),\ \beta_j(t) = \beta(-ct - (j-1)L) \tag{18}$$

and for $t \in (0, \Delta)$ define

$$\alpha_{k+1}(t) = \alpha(ct + (k+1)L),\ \beta_{k+1}(t) = \beta(-ct - kL). \tag{19}$$

The functions α_j, β_j are useful as decision variables in the transformed optimization problem. We will state the constraints in terms of the functions α_j, β_j: Since

$$[cT, cT + L] = [kL + c\Delta, (k+1)L + c\Delta]$$

$$= [kL + c\Delta, (k+1)L] \cup [(k+1)L, (k+1)L + c\Delta]$$

and

$$[-cT, -cT + L] = [-kL - c\Delta, -(k-1)L - c\Delta]$$

$$= [-kL - c\Delta, -kL] \cup [-kL, -(k-1)L - c\Delta]$$

the constraints (12) are equivalent to the conditions

$$\alpha_k(t) = r,\ t \in (\Delta, t_0),\ \alpha_{k+1}(t) = r,\ t \in (0, \Delta), \tag{20}$$

$$\beta_k(t) = -r,\ t \in (\Delta, t_0),\ \beta_{k+1}(t) = -r,\ t \in (0, \Delta). \tag{21}$$

This means that the functions α_{k+1}, β_{k+1} are constant on $(0, \Delta)$ and the functions α_k, β_k are constant on (Δ, t_0) with the same absolute values but with opposite signs.

Conditions (16) and (17) are respectively equivalent to

$$\alpha_0(t) = y_0(ct) + (1/c)\int_0^{ct} y_1(s)\,ds,\ t \in (0,t_0), \tag{22}$$

$$\beta_0(t) = y_0(L-ct) - (1/c)\int_0^{L-ct} y_1(s)\,ds,\ t \in (0,t_0), \tag{23}$$

so the values of the functions α_0, β_0 are prescribed by the initial conditions.

We can represent the control functions u_1, u_2 in terms of α_j, β_j in the following way. Define the intervals

$$I_j^1 = [jt_0, jt_0 + \Delta],\ j \in \{0,1,2,...,k\}, \tag{24}$$

$$I_j^2 = [jt_0 + \Delta, (j+1)t_0],\ j \in \{0,1,2,...,k-1\}. \tag{25}$$

Then for $t \in I_j^1$ or $t \in I_j^2$ we have

$$u_1(t) = [\,\alpha_j(t - jt_0) + \beta_{j+1}(t - jt_0)\,]/2, \tag{26}$$

$$u_2(t) = [\,\alpha_{j+1}(t - jt_0) + \beta_j(t - jt_0)\,]/2. \tag{27}$$

Now we complete the proof of Theorem 1. Assume that controls u_1, $u_2 \in L^p(0,T)$ are given such that the travelling waves solution (8) satisfies the initial conditions (2) and the end conditions (5). The end conditions (20), (21) imply that the functions α_{k+1}, β_{k+1} are in $L^p(0,\Delta)$. Then (26) and the fact that u_1 is in $L^p(0,T)$ imply that α_k is also in $L^p(0,\Delta)$. Equation (27) and the fact that $u_2 \in L^p(0,T)$ imply that β_k is also in $L^p(0,\Delta)$. Analogous arguments show that α_{k-1}, β_{k-1} are in $L^p(0,\Delta)$ and repeating the argument shows that α_0, β_0 is in $L^p(0,\Delta)$.

The end conditions (20), (21) imply that the functions α_k, β_k are in $L^p(\Delta,t_0)$. Then (26) and $u_1 \in L^p(0,T)$ imply that α_{k-1} is also in $L^p(\Delta,t_0)$. Equation (27) and the fact that $u_2 \in L^p(0,T)$ imply that β_{k-1} is also in $L^p(\Delta,t_0)$. Repeating the argument implies that α_0, β_0 are in $L^p(\Delta,t_0)$.

Thus we have shown that α_0, β_0 are in $L^p(0,t_0)$. Equations (22), (23) imply that y_0 is in $L^p(0,L)$ and that Y_1 is in $L^p(0,L)$.

7 Reformulation of the Optimization Problem in terms of α_j, β_j

We start by transforming our objective function

$$J(u_1,u_2) = \int_0^T |u_1(t)| + |u_2(t)|\,dt. \tag{28}$$

We have

$$J(u_1,u_2) = \sum_{j=0}^{k}\int_{jt_0}^{jt_0+\Delta} |u_1(t)| + |u_2(t)|\,dt + \sum_{j=0}^{k-1}\int_{jt_0+\Delta}^{(j+1)t_0} |u_1(t)| + |u_2(t)|\,dt$$

$$= \sum_{j=0}^{k} \int_0^{\Delta} |u_1(t+jt_0)| + |u_2(t+jt_0)|\, dt + \sum_{j=0}^{k-1} \int_{\Delta}^{t_0} |u_1(t+jt_0)| + |u_2(t+jt_0)|\, dt$$

$$= \int_0^{\Delta} \sum_{j=0}^{k} |u_1(t+jt_0)| + |u_2(t+jt_0)|\, dt + \int_{\Delta}^{t_0} \sum_{j=0}^{k-1} |u_1(t+jt_0)| + |u_2(t+jt_0)|\, dt$$

$$= \int_0^{\Delta} \sum_{j=0}^{k} \left[\frac{1}{2}|\alpha_j(t) + \beta_{j+1}(t)| + \frac{1}{2}|\alpha_{j+1}(t) + \beta_j(t)| \right] dt$$

$$+ \int_{\Delta}^{t_0} \sum_{j=0}^{k-1} \left[\frac{1}{2}|\alpha_j(t) + \beta_{j+1}(t)| + \frac{1}{2}|\alpha_{j+1}(t) + \beta_j(t)| \right] dt$$

$$=: F(\alpha_j|_{(0,\Delta)},\, \beta_j|_{(0,\Delta)}, j \in \{1,...,k\};\ \alpha_j|_{(\Delta,t_0)},\, \beta_j|_{(\Delta,t_0)}, j \in \{1,...,k-1\}). \tag{29}$$

Now we write down our optimization problem in terms of the unknown functions α_j, β_j. If we have determined a solution pair α_j, β_j, we obtain the corresponding controls u_1, u_2 from (26), (27). In this sense Problem $\mathcal{P}$ is equivalent to the problem:

$$\text{minimize the objective function } F \text{ given in (29)} \tag{30}$$

over the functions

$$\alpha_j|_{(0,\Delta)},\ \beta_j|_{(0,\Delta)} \in L^1(0,\Delta),\ j \in \{1,...,k\},$$

$$\alpha_j|_{(\Delta,t_0)},\ \beta_j|_{(\Delta,t_0)} \in L^1(\Delta,t_0),\ j \in \{1,...,k-1\}$$

where α_0, β_0 are given in (22), (23) and $\alpha_k|_{(\Delta,t_0)}$, $\beta_k|_{(\Delta,t_0)}$, $\alpha_{k+1}|_{(0,\Delta)}$, $\beta_{k+1}|_{(0,\Delta)}$ are given by (20), (21).

7.1 Definition of a Time–Parametric Optimization Problem

For $t \in (0,t_0)$ and a natural number m consider the optimization problem

$$H(t,m): \quad \min \sum_{j=0}^{m-1} \frac{1}{2}|\alpha_j(t) + \beta_{j+1}(t)| + \frac{1}{2}|\alpha_{j+1}(t) + \beta_j(t)| \tag{31}$$

where the numbers $\alpha_0(t)$, $\beta_0(t)$ and $\alpha_m(t)$, $\beta_m(t)$ are given and the decision variables are $\alpha_1(t)$,...,$\alpha_{m-1}(t)$, $\beta_1(t)$,...,$\beta_{m-1}(t)$. If $m = 1$ there are no decision variables. The objective function of $H(t,m)$ is the integrand of the function F given in (29) at a single point $t \in (0,t_0)$, so the idea of $H(t,m)$ is to minimize the integrand of Problem (30) at a single point in time.

We obtain solutions of Problem (30) by solving the optimization problems $H(t,k+1)$ for $t \in (0,\Delta)$ and $H(t,k)$ for $t \in (\Delta,t_0)$ almost everywhere, that is

we minimize the integrand in the objective function J pointwise a.e.. Consider solutions $\alpha_j(t)$, $\beta_j(t)$ of these optimization problems as functions of t. If these functions are Lebesgue integrable, they are condidates for a solution of the optimization problem (30) and thus yield solutions of the optimal control problem $\mathcal{P}$. In fact, the solutions $\alpha_j(t)$, $\beta_j(t)$ are coupled by the real parameter r from (20), (21). So we reduce the original infinite–dimensional problem to the problem to find the value of the real number r for which the objective function evaluated at the corresponding solutions $\alpha_j(t)$, $\beta_j(t)$ has minimal value.

7.2 Solution of a Time–Parametric Optimization Problem

Consider problem $H(t,m)$ for a fixed time $t \in (0,t_0)$. Since t is fixed, we call the decision variables $\alpha_1,...,\alpha_{m-1}$, $\beta_1,...,\beta_{m-1}$ and omit the parameter t. We introduce new variables:

$$\gamma_j = \alpha_j + \beta_{j+1}\ ,\ \delta_j = \alpha_{j+1} + \beta_j \quad \text{for } j \text{ even},$$
$$\gamma_j = \alpha_{j+1} + \beta_j\ ,\ \delta_j = \alpha_j + \beta_{j+1} \quad \text{for } j \text{ odd}.$$

We have

$$\sum_{j=0}^{m-1}(-1)^j\gamma_j = \begin{cases} \alpha_0 - \alpha_m & \text{if } m \text{ is even}, \\ \alpha_0 + \beta_m & \text{if } m \text{ is odd}. \end{cases}$$

If $\alpha_m = r = -\beta_m$ as in (20), (21), this yields for all m the equation

$$\sum_{j=0}^{m-1}(-1)^j\gamma_j = \alpha_0 - r = c_1.$$

Similarly,

$$\sum_{j=0}^{m-1}(-1)^j\delta_j = \begin{cases} \beta_0 - \beta_m & \text{if } m \text{ is even}, \\ \beta_0 + \alpha_m & \text{if } m \text{ is odd}. \end{cases}$$

If $\alpha_m = r = -\beta_m$ as in (20), (21), this yields for all m the equation

$$\sum_{j=0}^{m-1}(-1)^j\delta_j = \beta_0 + r = c_2.$$

We also have

$$\sum_{j=0}^{m-1}|\alpha_j(t) + \beta_{j+1}(t)| + |\alpha_{j+1}(t) + \beta_j(t)| = \sum_{j=0}^{m-1}|\gamma_j| + |\delta_j|.$$

This means that we can decouple problem $H(t,m)$ into two problems

$$P_1: \quad \min \sum_{j=0}^{m-1}|\gamma_j| \ \text{ s.t. } \ \sum_{j=0}^{m-1}(-1)^j\gamma_j = c_1, \tag{32}$$

$$P_2: \quad \min \sum_{j=0}^{m-1}|\delta_j| \ \text{ s.t. } \ \sum_{j=0}^{m-1}(-1)^j\delta_j = c_2. \tag{33}$$

Lemma 1. *The optimal value of P_1 is $|c_1|$. If $c_1 = 0$, the solution of P_1 is uniquely determined. If $c_1 \neq 0$, the solution of P_1 is not uniquely determined. In fact, $(\gamma_0, ..., \gamma_{m-1})$ is a solution of P_1 if and only if*

$$\gamma_j = (-1)^j \lambda_j c_1, \; j \in \{0, ..., m-1\},$$

where for $j \in \{0, ..., m-1\}$ we have $\lambda_j \geq 0$ and $\sum_{i=0}^{m-1} \lambda_i = 1$. The corresponding assertions for P_2 also hold.

Proof. The point with the components γ_j as defined in the Lemma satisfies the equality constraint of P_1 and has the objective value $|c_1|$. Thus the objective value is less than or equal to $|c_1|$. Now take an arbitrary point that satisfies the equality constraint of P_1. Then the triangle inequality implies

$$\sum_{j=0}^{m-1} |\gamma_j| = \sum_{j=0}^{m-1} |(-1)^j \gamma_j| \geq |\sum_{j=0}^{m-1} (-1)^j \gamma_j| = |c_1|.$$

Hence the optimal value of P_1 is greater than or equal to $|c_1|$ and we have proved the assertion. Assume that $c_1 \neq 0$. Let an arbitrary solution of P_1 with the components η_0, η_1,...,η_{m-1} be given. Then we have

$$\sum_{j=0}^{m-1} |\eta_j| = |c_1|.$$

Define $\lambda_j = |\eta_j|/|c_1|$. Then $\lambda_j \geq 0$, $\sum_{i=0}^{m-1} \lambda_i = 1$, and $\eta_j = |c_1| \lambda_j \text{sign}(\eta_j)$. The equation

$$\sum_{j=0}^{m-1} (-1)^j \eta_j = \sum_{j=0}^{m-1} \lambda_j |c_1| (-1)^j \text{sign}(\eta_j) = |c_1| \sum_{j=0}^{m-1} \lambda_j (-1)^j \text{sign}(\eta_j) = c_1$$

holds. Thus

$$\sum_{j=0}^{m-1} \lambda_j (-1)^j \text{sign}(\eta_j) = \text{sign}(c_1).$$

This equation can only hold if for all $j \in \{0, ..., m-1\}$ we have

$$(-1)^j \text{sign}(\eta_j) = \text{sign}(c_1),$$

which implies $\text{sign}(\eta_j) = (-1)^j \text{sign}(c_1)$. Thus we have $\eta_j = (-1)^j \lambda_j c_1$, and the assertion follows. □

7.3 Solution of the Optimal Control Problem

Consider the functions $\alpha_j(t)$, $\beta_j(t)$ defined as the solutions of $H(t, k+1)$ for $t \in (0, \Delta)$ and of $H(t, k)$ for $t \in (\Delta, t_0)$ almost everywhere. Lemma 1 gives

values $\alpha_j(t)+\beta_{j+1}(t)$ and $\alpha_{j+1}(t)+\beta_j(t)$ for the solution of problem $H(t,m)$ explicitly. The general solution given in Lemma 1 yields solutions of the form

$$\gamma_j(t) = (-1)^j\lambda_j(t)[\alpha_0(t)-r] \text{ for } t\in(0,\Delta),\ j\in\{0,...,k\}$$
$$\gamma_j(t) = (-1)^j\mu_j(t)[\alpha_0(t)-r] \text{ for } t\in(\Delta,t_0),\ j\in\{0,...,k-1\}$$
$$\delta_j(t) = (-1)^j\nu_j(t)[\beta_0(t)+r] \text{ for } t\in(0,\Delta),\ j\in\{0,...,k\}$$
$$\delta_j(t) = (-1)^j\omega_j(t)[\beta_0(t)+r] \text{ for } t\in(\Delta,t_0),\ j\in\{0,...,k-1\}.$$

Here λ_j and ν_j are functions defined almost everywhere on $(0,\Delta)$ such that

$$\lambda_j(t)\geq 0,\ \nu_j(t)\geq 0,\ \sum_{j=0}^{k}\lambda_j(t)=1=\sum_{j=0}^{k}\nu_j(t),$$

and such that the functions $\lambda_j(\alpha_0-r)$ and $\nu_j(\beta_0+r)$ are in $L^1(0,\Delta)$ for all $j\in\{0,...,k\}$. Moreover μ_j and ω_j are functions defined almost everywhere on (Δ,t_0) such that

$$\mu_j(t)\geq 0,\ \omega_j(t)\geq 0,\ \sum_{j=0}^{k-1}\mu_j(t)=1=\sum_{j=0}^{k-1}\omega_j(t),$$

and such that the functions $\mu_j(\alpha_0-r)$ and $\omega_j(\beta_0+r)$ are in $L^1(\Delta,t_0)$ for all $j\in\{0,...,k-1\}$.

Equations (26), (27) and the definition of γ_j, δ_j imply that the control values corresponding to these functions are given as

$$u_1(t+jt_0) = \gamma_j(t)/2 \text{ if } j \text{ is even}, \tag{34}$$
$$u_1(t+jt_0) = \delta_j(t)/2 \text{ if } j \text{ is odd}, \tag{35}$$

$$u_2(t+jt_0) = \delta_j(t)/2 \text{ if } j \text{ is even}, \tag{36}$$
$$u_2(t+jt_0) = \gamma_j(t)/2 \text{ if } j \text{ is odd}. \tag{37}$$

Now both for u_1 and u_2 we have to consider four different cases, depending on whether t is in the interval $(0,\Delta)$ or the interval (Δ,t_0) and on whether j is even of j is odd. The general solutions given in Lemma 1 correspond to optimal controls of the form

$$u_1(t+jt_0) = \lambda_j(t)[\alpha_0(t)-r]/2 \text{ if } j \text{ is even and } t\in(0,\Delta), \tag{38}$$
$$u_1(t+jt_0) = \mu_j(t)[\alpha_0(t)-r]/2 \text{ if } j \text{ is even and } t\in(\Delta,t_0), \tag{39}$$
$$u_1(t+jt_0) = -\nu_j(t)[\beta_0(t)+r]/2 \text{ if } j \text{ is odd and } t\in(0,\Delta), \tag{40}$$
$$u_1(t+jt_0) = -\omega_j(t)[\beta_0(t)+r]/2 \text{ if } j \text{ is odd and } t\in(\Delta,t_0), \tag{41}$$

$$u_2(t+jt_0) = \nu_j(t)[\beta_0(t)+r]/2 \text{ if } j \text{ is even and } t\in(0,\Delta), \tag{42}$$
$$u_2(t+jt_0) = \omega_j(t)[\beta_0(t)+r]/2 \text{ if } j \text{ is even and } t\in(\Delta,t_0), \tag{43}$$
$$u_2(t+jt_0) = \lambda_j(t)[-\alpha_0(t)+r]/2 \text{ if } j \text{ is odd and } t\in(0,\Delta), \tag{44}$$
$$u_2(t+jt_0) = \mu_j(t)[-\alpha_0(t)+r]/2 \text{ if } j \text{ is odd and } t\in(\Delta,t_0). \tag{45}$$

If $T = k\,t_0$, that is if $\Delta = 0$ the intervals $(0, \Delta)$ vanish.

It only remains to determine the value of the real number r. For this purpose, the control given above is inserted in the objective function $J(u_1, u_2)$ and r is chosen such that $J(u_1, u_2)$ is minimized.

8 Main Result

In this section we state the main result of this paper, which provides an explicit solution to the optimization problem $\mathcal{P}$, that is to say, to the L^1–norm optimal two–point Dirichlet boundary control of the wave equation to the zero position.

Theorem 2. *Assume that T is greater than or equal to $t_0 = L/c$. Consider the Problem $\mathcal{P}$ defined in (1)–(5). Choose a real number r that minimizes*

$$\frac{1}{2}\int_0^{t_0} |\alpha_0(t) - r| + |\beta_0(t) + r|\, dt \tag{46}$$

where α_0 is given by (22) and β_0 is given by (23).

Then a solution of Problem $\mathcal{P}$ is given by controls u_1, u_2 defined in (38)–(45) and, conversely, every solution has this form.

The minimal value of Problem $\mathcal{P}$ is given by the integral (46) with an optimal choice of r. Problem $\mathcal{P}$ admits a unique solution if and only if the minimal value of Problem $\mathcal{P}$ is zero.

Proof. We have presented controls u_1, $u_2 \in L^1(0,T)$ such that the generated state satisfies the end conditions and the corresponding value of the objective function is

$$J(u_1, u_2) = \min_r \frac{1}{2}\int_0^{t_0} |\alpha_0(t) - r| + |\beta_0(t) + r|\, dt.$$

Let v_1, $v_2 \in L^1(0,T)$ be control functions for which the generated state satisfies the end conditions. Then there exists a real number $r = r_0$ such that (12) holds. Suppose that the corresponding functions γ_j, δ_j (as in (34), (37)) do not solve the problem $H(t, k+1)$ almost everywhere on $(0, \Delta)$ (with $\alpha_{k+1} = r_0$, $\beta_{k+1} = -r_0$) or do not solve the problem $H(t,k)$ almost everywhere on (Δ, t_0) (with $\alpha_k = r_0$, $\beta_k = -r_0$). For $t \in (0, \Delta)$, let $h_1(t)$ denote the optimal value of $H(t, k+1)$. Lemma 1 implies that $h_1(t) = [|\alpha_0(t) - r_0| + |\beta_0(t) + r_0|]/2$. For $t \in (\Delta, t_0)$ let $h_2(t)$ denote the optimal value of $H(t,k)$. Lemma 1 implies that $h_2(t) = [|\alpha_0(t) - r_0| + |\beta_0(t) + r_0|]/2$. Then we have

$$\begin{aligned} J(v_1, v_2) &> \int_0^{\Delta} h_1(t)\, dt + \int_{\Delta}^{t_0} h_2(t)\, dt \\ &= \frac{1}{2}\int_0^{t_0} |\alpha_0(t) - r_0| + |\beta_0(t) + r_0|\, dt \\ &\geq J(u_1, u_2). \end{aligned}$$

Hence v_1, v_2 cannot be a solution of $\mathcal{P}$. This yields the assertion that the optimal controls are of the form as stated in the theorem, that is they solve the problem $H(t, k+1)$ almost everywhere on $(0, \Delta)$ (with $\alpha_{k+1} = r$, $\beta_{k+1} = -r$) and solve the problem $H(t, k)$ almost everywhere on (Δ, t_0) (with $\alpha_k = r$, $\beta_k = -r$), where r is chosen as to minimize (46). □

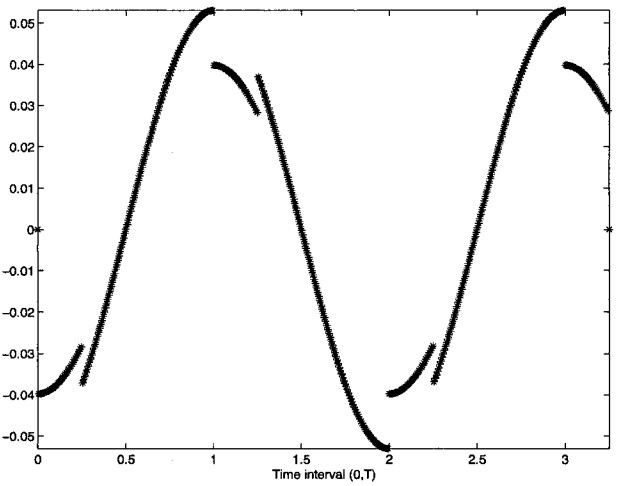


Fig. 1. The optimal control $u_1 = u_2$ in Example 2

9 Examples

In general the value of r for which the integral (46) attains its minimal value is *not* uniquely determined.

Example 1. Assume that $y_0 = c_0$ is constant and $y_1 = 0$, that is the string is initially at rest. Then (22) implies that $\alpha_0(t) = c_0$ and (23) implies that $\beta_0(t) = c_0$, hence we have $\alpha_0(t) = \beta_0(t) = c_0$, and the number r from Theorem 2 minimizes

$$\int_0^{t_0} |c_0 - r| + |c_0 + r| \, dt = t_0(|c_0 - r| + |c_0 + r|).$$

The value $r = 0$, minimizes the integral, since with $r = 0$ the integrand equals $2|c_0|$ and the triangle inequality implies that for all real numbers s we have

$$2|c_0| = |c_0 - s + c_0 + s| + \leq |c_0 - s| + |c_0 + s|.$$

So the optimal value of Problem $\mathcal{P}$ is $|c_0|t_0$. In this case, (38)–(45) imply that for all $j \in \{0, ..., k\}$, $t \in (0, \Delta)$ optimal controls are given by

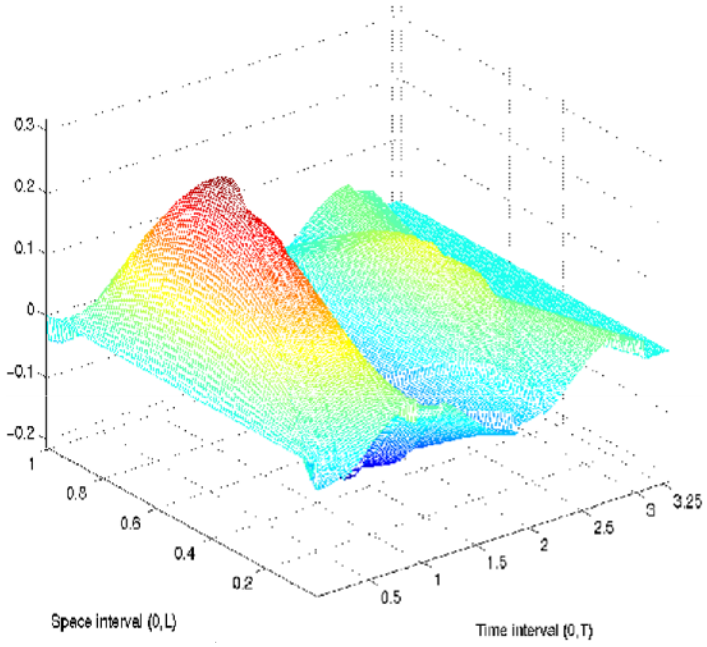


Fig. 2. The optimal state y in Example 2

$$u_1(t + jt_0) = u_2(t + jt_0) = (-1)^j \frac{c_0}{2(k+1)}$$

and for all $j \in \{0, ..., k-1\}$, $t \in (\Delta, t_0)$ optimal controls are given by

$$u_1(t + jt_0) = u_2(t + jt_0) = (-1)^j \frac{c_0}{2k}.$$

If $c_0 > 0$, the optimal value is $t_0 c_0$. With $r \in [-c_0,\ c_0]$, the integral (46) has the value

$$\frac{1}{2} \int_0^{t_0} c_0 - r + c_0 + r \, dt = t_0 c_0,$$

thus also for all $r \in [-c_0, c_0]$ the controls given by (38)–(45) are optimal. Only in the trivial case $c_0 = 0$ where the initial state is already zero, the choice $r = 0$ represents the unique solution.

Example 2. Assume that $y_0(x) = 0$ and $y_1(x) = \sin(x\pi/L)$. Then (22) and (23) imply respectively

$$\alpha_0(t) = \frac{2L}{c\pi} \left[\sin(t\frac{c\pi}{2L})\right]^2, \quad \beta_0(t) = -\frac{2L}{c\pi} \left[\cos(t\frac{c\pi}{2L})\right]^2.$$

Since $\alpha_0(t) = [(2L)/(c\pi)] + \beta_0(t)$, for all real numbers $s \neq 0$ we have

$$|\alpha_0(t) - \frac{L}{c\pi}| + |\beta_0(t) + \frac{L}{c\pi}| = 2|\frac{L}{c\pi} + \beta_0(t)| = |\frac{L}{c\pi} + \frac{L}{c\pi} + \beta_0(t) + \beta_0(t)|$$

$$= |\frac{2L}{c\pi} - s + \beta_0(t) + s + \beta_0(t)| \leq |\frac{2L}{c\pi} - s + \beta_0(t)| + |s + \beta_0(t)| = |\alpha_0(t) - s| + |\beta_0(t) + s|.$$

Hence with the value $r = L/(c\pi)$ the integral from Theorem 2 attains its minimal value, namely

$$\frac{1}{2}\int_0^{t_0} 2|\frac{L}{c\pi} + \beta_0(t)|\, dt$$

and optimal controls are given by (38)–(45) with $r = L/(c\pi)$. Note that since $\alpha_0(t) - r = \beta_0(t) + r$ we have $u_1 = u_2$.

Now let $L = 1$, $c = 1$ and $T = 3.25$, hence $k = 3$. Figure 1 shows the corresponding optimal control $u_1 = u_2$ with $r = 1/\pi$ and Figure 2 shows the state y generated by u_1 and u_2.

References

1. S. A. Avdonin and S. S. Ivanov. Families of Exponentials. Cambridge University Press, 1995.
2. M. Gugat. Analytic solutions of L^∞–optimal control problems for the wave equation. J. Optim. Theory Appl., 114:397–421, 2002.
3. M. Gugat. Boundary controllability between sub– and supercritical flow. SIAM J. Control Optim., 42:1056–1070, 2003.
4. M. Gugat and G. Leugering. Solutions of L^p–norm–minimal control problems for the wave equation. Comput. and Applied Math., 21:227–244, 2002.
5. M. Gugat and G. Leugering. Global boundary controllability of the de St. Venant equations between steady states. Ann. Inst. Henri Poincaré, Nonlinear Analysis, 20:1–11, 2003.
6. M. Gugat, G. Leugering, and G. Sklyar. L^p-optimal boundary control for the wave equation. SIAM J. Control Optim., pp. xx–xx, 2005.
7. W. Krabs. Optimal control of processes governed by partial differential equations. Part II: Vibrations. Z. Oper. Res., 26:63–86, 1982.
8. W. Krabs. On moment theory and controllability of one–dimensional vibrating systems and heating processes. Lect. Notes in Control and Information Science 173, Springer–Verlag, Heidelberg, 1992.
9. R. J. LeVeque. Numerical Methods for Conservation Laws. Birkhaeuser, Basel, 1999.
10. J. L. Lions. Exact controllability, stabilization and perturbations of distributed systems. SIAM Review, 30: 1–68, 1988.
11. K. Malanowski. On time–optimal control of a vibrating string (Polish). Arch. Automat. Telemech., 14: 33–44, 1969.
12. D. L. Russell. Nonharmonic Fourier series in the control theory of distributed parameter systems. J. Math. Anal. Appl., 18: 542–560, 1967.
13. V. V. Tikhomirov. The wave equation with a boundary control in the case of elastic fixing: I. Differential Equations, 38: 413–424, 2002.
14. E. Zuazua. Optimal and approximate control of finite-difference approximation schemes for the 1D wave equation. To appear.

An Application of PL Continuation Methods to Singular Arcs Problems

Pierre Martinon and Joseph Gergaud

ENSEEIHT-IRIT, UMR CNRS 5505, 2 rue Camichel, F-31071 Toulouse, France.
martinon,gergaud@enseeiht.fr

Summary. Among optimal control problems, singular arcs problems are interesting and difficult to solve with indirect methods, as they involve a multi-valued control and differential inclusions. Multiple shooting is an efficient way to solve this kind of problems, but typically requires some a priori knowledge of the control structure. We limit here ourselves to the case where the Hamiltonian is linear with respect to the control u, and primarily use a quadratic (u^2) perturbation of the criterion. The aim of this continuation approach is to obtain an approximate solution that can provide reliable information concerning the singular structure. We choose to use a PL (simplicial) continuation method, which can be more easily adapted to the multi-valued case. We will first present some convergence results regarding the continuation, and then study the numerical resolution of two example problems. All numerical experiments were conducted with the Simplicial package we developed.

1 Introduction

In indirect methods, applying the Pontryagin's Maximum Principle to a problem with singular arcs leads to a Boundary Value Problem with a differential inclusion. We denote this BVP with the following notations, that will be used throughout this paper. Let us denote x the state, p the costate, u the control, and φ the state-costate dynamics. If $y = (x, p) \in \mathbf{R}^n$ (n is thus twice the state dimension) and Γ denotes the set valued map of optimal controls, then one has

$$\text{(BVP)} \begin{cases} \dot{y}(t) \in \Phi(y(t)) = \varphi(y(t), \Gamma(y(t))) & \text{a.e. in } [0, t_f] \\ \text{Boundary Conditions} \end{cases}$$

First, we want to obtain some information regarding the structure of the solutions, ie the number and approximate location of singular arcs. We use for this a perturbation of the original problems by a quadratic (u^2) term, as done for instance in [10]. We will show some convergence properties of this continuation scheme, that are mainly derived from the results in [4] by J.P. Aubin and A. Cellina, and [12] by A.F. Filippov. This continuation method

involves following the zero path of a multi-valued homotopy, which is why we chose a simplicial method rather than a differential continuation method (extensive information about continuation methods can be found in E. Allgower and K. Georg [2, 3], and M.J. Todd [20, 21]). Then we will study the practical path following method, and some of the numerical difficulties we encountered, which led us to introduce a discretized formulation of the Boundary Value Problem. Finally, we use the information from these two continuation approaches to solve the optimal control problems with a variant of the multiple shooting method. We will study in this paper two examples (from [11] and [10]) in parallel, whose similar behaviour indicates that our approach is not too problem-dependant.

1.1 First Example

The first example we consider is a fishing problem described in [11]. The state ($x(t) \in \mathbf{R}$) represents the fish population, the control ($u(t) \in \mathbf{R}$) is the fishing activity, and the objective is to maximize the fishing product over a fixed time interval:

$$(P_1) \begin{cases} \text{Max } \int_0^{10} (E - \frac{c}{x(t)})\, u(t)\, U_{max}\, dt \\ \dot{x}(t) = rx(t)(1 - \frac{x(t)}{k}) - u(t)\, U_{max} \\ 0 \leq u(t) \leq 1 \quad \forall t \in [0, 10] \\ x(0) = 70.10^6 \quad x(10) \text{ free} \end{cases}$$

with $E = 1$, $c = 17.5\ .10^6$, $r = 0.71$, $k = 80.5\ .10^6$ and $U_{max} = 20\ .10^6$.

First we transform this problem into the corresponding minimization problem with the objective Min $\int_0^{10} (\frac{c}{x(t)} - E)\, u(t)\, U_{max}\, dt$ (note that the numerical values of the problem are such that we always have $\frac{c}{x(t)} - E < 0$, which corresponds to a positive fishing product). Applying the Maximum Principle of Pontryagin then gives the following hamiltonian system for the state x and costate p:

$$\begin{cases} \dot{x}(t) = r\ x(t)\ (1 - \frac{x(t)}{k}) - u(t) U_{max} \\ \dot{p}(t) = \frac{c}{x^2(t)}\ u(t) U_{max} - p(t)\ r\ (1 - \frac{2x(t)}{k}). \end{cases}$$

In terms of the switching function ψ

$$t \in [0, 10] \mapsto \psi(t) = \frac{c}{x(t)} - E - p(t),$$

the Hamiltonian minimization gives the optimal control

$$\begin{cases} u^*(t) = 0 \quad if\ \psi(t) > 0 \\ u^*(t) = 1 \quad if\ \psi(t) < 0 \\ u^*(t) \in [0, 1] \quad if\ \psi(t) = 0. \end{cases}$$

Over a singular arc, the relations $\dot{\psi} = 0$ and $\ddot{\psi} = 0$ give the expression of the singular control (t is omitted for clarity)

$$u^*_{singular} = \frac{k\ r}{2(\frac{c}{x} - p)U_{max}}(\frac{c}{x} - \frac{c}{k} - p + \frac{2px}{k} - \frac{2px^2}{k^2}).$$

More precisely, the control actually vanishes in the equation $\dot{\psi} = 0$, so it is necessary to use the second derivative of the switching function. On a side note, these relations also lead to $\dot{x} = \dot{p} = 0$, so the state, costate, and therefore also the control are constant over a singular arc for this problem, which is of course not a general property. Another remark is the important difference of magnitude between the state and costate (about 10^8), which requires the use of a proper scaling.

1.2 Second Example

The second example is the quadratic regulator problem studied by Y. Chen and J. Huang [10]:

$$(P_2)\begin{cases} \text{Min } \frac{1}{2}\int_0^5 (x_1^2(t) + x_2^2(t))\ dt \\ \dot{x_1}(t) = x_2(t) \\ \dot{x_2}(t) = u(t) \\ -1 \leq u(t) \leq 1 \quad \forall t \in [0,5] \\ x(0) = (0,1) \quad x(5) \text{ free} \end{cases}$$

We have the state and costate dynamics

$$\begin{cases} \dot{x_1}(t) = x_2(t) \\ \dot{x_2}(t) = u(t) \\ \dot{p_1}(t) = -x_1(t) \\ \dot{p_2}(t) = -p_1(t) - x_2(t) \end{cases}$$

and the following switching function and optimal control:

$$\psi(t) = p_2(t),$$

$$\begin{cases} u^*(t) = -\text{ sign } p_2(t) & \text{if } \psi(t) \neq 0 \\ u^*(t) \in [-1,1] & \text{if } \psi(t) = 0. \end{cases}$$

In that case again, the control disappears from the equation $\dot{\psi} = 0$, but the relation $\ddot{\psi} = 0$ still gives the singular control $u^*_{singular}(t) = x_1(t)$.

2 Continuation Method: A Quadratic Perturbation

Solving these problems directly by single shooting is not possible due to the presence of singular arcs, so we use a continuation method. Like the approach in [10], we try to regularize these problems with a quadratic $(u^2(t))$ perturbation, and consider the following objectives:

$$\text{Min} \int_0^{10} (\frac{c}{x(t)} - E)(u(t) - (1-\lambda)u^2(t))U_{max}\ dt, \quad \lambda \in [0,1],$$

$$\text{Min}\ \frac{1}{2}\int_0^5 (x_1^2(t)\ +\ x_2^2(t))\ +\ (1-\lambda)u^2(t)\ dt, \quad \lambda \in [0,1].$$

For problem (P_1), as mentioned before the term $\frac{c}{x(t)} - E$ is always negative, so the minus sign before $(1-\lambda)u^2(t)U_{max}$ actually results in "adding" a quadratic term, as for problem (P_2). We then obtain two families of boundary value problems parametrized by λ, denoted by $(BVP_1)_\lambda$ and $(BVP_2)_\lambda$ respectively. The original problems correspond to $\lambda = 1$. For $\lambda = 0$, the problems are much more regular, and can be solved directly by single shooting without any difficulties. The principle of the continuation method is to start from the solution at $\lambda = 0$ to attain $\lambda = 1$, where we have the solution of the original problem. The first idea is to try to solve a sequence of problems $(BVP)_{\lambda_k}$, with a sequence (λ_k) ranging from 0 to 1. However, finding a suitable sequence (λ_k) is often problematic in practice, and here the low regularity of the homotopy (for $\lambda = 1$ especially) led us to rather consider a full path following method. More precisely, if we note S_λ the shooting function related to the parametrized problems, we will follow the zero path of the homotopy $h : (z, \lambda) \mapsto S_\lambda(z)$, from $\lambda = 0$ to $\lambda = 1$. There are two main families of path following methods: Predictor-Corrector methods, which are fast but require that the zero path be C^2, and the slower but more robust Piecewise Linear methods. In the present case, we have to deal for $\lambda = 1$ with a multi-valued homotopy, which is why we use a simplicial method, whose general principle will be described below.

Remark. From now on, we will use the subscripts $_1$ and $_2$ for notations specific to Problems 1 and 2, and keep unsubscripted notations for the general case.

2.1 Hamiltonian Minimization Properties

We begin with some results concerning the Hamiltonian minimization, that were presented in [14]. We first recall a standard result (in the following we keep the notation $y = (x, p)$, with y of dimension n):

Theorem 1. *Assume that $U \subset \mathbf{R}^m$ is a convex compact set with nonempty interior, and that the Hamiltonian function $H : [a,b] \times \mathbf{R}^n \times U \to \mathbf{R}$ is continuous and convex with respect to the control u. We note $\Gamma(t,x,p)$ the set of solutions of $\min_{u\in U}\ H(t,x,p,u)$. Then Γ has nonempty compact convex values.*

Lemma 1. *A compact-valued map G is upper-semicontinuous (in the sense of Berge [5, p.114]) if and only if for all sequence (x_k) that converges to x, $(G(x_k))$ converges to $G(x)$ according to*

$$\forall \epsilon > 0,\ \exists k_0 > 0 \text{ such that } \forall k > k_0,\ G(x_k) \subset G(x) + \epsilon B(0,1),$$

with $B(0,1)$ standing for the closed unit ball of center 0 and radius 1.

Proof. See [17, p.66]. □

Recall that a function $f : \mathbf{R}^m \times \overline{\mathbf{R}}$ is said to be inf-compact (in the sense of [19]) if for all $(y, a) \in \mathbf{R}^m \times \mathbf{R}$, the set $\{u \in \mathbf{R}^m : f(u) - (u|y) \leq a\}$ is compact. Here $\overline{\mathbf{R}}$ denotes the extended-real line and $(\cdot|\cdot)$ stands for the usual inner product in $\mathbf{R}^m$.

Lemma 2. *Let $(f_k)_{k\in\mathbf{N}}$ be proper convex lower-semicontinuous functions defined over $\mathbf{R}^m$. Assume the following assumptions:*

(i) $(f_k)_{k\in\mathbf{N}}$ converges pointwisely to f,

(ii) $int(domf) \neq \emptyset$,

(iii) f is inf-compact.

Then,

$$\lim_{k\to\infty} \inf_{u\in\mathbf{R}^m} f_k(u) = \inf_{u\in\mathbf{R}^m} f(u)$$

and, $\forall \epsilon > 0$, there is $k_0 \in \mathbf{N}$ such that

$$argmin_{u\in\mathbf{R}^m}\ f_k(u) \subset argmin_{u\in\mathbf{R}^m}\ f(u) + \epsilon\ B(0,1) \quad \forall k \geq k_0.$$

Proof. See [19], page I.3.54. □

Theorem 2. *Consider the same hypotheses as in Theorem 1. Then, Γ has the following convergence property: if (t_k, x_k, p_k) is a sequence that converges to (t, x, p), then*

(i) $inf_{u\in\mathbf{R}^m}\ H(t_k, x_k, p_k, u) \to inf_{u\in\mathbf{R}^m}\ H(t, x, p, u)$ as $k \to +\infty$,

(ii) $\forall \epsilon > 0,\ \exists k_0 > 0$ s.t. $\forall k > k_0,\ \Gamma(t_k, x_k, p_k) \subset \Gamma(t, x, p) + \epsilon\ B(0, 1)$.

Proof. Let (t_k, x_k, p_k) be a sequence that converges to (t, x, p). We note $f_k(u) = H(t_k, x_k, p_k, u) + \delta(u/U)$ and $f(u) = H(t, x, p, u) + \delta(u/U)$ (where $\delta(u/U) = 0$ if $u \in U$, and $+\infty$ if $u \notin U$). For both example problems, it is clear from the expression of the Hamiltonian (see below) that the (f_k) are convex and lower-semicontinuous. Let us check the assumptions of Lemma 2:
(i) If $u \notin U$, then $f_k(u) = f(u) = +\infty\ \forall k$; if $u \in U$, then $f_k(u) = H(t_k, x_k, p_k, u)$, and as H is continuous we have $H(t_k, x_k, p_k, u) \to H(t, x, p, u)$, so $f_k(u) \to f(u)$.
(ii) One has $\text{int}(\text{dom } f) = \text{int}(U) \neq \emptyset$.
(iii) If $v \in \mathbf{R}^m$ and $a \in \mathbf{R}$, then $\{u\ |H(t, x, p, u) + \delta(u/U) - (u|v) \leq a\} = U \cap \{u\ |H(t, x, p, u) - (u|v) \leq a\}$. This set is compact because it is a closed subset of the compact set $U\mathbf{R}^m$. This shows the inf-compacity of f.
Now Lemma 2 proves the theorem. □

Corollary 1. *Consider the same hypotheses as in Theorem 1. Then, Γ is upper-semicontinuous (in short, usc).*

Proof. Theorem 1, Lemma 1 and Theorem 2 give this result. □

Remark. If H is strictly convex, then we have the well-known property (see e.g. [13, Theorem 6.1, p.75] and [6]) that u^* is a continuous function (as Γ is then a continuous function).

Back to the two examples, we have $U_1 = [0,1]$, $U_2 = [-1,1]$, and the Hamiltonians are (t is omitted for clarity):

$$H_1(t,x,p,u) = (\frac{c}{x} - E)(u - (1-\lambda)u^2)U_{max} + p(rx\,(1 - \frac{x}{k}) - u\,U_{max}),$$

$$H_2(t,x,p,u) = \frac{1}{2}(x_1^2 + x_2^2 + (1-\lambda)u^2) + p_1 x_2 + p_2 u.$$

Both H_1 and H_2 are continuous, and convex with respect to u (for Problem 1 we can numerically check a posteriori that $\frac{c}{x(t)} - E < 0 \quad \forall t \in [0,10]$). So for both problems (P_1) and (P_2), Theorem 1 and Corollary 1 apply, thus Γ_1 and Γ_2 are upper-semicontinuous, and non empty compact convex valued. These properties will be useful for the following convergence results concerning the continuation. We can also note that for $\lambda < 1$, both Hamiltonians are strictly convex, thus the optimal control are continuous functions.

2.2 Convergence Properties

The following results were presented in [8], and are primarily derived from the books by J.P. Aubin and A. Cellina [4], and by A.F. Filippov [12], whose notations we will keep. In particular,

$\overline{co}\ K$ = closed convex hull of K,
$M^\delta = \{m : d(m,M) \leq \delta\}$.

Definition 1 ([12]). *A function y is called a δ-solution of $\dot{y}(t) \in F(t,y(t))$, with $F : [a,b] \times \mathbf{R}^n \rightarrow \mathbf{R}^n$ an upper-semicontinuous set-valued map, if over an interval $[a,b]$, y is absolutely continuous and*

$$\dot{y}(t) \in F_\delta(t,y(t)) = [\overline{co}\ F(t^\delta, y^\delta)]^\delta$$

where $F(t^\delta, y^\delta) = \cup_{s \in t^\delta, z \in y^\delta}\ F(s,z)$.

Lemma 3. *Let (y_k) be a sequence in $AC_n([a,b])$ such that:*
(i) $\forall t \in [a,b]$, $\{y_k(t)\}_k$ is relatively compact,
(ii) $\exists\ l$ such that $|\dot{y}_k(t)| \leq l$ almost everywhere in $[a,b]$.
Then, there exists a subsequence still noted (y_k) that converges uniformly to an absolutely continuous function $y : [a,b] \rightarrow \mathbf{R}^n$, and for which the sequence $(\dot{y}_k)$ converges weakly- to $\dot{y}$ in $L_n^\infty([a,b])$.*

Proof. The proof follows the principle of the demonstration of Theorem 4, pp. 14-15 in [4]. The sequence (y_k) is equicontinuous as

$$|y_k(t') - y_k(t'')| = \int_{t'}^{t''} \dot{y}_k(t)\ dt\ \leq l|t' - t''|.$$

The Arzelá-Ascoli theorem implies the existence of a subsequence, still noted (y_k), that converges uniformly to y in $C_n([a,b])$. Moreover, $\dot{y}_k \in \overline{B}(0,c) \subset$

$L_n^\infty([a,b])$, with $L_n^\infty([a,b])$ being the dual of $L_n^1([a,b])$. Thus the Alaoglu theorem implies that this closed ball is weak-* compact. As $L_n^1([a,b])$ is separable this closed ball is metrizable for the weak-* topology (cf [7]). Then there exists a subsequence, still noted $(\dot{y}_k)$, that converges weakly-* to z in $L_n^\infty([a,b])$. We now have to prove that y is absolutely continuous and that $\dot{y} = z$. First, y_k is absolutely continuous, thus

$$y_k(t') - y_k(t'') = \int_{t'}^{t''} \dot{y}_k(s)\ ds. \tag{1}$$

The sequence (y_k) converges uniformly to y, so the left hand side converges to $y(t') - y(t'')$. As $(\dot{y}_k)$ converges weakly-* to z, for all components i we have

$$< \mathbf{1}, \dot{y}_{k,i} >_{L^1,L^\infty} = \int_a^b \dot{y}_{k,i}(s)ds \to < \mathbf{1}, z_i >_{L^1,L^\infty} = \int_a^b z_i(s)ds$$

(where $\mathbf{1}$ is the constant map equal to 1). So the right hand side of (1) converges to $\int_{t'}^{t''} z(s)ds$. We have then

$$y(t') - y(t'') = \int_{t'}^{t''} z(s)ds$$

with z in $L_n^\infty([a,b])$, thus in $L_n^1([a,b])$. This means that y is absolutely continuous and that $\dot{y}(t) = z(t)$ almost everywhere. □

Theorem 3. *Let (y_k) be a sequence in $AC_n([a,b])$ that converges to y and verifies $\dot{y}_k(t) \in K$ for all k and t, with K compact. Then y is absolutely continuous and $\dot{y}_k(t) \in \overline{co}\ K$ for all t.*

Proof. The proof is based on Filippov's Lemma 13, p. 64 of [12]. □

Theorem 4. *Let F be a nonempty compact convex valued map, defined on an open set $\Omega \subset \mathbf{R}^{n+1}$. Let (y_k) be a sequence of δ_k-solutions defined on $[a,b]$ that converges uniformly to $y : [a,b] \to \mathbf{R}^n$ when $\delta_k \to 0$, and such that the graph of y is in Ω. Then y is a solution of the differential inclusion $y(t) \in F(t, y(t))$.*

Proof. See Filippov's Lemma 1, p. 76 of [12] □

Lemma 4. *Let $\varphi : \Omega \times [0,1] \to \mathbf{R}^n$, with Ω an open subset of $\mathbf{R}^n$, be a set-valued map verifying*

(i) φ is upper-semicontinuous with nonempty compact convex values,

(ii) $\varphi_\lambda = \varphi(\cdot, \lambda)$ is a is piecewise $- C^1$ function for $0 \leq \lambda < 1$.

Let us assume that the solutions of $\dot{y}_\lambda(t) = \varphi_\lambda(y(t))$ remain in a fixed compact K and are defined on an interval $[0, t_f]$. Then y_λ is a δ-solution of the differential inclusion $\dot{y}(t) \in \varphi(y(t), 1)$, and δ tends to 0 when $\lambda \to 1$.

Proof. φ is usc at $(y^*,1)$ for all $(y^*,1) \in K$, thus for $\epsilon = \delta$, there exists η such that for all $|y-y^*| < \eta_{y^*} < \delta$, $|\lambda - 1| < \eta_{y^*} < \delta$, we have $\varphi_\lambda(y,\lambda) \in \varphi(y^*,1)^\epsilon$. Thus $K \subset \cup_{y^* \in K} B(y^*, \eta_{y^*})$ and as K is compact, we have $K \subset \cup_{i=1}^q B(y_i,\eta_i)$. For $\epsilon = \delta, \forall y_i, \exists \eta_i$ $|y-y_i| < \eta_i < \delta$ and $|\lambda - 1| < \eta_i < \delta$, $\varphi_\lambda(y) \in \varphi(y_i,1)^\epsilon$. For all $y \in K$, there exists y_i such that $y \in B(y_i,\eta_i)$ and thus

$$\varphi_\lambda(y) \in (\varphi(y_i,1))^\epsilon \subset (\varphi(y_i^\eta,1))^\epsilon \subset (\varphi(y^\delta,1))^\epsilon.$$

Then for all λ such that $|\lambda - 1| < \eta$ and for all $y \in K$, we have $\varphi_\lambda(y) \in (\varphi(y^\delta,1))^\delta$. □

Theorem 5. *Let us assume that the solutions of $(BVP)_\lambda$ remain in a fixed compact of $[0,t_f] \times K$, $K \subset \Omega$, with Ω an open subset of $\mathbf{R}^n$. Then from any sequence (y_{λ_k}) of solutions of $(BVP)_{\lambda_k}$, such that $\lambda_k \to 1$ as $k \to +\infty$, we can extract a subsequence (y_k) verifying:*

(i) (y_k) converges uniformly to y solution of $(BVP)_1$,

(ii) $(\dot{y}_k)$ converges weakly- to $\dot{y}$ in $L_n^\infty([0,t_f])$.*

Proof. φ is usc, thus $\varphi(K,[1-\epsilon,1])$ is compact. There exists l such that $|\dot{y}_\lambda(t)| < l$ for $\lambda \in [1-\epsilon,1]$. The y_λ are absolutely continuous, and Lemma 3 says that we can extract a subsequence (y_k) that converges uniformly to y, and such that $(\dot{y}_k)$ converges weakly-* to $\dot{y}$ in $L_n^\infty([0,t_f])$. As per Lemma 4, (y_k) is a δ_k-solution. $\varphi(y,1)$ is non empty compact convex valued, so Theorem 3 says that y is a solution of the differential inclusion $\dot{y}(t) \in \varphi(y(t),1)$. Initial and terminal conditions can be written as $h_0(y(0)) = 0$ and $h_f(y(t_f)) = 0$ with h_0 and h_f continuous. The uniform convergence of (y_k) implies that y verifies the boundary conditions, thus y is a solution of $(BVP)_1$. □

Corollary 2. *Under the hypotheses of Theorem 5, assume that $\dot{x} = f(t,x,u)$ provides a control of the form $u = S(t,x) + R(t,x)\dot{x}$, with R and S continuous and R linear. Consider the subsequence $y_k = (x_k,p_k)$ from Theorem 5, and let $u_k = S(t,x_k) + R(t,x_k)\dot{x}_k$. Then (u_k) converges weakly-* in $L_n^\infty([0,tf])$.*

Proof. See [9], proof of Proposition 3.2, pp. 551-552. □

Back to our families of problems $(BVP_1)_\lambda$ and $(BVP_2)_\lambda$, we have the state-costate dynamics

$$\varphi_1(y(t),u(t),\lambda) = \begin{pmatrix} r\ x(t)\ (1 - \frac{x(t)}{k}) - u(t)U_{max} \\ \frac{c}{x^2(t)}\ (u(t) - (1-\lambda)u^2(t))U_{max} - p(t)\ r\ (1 - \frac{2x(t)}{k}) \end{pmatrix},$$

$$\varphi_2(y(t),u(t),\lambda) = \begin{pmatrix} x_2(t) \\ u(t) \\ x_1(t) \\ x_2(t) \end{pmatrix}.$$

We consider the set valued dynamic $\Phi(y(t),\lambda) = \varphi(y(t),\Gamma(y(t)),\lambda)$. From the expression of φ_1 and φ_2, and the fact that Γ_1 and Γ_2 are usc with nonempty

compact convex values, we obtain that Φ_1 and Φ_2 are also usc with nonempty compact convex values. Now we make the assumption that the solutions of $(BVP_1)_\lambda$ and $(BVP_2)_\lambda$ remain in some fixed compact sets, which has been validated by the numerical experiments. Then Theorem 5 applies and gives the convergence result for the continuation approach.

Moreover, we have the following expression of the control:

$$\begin{cases} u_\lambda(t) = \frac{1}{U_{max}}(r\ x_\lambda(t)\ (1 - \frac{x_\lambda(t)}{k}) - \dot{x}_\lambda(t)) & \text{for Problem 1} \\ u_\lambda(t) = \dot{x}_{2_\lambda}(t) & \text{for Problem 2.} \end{cases}$$

Thus Corollary 2 applies and gives the convergence for the control.

2.3 PL Continuation - Simplicial Method

We will here recall very briefly the principle of a PL continuation method. Extensive documentation about path following methods can be found in E. Allgower and K. Georg [2, 3], as well as Todd [20, 21], to mention only a few. The idea of a continuation is to solve a difficult problem by starting from the known solution of a somewhat related, but easier problem. By related we mean here that there must exist an application h, called a *homotopy*, with the right properties connecting the two problems. For the following definitions we consider that h is a function, the multi-valued case will be treated afterwards.

PL continuation methods actually follow the zero path of the homotopy h by building a piecewise linear approximation of h, hence their name. Towards this end, the search space is subdivided into cells, most often in a particular way called a *triangulation* in *simplices*. This is why PL continuation methods are often referred to as simplicial methods. The main advantage of this approach is that it puts extremely low requirements on the homotopy h: as no derivatives are used, continuity is in particular sufficient, and should not even be necessary in all cases.

First, we recall some useful definitions.

Definition 2. *A simplex is the convex hull of $n+1$ affinely independent points (called the vertices) in $\mathbf{R}^n$, while a k-face of a simplex is the convex hull of k vertices of the simplex (note: k is typically omitted for n-faces, which are just called faces).*

Definition 3. *A triangulation is a countable family T of simplices of $\mathbf{R}^n$ verifying:*
- *The intersection of two simplices of T is either a face or empty,*
- *T is locally finite (a compact subset of $\mathbf{R}^n$ meets finitely many simplices).*

Definition 4. *We call labeling a map l that associates a value to the vertices v_i of a simplex. We label here the simplices by the homotopy h: $l(v^i) = h(z^i, \lambda^i)$, where $v^i = (z^i, \lambda^i)$. Affine interpolation on the vertices thus gives a PL approximation h_T of h.*

Definition 5. *A face* $[v_1, .., v_n]$ *of a simplex is said completely labeled iff it contains a solution* v_ϵ *of the equation* $h_T(v) = \epsilon$ *for all sufficiently small* $\epsilon > 0$ *(where* $\epsilon = (\epsilon, .., \epsilon^n)$*).*

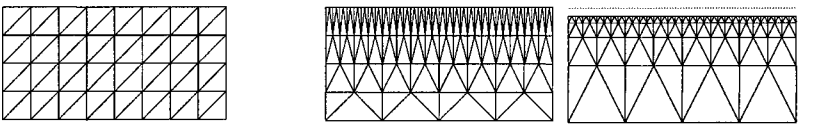


Fig. 1. Illustration of some well known triangulations of $\mathbf{R} \times [0, 1]$ ($[0, 1[$ for J_3): Freudenthal's uniform K_1, and Todd's refining J_4 and J_3

Lemma 5. *Each simplex possesses either zero or exactly two completely labeled faces (being called a* transverse *simplex in the latter case).*

Proof. See [2], Chapter 12.4. □

The constructive proof of this property, which gives the other completely labeled face of a simplex that already has a known one, is often referred to as PL step, linear programming step, or lexicographic minimization. Then there exists a unique transverse simplex that shares this second completely labeled face, that can be determined via the pivoting rules of the triangulation.

A simplicial algorithm thus basically follows a sequence of transverse simplices, from a given first transverse simplex with a completely labeled face at $\lambda = 0$, to a final simplex with a completely labeled face at $\lambda = 1$ (or $1 - \epsilon$ for some refining triangulations that never reach 1, such as J_3), which contains an approximate solution of $h(z, 1) = 0$.

For a multi-valued homotopy h, we have the following convergence property.

Theorem 6. *We consider a PL continuation algorithm using a selection of* h *for labeling and a refining triangulation of* $\mathbf{R}^n \times [0, 1[$ *(such as* J_3 *for instance). We make two assumptions regarding the path following:*

(i) all the faces generated by the algorithm remain in $K \times [0, 1]$*, with* K *compact.*

(ii) the algorithm does not go back to $\lambda = 0$*.*

Then, if h *is usc with compact convex values, the algorithm generates a sequence* (z_i, λ_i) *such that* $\lambda_i \to 1$*, and there exists a subsequence still noted* (z_i, λ_i) *converging to* $(z, 1)$ *such that* $0 \in h(z, 1)$*.*

Proof. The proof comes from [1], chapter 4, page 56. □

For the two examples under consideration, the two assumptions concerning the path following are numerically verified for both problems. However, the assumption of the homotopy convexity only holds for Problem 1, but not for Problem 2 a priori.

2.4 Path Following - Singular Structure Detection

In order to initialize the continuation, we need to solve both problems for $\lambda = 0$, which is easily done by single shooting from an extremely simple initial point ($x(0) = -0.1$ and $x(0) = (0,0)$ respectively). The objective shown is the original, unperturbed criterion, and the results are summarized on Table 1. For both examples, the path following goes smoothly at first, and the switching function and control evolution as λ increases is quite interesting, as shown on Figures 2 and 3.

Table 1. Solutions for $\lambda = 0$

	λ	z^*	$\lvert S_0(z^*)\rvert$	objective	iter	time
Problem 1	0	$-4.0935\ 10^{-2}$	$3.6295\ 10^{-16}$	69374046	39	$< 1s$
Problem 2	0	$(1.2733, 2.2715)$	$3.3596\ 10^{-14}$	0.4388	134	$< 1s$

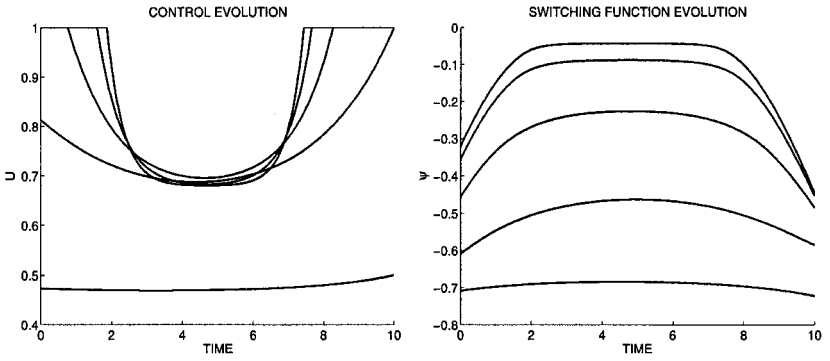


Fig. 2. Problem 1: Control and Switching function for $\lambda = 0, 0.5, 0.75, 0.9, 0.95$

We can see that for both problems, the switching function ψ comes closer to 0 over some time intervals, which strongly suggests the presence of singular arcs at the solution for $\lambda = 1$. For the first problem, the control structure seems to be *regular-singular-regular*, with the singular arc boundaries near $[2, 7.5]$, and for the second problem *regular-singular*, the singular arc beginning around 1.5. Meanwhile, we can see that outside the suspected singular arcs the control tends to a bang-bang structure coherent with the necessary conditions, more precisely $+1$ before and after the arc for the first problem, and -1 before the arc for the second problem. An interesting fact is that the control keeps on taking intermediate values over the time intervals where ψ tends to 0, which confirms the assumption of a singular arc. On these two examples, the continuation based on the quadratic perturbation gives a strong

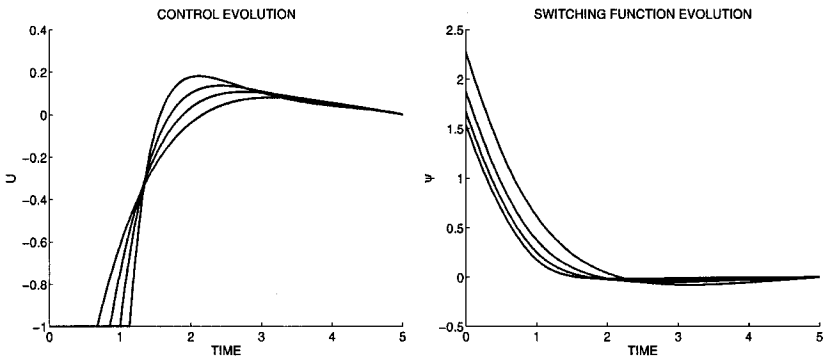


Fig. 3. Problem 2: Control and Switching function for $\lambda = 0, 0.5, 0.75, 0.9$

indication about the control structure, with an approximate location of the singular arcs. So as far as the detection of the singular structure is concerned, this approach seems rather effective.

2.5 Numerical Difficulties - Different Control Structures

However, as λ tends to 1, the path following encounters some difficulties: above a certain point, the PL approximation of the shooting function becomes increasingly inaccurate. We can also note that from this point on, the objective value does not improve any longer (here again, the objective values displayed correspond to the original non-perturbed problems, thus Max $\int_0^{10} (E - c/x(t))\ u(t)\ U_{\max}\ dt$ and Min $\frac{1}{2} \int_0^5 (x_1^2(t) + x_2^2(t))\ dt$). Fig.4 shows the evolution of the homotopy norm and the criterion value along the zero path.

Difficulties are often expected at the end of continuation strategies, and for simplicial methods there exists for instance some refining triangulations (such as J_3 or J_4) whose meshsize decreases progressively, in order to ensure an accurate path following near the convergence. Yet in our case, using this kind of techniques only delays this degradation a little, and does not prevent it from appearing eventually. The instability threshold is problem dependant: it appears at best (via refining triangulations) beyond $\lambda = 0.975$ for Problem 1 and $\lambda = 0.95$ for Problem 2.

The reason behind this phenomenon can be found if one looks at the control structures corresponding to the vertices of the completely labeled faces (which are supposed to contain a zero of the PL approximation of the shooting function). Depending on the vertices, we find two different control structures: the interval on which the switching function is close to 0, that we call a *pseudo singular arc*, is not stable. At some point, the switching function leaves the proximity of 0 and increases in absolute value, either with positive or negative

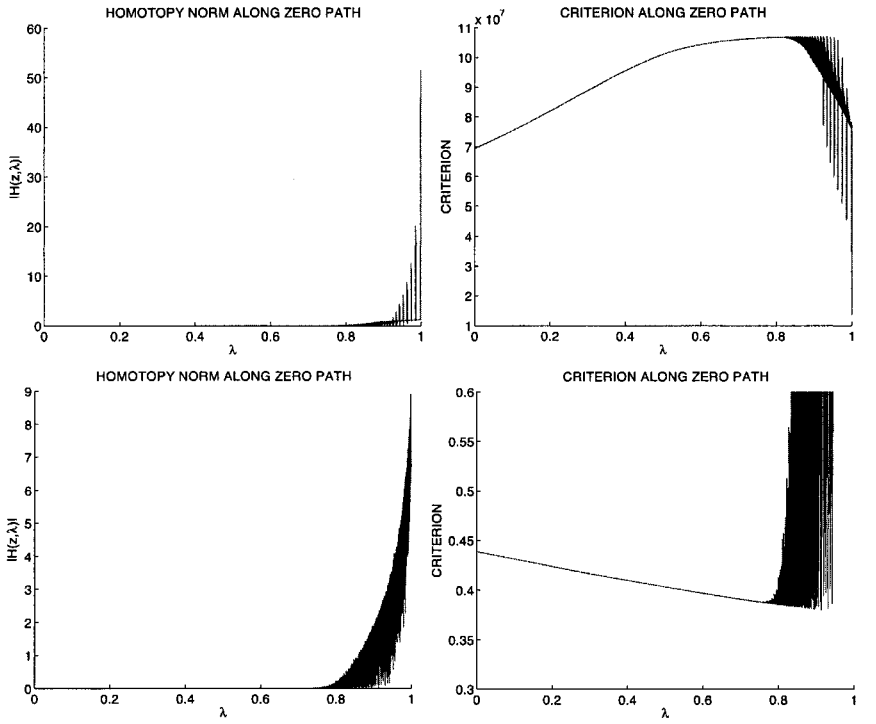


Fig. 4. Homotopy norm and objective value along zero paths, for Problem 1 and 2

sign. Depending on the sign of the switching function at the exit of the pseudo singular arc, we obtain two different control structures, either with "crossing" or "turning back". What happens is that these two structures keep appearing among the vertices of the labeled faces, however small a meshsize we use, which is why refining triangulations are useless. We also note that this instability of the switching function near zero becomes worse as λ tends to 1: the length of the pseudo singular arc decreases as the exit occurs earlier and earlier, as illustrated on Fig.5.

At the convergence for $\lambda = 1$, all that is left from the pseudo singular arc is a contact point, once again with two possible control structures depending on the sign of the switching function after it reaches 0. For both problems, the control and switching functions (but also the state and costate) are identical for the two structures before the contact point. After that point, which corresponds to the beginning of the supposed singular arc, the switching function goes either positive or negative, with the two corresponding bang-bang controls. More precisely, if the switching function crosses 0 and changes sign, there is a control switch, while it remains the same if the switching function turns back with the same sign after the contact point. Anyway, in both cases

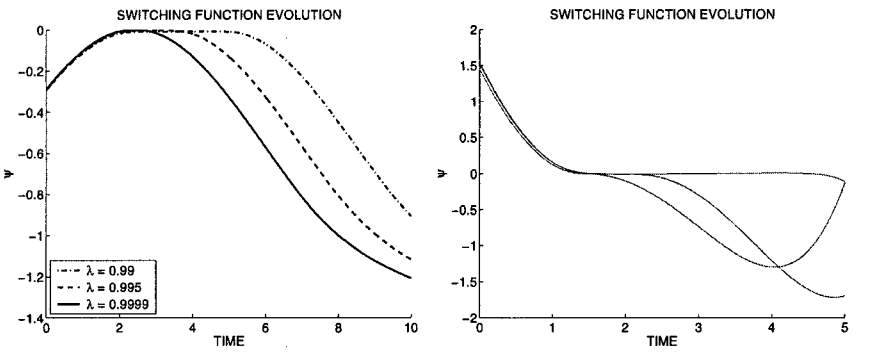


Fig. 5. Switching function evolution for Problem 1 and 2 ($\lambda = 0.99, 0.995, 0.9999$, and $\lambda = 0.95, 0.975, 0.9999$ respectively)

we have lost the singular structure at the convergence. Figures 6 and 7 show these two distinct control structures for each problem.

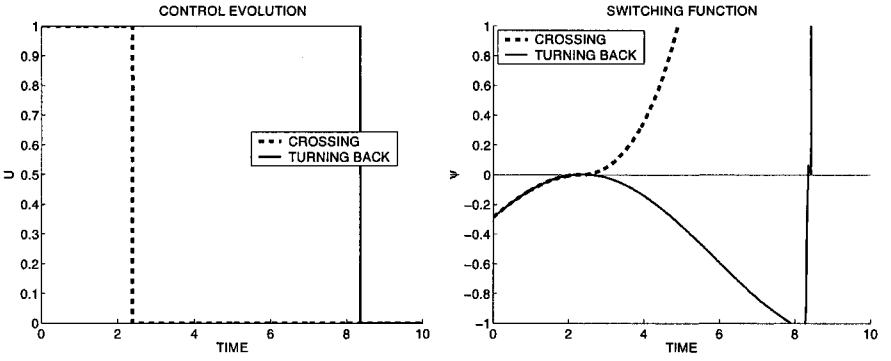


Fig. 6. Problem 1: Control structures according to switch exit sign

The existence of these two very close (with respect to the shooting function unknown z) and yet completely different control structures corresponds to a discontinuity of the shooting function at the solution, which is illustrated on Fig.8 for both Problem 1 (first graph) and Problem 2 (second and third graphs).

We use here a basic Runge Kutta 4th order method with 1000 integration steps. We tried various other fixed step integration methods, such as Euler, Midpoint, Runge Kutta 2 or 3, and increased the number of steps to 10000. We also used variable step integrators, namely Runge Kutta Fehlberg 4-5, Dormand Prince 8-5-3, and Gragg Bulirsch Stoer extrapolation method (see [15]), with similar results.

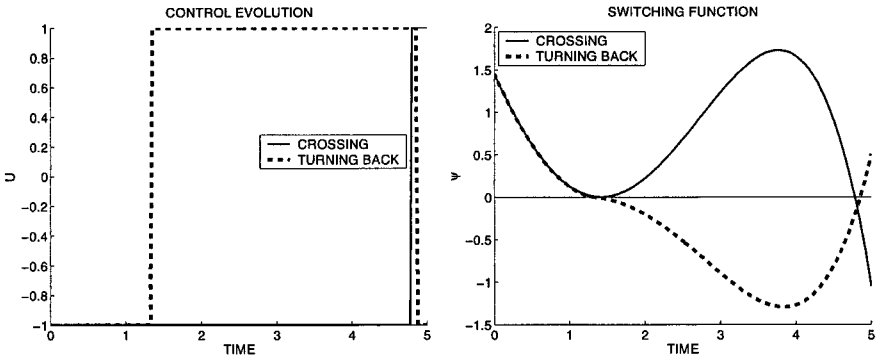


Fig. 7. Problem 2: Control structures according to switch exit sign

For Problem 1, the path following always converges to the correct z^* and locates this discontinuity precisely, which is not surprising as Theorem 6 applies. This is not the case for Problem 2, and the experiments indeed show that the path following can converge to different solutions, depending on the integration used. We also note that the convergence is more difficult to attain for Problem 2, as we often had to use the less accurate triangulation J_4 instead of J_3.

3 Continuation: Discretized BVP Formulation

We now try to circumvent the previously encountered difficulties by discretizing the equations of the Boundary Value Problem. We use here a basic Euler scheme for the state and costate, and consider a piecewise constant control. The values of the state and costate at the interior discretization nodes become additional unknowns of the shooting function, while we have the following matching conditions at these nodes:

$$\begin{cases} x_{i+1} - (x_i + h\frac{\partial x}{\partial t}(t_i, x_i, p_i, u_i^*)) \\ p_{i+1} - (p_i + h\frac{\partial p}{\partial t}(t_i, x_i, p_i, u_i^*)) \end{cases}$$

where the optimal control u_i^* is obtained from (x_i, p_i) by the usual necessary conditions. The idea is, that even if the control obtained on the singular arc is irrelevant, we hope to have a good approximation of the state and costate values. This formulation corresponds to a particular case of multiple shooting, with a 1-step Euler integration between two successive discretization nodes. Thanks to this integration choice, the discretized version of the shooting function is compact convex valued. This allows us to hope a good behaviour of the path following, according to Theorem 6.

Here are the discretized shooting function unknown and value layouts:

Unknown z | IVP unknown at t_0 | (x^1, p^1) | (x^2, p^2) | ... |

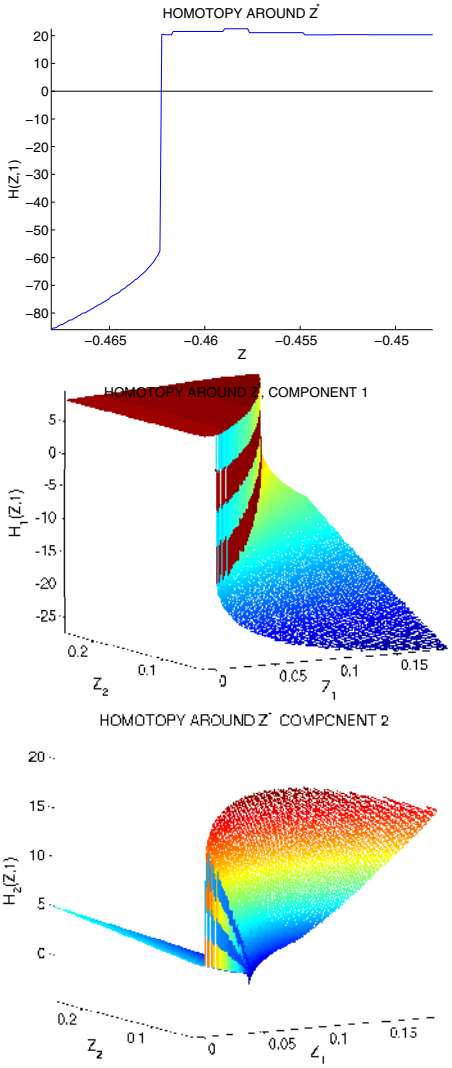


Fig. 8. Shooting functions discontinuity at $\lambda = 1$ for Problem 1 and 2

- IVP unknown at t_0 (same as in single shooting method)
- values of (x^i, p^i) at interior times t_i

Value $S_D(z)$	$Match_{cond}(t_1)$	$Match_{cond}(t_2)$	...	Conditions at t_f

- matching conditions at interior times
- terminal and transversality conditions at t_f (same as single shooting)

Remark. A major drawback of this formulation is that the full state and costate are discretized. This drastically limits the number of discretization nodes, else

the high dimension of the unknown leads to prohibitive execution times. As a side effect, this also puts some restrictions on the use of small meshsizes or refining triangulations, for the same computational cost reasons.

Then we apply the same continuation with the quadratic perturbation as before. Once again, solving both problems for $\lambda = 0$ is done immediately by single shooting, and we follow the zero path until $\lambda = 1$. The instability observed with the single shooting method does not occur. Here on Fig.9 are the solutions obtained with 50 discretization nodes for Problem 1 and 20 for Problem 2 (whose state and costate are in $\mathbf{R}^2$ instead of $\mathbf{R}$). This time

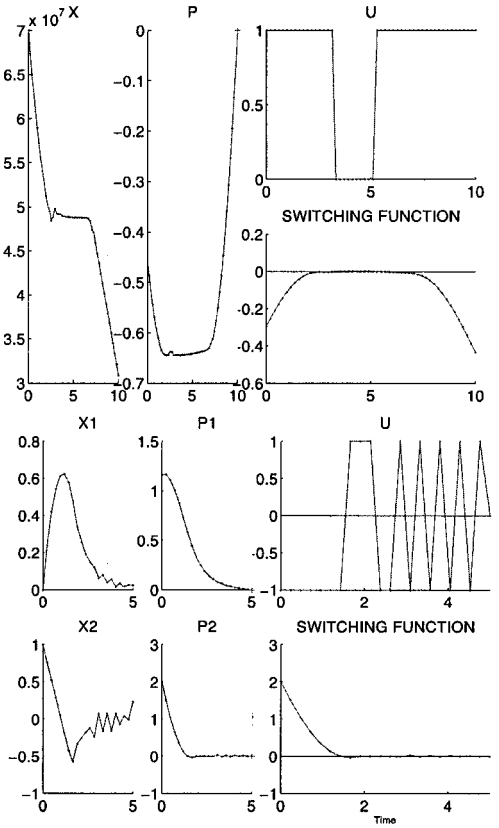


Fig. 9. Discretized BVP solutions at $\lambda = 1$ for Problem 1 and 2

both switching functions clearly show the presence of a singular arc, located near $[2, 7]$ and $[1.5, 5]$ respectively. We note that the switching function for Problem 1 is much closer to zero than the best solution we could obtain with the previous approach. But now, an annoying fact is the presence of some

oscillations located within the bounds of the singular arcs. Trying to get rid of these oscillations by conventional means, such as smaller meshsizes and/or refining triangulations, or increasing the number of discretization nodes, turns out to be ineffective, especially for Problem 2.

These difficulties might come from the expression of the control, which is still given by the necessary conditions. On a time interval $[t_i, t_{i+1}]$ located within a singular arc, let us assume that the continuation has found the correct values of the state and costate (x_i, p_i) and (x_{i+1}, p_{i+1}). The switching function $\psi(t_i)$ should be near zero as we are supposed to be on a singular arc, but numerically it will not be exactly zero, mostly due to the rough discretization scheme used. The necessary conditions then give a bang-bang control u_i that is different from the actual singular control u_i^*, so the matching conditions on the state and costate at t_{i+1} may not be satisfied. So on singular arcs, the algorithm may deviate to values of (x, p) that try to verify these incorrect matching conditions given by wrong control values.

If we look closer at the value of S_D at the solution, we indeed notice that non-zero components are found only for discretization times corresponding to singular arcs. Moreover, for Problem 2 matching errors only occur for x_2, whose derivative is the only one in which the control appears. Other components x_1, p_1, p_2, whose derivatives do not depend on u, always have correct matching conditions, even on the singular arc. For Problem 1, both derivatives of x and p involve the control, so it is not surprising that both matching conditions are non-zero on the singular arc. We also note that the sign of non-zero matching conditions changes accordingly to the sign of the switching function. This is coherent with the fact that it also corresponds to the switchings of the incorrect bang-bang control, and therefore possibly the changes of sign of $u_i - u_i^*$. All this confirms that these oscillations observed on singular arcs are related to the wrong control value given by the necessary conditions. Fig.10 shows these matching conditions, with the switching function.

4 Numerical Resolution

Now we have gathered some knowledge concerning the singular structure of the problems, and we try to solve them more precisely. Based on the solutions of the continuation with the continuous and discretized formulations, we will assume that we have the following control structures: *regular-singular-singular* for Problem 1, and *regular-singular* for Problem 2.

We use a variant of the classical multiple shooting method, that we call "structured shooting". It shares the same principle as the well known code BNDSCO from H.J. Oberle (see [18]), slightly simplified and adapted to the singular case instead of the state constraints. The control structure is here described by a fixed number of interior switching times, that correspond to the junction between a regular and a singular arc. This times $(t_i)_{i=1..n_{switch}}$ are part of the unknowns and must satisfy some switching conditions. Each

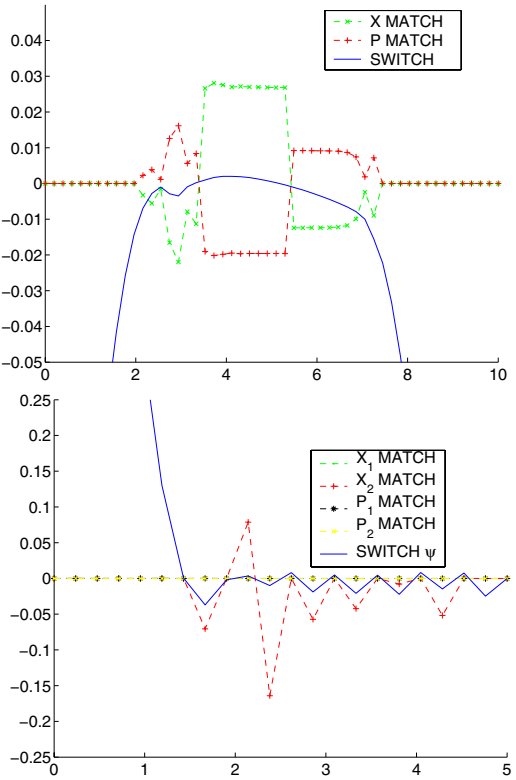


Fig. 10. Matching conditions at $\lambda = 1$ for Problems 1 and 2

control arc is integrated separately, and matching conditions must be verified at the switching times. The switching condition indicates a change of structure, that is the beginning or end of a singular arc where $\psi = 0$, and thus can be defined for instance as $\psi^2(t_i, x_i, p_i) = 0$. Matching conditions basically consist in state and costate continuity at the switching times.

Summary: structured shooting function unknown and value layout:

Unknown z | IVP unknown at t_0 | (x^1, p^1) | (x^2, p^2) | ... | Δ_1 | Δ_2 | ... |

- IVP unknown at t_0 (same as in single shooting method)
- values of (x^i, p^i) at interior switching times t_i
- switching times intervals Δ_i, such that $t_i = t_{i-1} + \Delta_i \quad , \quad \forall i \in [1..nswitch]$

Value $S_{Struct}(z)$ | $Switch_{cond}(t_1)$ | $Match_{cond}(t_1)$ | ... | Conditions at t_f |

- switching and matching conditions at interior times
- terminal and transversality conditions at t_f (same as single shooting)

Structured shooting initialization:
Based upon the solutions obtained with the two continuations, we have two switchings times for Problem 1, and one switching time for Problem 2. The structured shooting unknowns, and the initialization sets corresponding to the continuous and discretized continuations are summarized below on Table 2 and 3 (for the single shooting we use a solution for $\lambda = 0.95$, before instability occurs).

Table 2. Problem 1: control structure regular-singular-regular

Continuation	$p(0)$	$t_1; t_2$	(x^1, p^1)	(x^2, p^2)
$BVP_{0.95}$	-0.429	2.5 ; 7	$(4.996\ 10^7, -0.600)$	$(4.825\ 10^7, -0.587)$
BVP_D	-0.453	2.55 ; 7.06	$(4.839\ 10^7, -0.637)$	$(4.741\ 10^7, -0.621)$

Table 3. Problem 2: control structure regular-singular

Continuation	$(p_1(0), p_2(0))$	t_1	$(x^1\ ,\ p^1)$
$BVP_{0.95}$	(0.974,1.512)	1.5	$(0.398, -0.309\ ,\ 0.401, 0.00358)$
BVP_D	(1.167,2.024)	1.429	$(0.578, -0.429\ ,\ 0.586, 0.000505)$

Now we try to solve directly $S_{struct}(z) = 0$ with these initializations. For both problems, convergence is immediate for the two initializations. Fig.11 shows the solutions obtained, with the expected singular arcs. The solutions are the same for the two initialization sets, see Table 4 and 5.

Table 4. Solution comparison for Problem 1 for the two initialization sets.

Initialization	z^*	$t_1^*; t_2^*$	$\lvert S_{Struct}(z^*)\rvert$	objective	iter	time
$BVP_{0.95}$	-0.46225	2.3704 ; 6.9888	$1.1\ 10^{-13}$	106905998	110	$< 1s$
BVP_D	-0.46225	2.3704 ; 6.9888	$3.1\ 10^{-11}$	106905998	88	$< 1s$

Table 5. Solution comparison for Problem 2 for the two initialization sets.

Initialization	z^*	t_1^*	$\lvert S_{Struct}(z^*)\rvert$	objective	iter	time
$BVP_{0.95}$	(0.9422,1.4419)	1.4138	$2.4\ 10^{-14}$	0.37699	93	$< 1s$
BVP_D	(0.9422,1.4419)	1.4138	$9.2\ 10^{-15}$	0.37699	116	$< 1s$

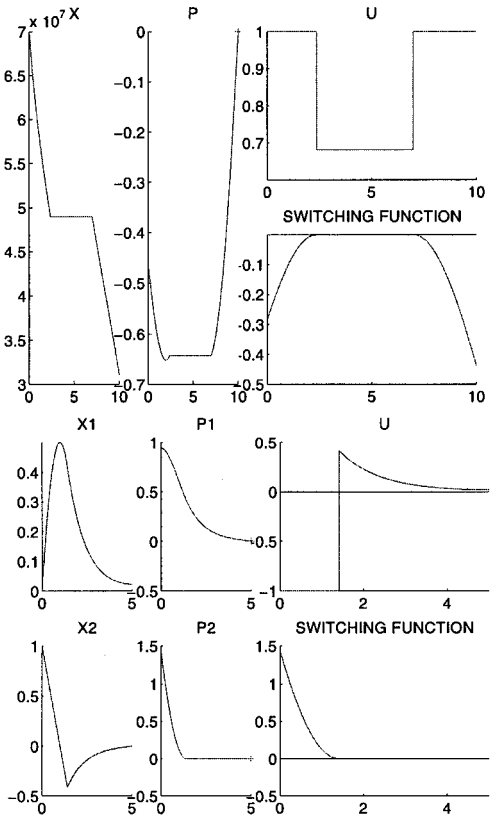


Fig. 11. Solutions obtained by Structured Shooting for Problems 1 and 2

Remark. It should be noted that the resolution can be quite sensitive with respect to the initial point. For Problem 1, a deviation of 0.1 of the costate values can be enough to prevent the convergence.

As a conclusion, we give here a comparison of the solutions obtained by the two continuation approaches (single shooting at $\lambda = 0.95$ and discretized BVP) with the reference solution from structured shooting, see Fig.12 and 13.

We can see that both continuation solutions are rather close to the reference solution for the state, costate and switching function. For Problem 1 the discretized solution is quite good, with very little oscillations, while the continuous solution at $\lambda = 0.95$ is less accurate for p and ψ. For Problem 2, the oscillations are much more important on the discretized solution, whereas the continuous solution at $\lambda = 0.95$ is very close to the reference. Concerning the control, the continuous formulation gives an acceptable approximation of the singular control, the differences being localized around the switching times,

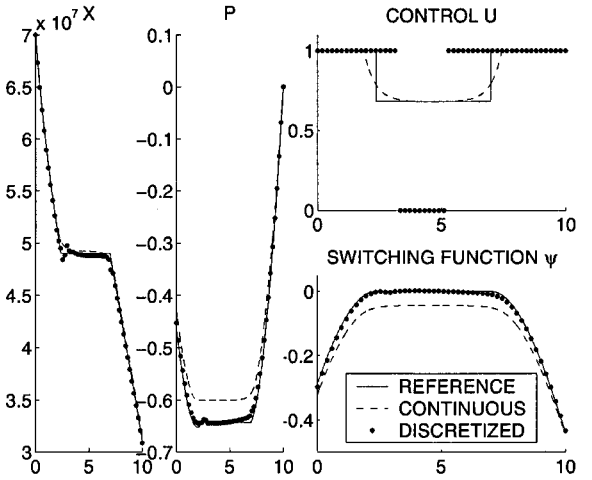


Fig. 12. Structured Shooting and continuation solutions comparison for Problem 1

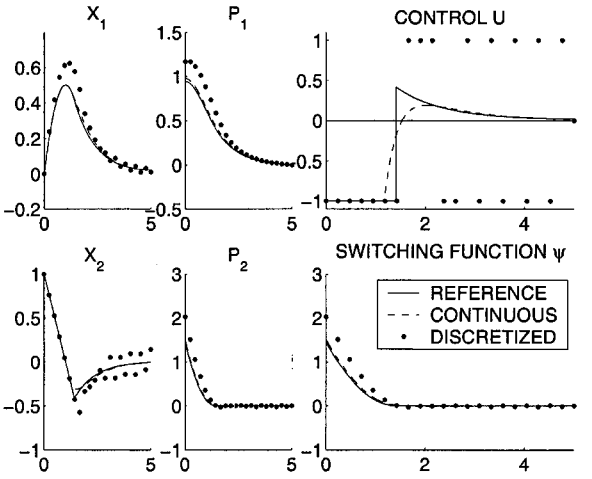


Fig. 13. Structured Shooting and continuation solutions comparison for Problem 2

which is not surprising.

Notes and Numerical precisions:

- There is no path following for structured shooting: we solve $S_{Struct}(z) = 0$ directly, which is possible because we have a quite good initial guess.
- For non-discretized formulations (single and structured shooting), a basic fixed step Runge Kutta 4th order integration was used, with 1000 discretization steps for both problems. As said before, we obtained similar results with

other integration methods, either with fixed or variable step.
• Tests were run on a PC workstation (2.8GHz Pentium 4), using a build of the Simplicial package compiled with ifc (Intel Fortran Compiler).
• All numerical experiments were made using the Simplicial package we wrote, which implements a PL continuation for optimal control problems via indirect methods (www.enseeiht.fr/lima/apo/simplicial/, see also [16])

5 Conclusions and Perspectives

Starting with no a priori knowledge about the control structure, the two formulations (single shooting and discretized BVP) of the continuation allowed us to detect the singular arcs accurately. With this information and the approximate solutions obtained, we were able to solve the problems with a variant of multiple shooting.

Concerning the oscillations encountered with the discretized formulation, one could think of using the expression of the singular control when the switching function is close to 0, instead of the incorrect bang-bang control given by the hamiltonian minimization. However, the solutions obtained depend heavily on the practical implementation of this "close to 0" condition, which can artificially force a singular arc...

Another interesting idea is to discretize the control, in the same fashion as direct shooting (or "semi-direct") methods. This consists in integrating the state and costate with an Euler scheme and a piecewise constant control, whose value on the discretization nodes are part of the unknowns. Some conditions would be enforced on these values, such as satisfying the Hamiltonian minimization in the regular case and the singular control expression in the singular case.

Finally, it would be interesting to try to adapt the methods we used here for singular arcs to the case of state constraints, which also lead to low regularity problems.

References

1. E. Allgower and K. Georg. Simplicial and continuation methods for approximating fixed points and solutions to systems of equations. Siam Review, 22(1):28–85, 1980.
2. E. Allgower and K. Georg. Numerical Continuation Methods. Springer-Verlag, Berlin, 1990.
3. E. Allgower and K. Georg. Piecewise linear methods for nonlinear equations and optimization. J. Comput. Applied Math., 124:245–261, 2000.
4. J.P. Aubin and A. Cellina. Differential Inclusion. Springer-Verlag, 1984.
5. C. Berge. Espaces Topologiques. Dunod, Paris, 1959.
6. J.F. Bonnans. The shooting algorithm for optimal control problems: a review of some theoretical and numerical aspects. Tech. Rep., Univ. El Manar (Tunis), 2002. Lect. Notes, DEA de Mathématiques Appliquées de l'ENIT.

7. H. Brezis. Analyse Fonctionnelle. Masson, 1983.
8. JB. Caillau, R. Dujol, J. Gergaud, T. Haberkorn, P. Martinon, J. Noailles, and D. Preda. Mise au point d'une méthode de résolution efficace pour les problèmes de contrôle optimal à solution "bang-bang" - application au calcul de trajectoires à poussée faible. Tech. Rep., ENSEEIHT-IRIT, UMR CNRS 5505, Toulouse, 2004. Rapport de fin de phase 2 - Contrat 02/CNES/0257/00 - DPI 500.
9. JB. Caillau, J. Gergaud, and J. Noailles. 3D geosynchronous transfer of a satellite: continuation on the thrust. J. Optim. Theory Appl., 118(3):541–565, 2003.
10. Y. Chen and J. Huang. A numerical algorithm for singular optimal control synthesis using continuation methods. Optimal Control Applications & Methods, 15:223–236, 1994.
11. C.W. Clark. Mathematical Bioeconomics. Wiley, 1976.
12. A.F. Filippov. Differential Equations with Discontinuous Righthand Sides. Kluwer Academic Publishers, Dordrecht, 1988.
13. W.H. Flemming and R.W. Rishel. Deterministic and Stochastic Optimal Control. Springer-Verlag, 1975.
14. J. Gergaud. Résolution numérique de problèmes de commande optimale à solution Bang-Bang par des méthodes homotopiques simpliciales. PhD thesis, Institut National Polytechnique de Toulouse, 1989.
15. E. Hairer, S.P. Nørsett, and G. Wanner. Solving Ordinary Differential Equations I: Nonstiff Problems (Second Revised Edition). Springer-Verlag, Berlin, 1993.
16. P. Martinon and J. Gergaud. Simplicial package v1.0 - User Guide. Tech. Rep.(RT/APO/04/04), ENSEEIHT-IRIT, UMR CNRS 5505, 2004.
17. H. Nikaido. Convex Structures and Economic Theory. Academic Press, 1968.
18. H.J. Oberle and W. Grimm. BNDSCO - A program for the numerical solution of optimal control problems. Tech. Rep. 515, Inst. for Flight System Dynamics, Oberpfaffenhofen, German Aerospace Research Establishment DLR, 1989.
19. R. Robert. Contributions à l'analyse non linéaire. PhD thesis, Univ. Scientifique et Médicale de Grenoble et Institut National Polytechnique de Grenoble, 1976.
20. M.J. Todd. The computation of fixed points and applications. In: Lectures Notes in Economics and Mathematical Systems, vol. 124 (VII). Springer-Verlag, Heidelberg, 1976.
21. M.J. Todd. Union Jack triangulations. In: Fixed Points: Algorithms and Applications, pp. 315–336. Karamadian, Academic Press, New York, 1977.

On an Elliptic Optimal Control Problem with Pointwise Mixed Control-State Constraints *

Christian Meyer[1] and Fredi Tröltzsch[2]

[1] Institut für Mathematik, Technische Universität Berlin, D-10623 Berlin, Str. des 17. Juni 136, Germany. cmeyer@math.tu-berlin.de
[2] Institut für Mathematik, Technische Universität Berlin, D-10623 Berlin, Str. des 17. Juni 136, Germany. troeltz@math.tu-berlin.de

Summary. A nonlinear elliptic control problem with pointwise control-state constraints is considered. Existence of regular Lagrange multipliers, first-order necessary and and second-order sufficient optimality conditions are derived. The theory is verified by numerical examples.

1 Introduction

In this paper, we consider the following semilinear elliptic optimal control problem with distributed control and pointwise mixed control-state constraints

$$\text{(P)} \quad \begin{cases} \text{minimize } J(y,u) := \dfrac{1}{2}\displaystyle\int_\Omega (y-y_d)^2\,dx + \dfrac{\kappa}{2}\int_\Omega u^2(x)\,dx \\ \text{subject to } -\Delta y(x) + d(y(x)) = u(x) \quad \text{in } \Omega \\ \qquad\qquad \partial_\nu y(x) + y(x) = 0 \qquad \text{on } \Gamma \end{cases} \tag{1}$$

$$\text{and } y_a(x) \le \lambda\, u(x) + y(x) \le y_b(x) \quad \text{a.e. in } \Omega, \tag{2}$$

where $\Omega \subset \mathbb{R}^N$, $N = \{2,3\}$, is a bounded domain with $C^{0,1}$-boundary Γ and ν denotes the outward unit normal The function $d : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and monotonic increasing. Furthermore, the second derivative d'' is assumed to be locally Lipschitz-continuous. Moreover, $\kappa > 0$ and $\lambda \neq 0$ are real numbers, and the bounds y_a and y_b are fixed functions in $L^\infty(\Omega)$ with $y_a(x) \le y_b(x)$ a.e. in Ω.

This paper is a contribution to the theory of distributed optimal control problems with pointwise state-constraints. The associated numerical analysis

* Supported by the DFG Research Center "Mathematics for key technologies" (FZT 86) in Berlin.

is known to be quite complicated, since the Lagrange multipliers for the state-constraints are in general regular Borel measures. We refer, for instance, to Casas [4] for first-order necessary optimality conditions, Casas, Tröltzsch and Unger [7] for second-order sufficient conditions and to Bergounioux, Ito and Kunisch [1] or Bergounioux and Kunisch [3] for associated numerical methods.

The analysis is often simpler for problems with mixed pointwise control-state constraints, since Lagrange multipliers are more regular there. For the elliptic case with quadratic objective and linear equation, this has been shown in the recent paper [10]. However, the corresponding proofs are quite technical.

Here, we consider a particular class of constraints, where the analysis can be developed by a simple trick: Locally, the problem (P) is converted to one with pointwise box-constraints, where the analysis is easy to perform. We will show that problem (P) has regular Lagrange multipliers in $L^\infty(\Omega)$. In view of this, we are able to derive first- and second-oder optimality conditions for (P). Moreover, we report on associated numerical tests.

It should be underlined that we investigate the problem for a fixed parameter $\lambda \neq 0$. Though λ is used as a small regularization parameter in the numerical tests, we do not study here the complicated question of convergence of optimal solutions and multipliers as $\lambda \to 0$. The problem (P) is interesting in itself for λ fixed.

Remark 1. The theory below also works for $-\Delta y(x) + y(x) + d(y(x)) = u(x)$ in Ω, $\partial_\nu y(x) = 0$ on Γ instead of (1). This is the case studied in the numerical tests in Section 5.

2 Standard Results

In this section, we recall some well-known results on (P). We consider y in the state space $Y = H^1(\Omega) \cap C(\bar{\Omega})$ and the control u in $L^2(\Omega)$. Moreover, we introduce the control-to-state operator $G : L^2(\Omega) \to Y$ that assigns y to u. The following result is well known, [4]:

Theorem 1. *Under the assumptions on d and Ω stated in Section 1, the state equation (1) admits for all $u \in L^2(\Omega)$ exactly one solution $y = G(u) \in Y$, and the estimate*

$$\|y\|_{H^1(\Omega)} + \|y\|_{C(\bar{\Omega})} \leq c_\infty \|u\|_{L^2(\Omega)} \tag{3}$$

holds true with a constant c_∞ that only depends on Ω.

Due to $\dim \Omega \leq 3$, we obtain the following results for the derivatives of G (cf.[6]):

Lemma 1. *Under the assumptions on d, G is twice continuously Fréchet differentiable from $L^2(\Omega)$ to Y. Its first derivative, denoted by $w = G'(u)\,h$, $h \in L^2(\Omega)$, is given by the solution of the linearized equation*

$$\begin{aligned} -\Delta w + d'(y)\, w &= h \quad \text{in } \Omega \\ \partial_\nu w + w &= 0 \quad \text{on } \Gamma \end{aligned} \tag{4}$$

with $y = G(u)$. Moreover, the second derivative $z = G''(u)[u_1, u_2]$ solves the equation

$$\begin{aligned} -\Delta z + d'(y)\, z &= -d''(y)\, y_1\, y_2 \quad \text{in } \Omega \\ \partial_\nu z + z &= 0 \quad \text{on } \Gamma \end{aligned} \tag{5}$$

with y as defined above, and $y_i = G'(u)u_i$, $i = 1, 2$.

The next theorem states the existence of an optimal solution for (P).

Theorem 2. *If the admissible set is not empty, then* (P) *admits at least one global solution.*

Proof. The proof is more or less standard. In all what follows, we denote the global solution by $(\bar{y}, \bar{u})$, where $\bar{y} = G(\bar{u})$ and $\bar{u}$ is said to be an optimal control. By $\kappa > 0$, we find a bounded minimizing sequence $\{u_n\} \subset L^2(\Omega)$ and we can assume without loss of generality $u_n \rightharpoonup \bar{u}$, $n \to \infty$. By Theorem 1, the associated sequence $\{y_n\}$ is bounded in $H^1(\Omega)$, hence we are justified to assume $y_n \to \bar{y}$ in $L^2(\Omega)$. Together with the boundedness in $C(\bar{\Omega})$ that follows from (3), this yields $d(y_n) \to d(\bar{y})$ in $L^2(\Omega)$, $\bar{y} = G(\bar{u})$. The optimality of $\bar{u}$ is a standard conclusion. □

We should mention that our theory does not rely on this existence result. It is also applicable to any *local* solution $\bar{u}$.

Remark 2. Obviously, all admissible controls must be bounded and measurable, since $y_a, y_b \in L^\infty(\Omega)$ and $y \in C(\bar{\Omega})$ imply $u \in L^\infty(\Omega)$ because of the constraint (2).

3 First-Order Conditions - Regular Multipliers

We start by introducing the reduced objective functional by

$$J(y, u) = J(G(u), u) =: f(u).$$

Thus, (P) is equivalent to mimimizing $f(u)$ subject to

$$y_a(x) \le \lambda\, u(x) + (G(u))(x) \le y_b(x) \quad \text{a.e. in } \Omega. \tag{6}$$

Since J is of tracking type, it is twice continuously differentiable. Together with the differentiability of G (cf. Lemma 1), this yields the following lemma.

Lemma 2. *Under the assumptions of Lemma 1, f is twice continuously Fréchet differentiable from $L^2(\Omega)$ to $\mathbb{R}$. Its first derivative is given by*

$$f'(u)h = (\kappa u + q\,,\, h)_{L^2(\Omega)}, \tag{7}$$

where q solves the adjoint equation

$$\begin{aligned} -\Delta q + d'(y)\, q &= y - y_d \quad \text{in } \Omega \\ \partial_\nu q + q &= 0 \qquad\quad \text{on } \Gamma, \end{aligned} \tag{8}$$

with $y = G(u)$. For the second derivative, we obtain

$$f''(u)[u_1, u_2] = (y_1\,,\, y_2)_{L^2(\Omega)} + \kappa\,(u_1\,,\, u_2)_{L^2(\Omega)} - \int_\Omega d''(y)\, y_1\, y_2\, q\, dx, \tag{9}$$

where y and q are as defined above, and $y_i = G'(u)u_i$, $i = 1, 2$.

Proof. Although the arguments are standard, we recall the main ideas for convenience of the reader. From $f(u) = J(G(u), u) = 1/2\,\|G(u) - y_d\|^2_{L^2(\Omega)} + \kappa/2\,\|u\|^2_{L^2(\Omega)}$, we get

$$f'(u)h = (y - y_d\,,\, w)_{L^2(\Omega)} + \kappa(u, h)_{L^2(\Omega)},$$

where $y = G(u)$ and $w = G'(u)h$ denotes the weak solution of the linearized equation (4) with the right hand side h. Now, choosing q as test function in the weak formulation of (4), we obtain

$$\int_\Omega \nabla w \cdot \nabla q\, dx + \int_\Omega d'(y)\, w\, q\, dx + \int_\Gamma w\, q\, ds = \int_\Omega h\, q\, dx.$$

On the other hand, we insert w in the weak formulation of equation (8):

$$\int_\Omega \nabla q \cdot \nabla w\, dx + \int_\Omega d'(y)\, q\, w\, dx + \int_\Gamma q\, w\, ds = \int_\Omega (y - y_d)\, w\, dx.$$

Subtracting one equation from the other finally yields $(y - y_d\,,\, w)_{L^2(\Omega)} = (h\,,\, q)_{L^2(\Omega)}$. Applying again the chain rule, we arrive at

$$\begin{aligned} f''(u)[u_1, u_2] = {} & (G'(u)u_1\,,\, G'(u)u_2)_{L^2(\Omega)} + (G(u) - y_d\,,\, G''(u)[u_1, u_2])_{L^2(\Omega)} \\ & + \kappa(u_1\,,\, u_2)_{L^2(\Omega)}. \end{aligned}$$

A similar discussion as above, where $z = G''(u)[u_1, u_2]$ denotes the weak solution of (5), then gives $(y - y_d\,,\, z)_{L^2(\Omega)} = -(d''(y)\, y_1\, y_2\,,\, q)_{L^2(\Omega)}$. □

Remark 3. Notice that, for a given right hand side in $L^2(\Omega)$, equation (8) admits a solution q in Y, since the differential operator in (8) has the same form as the one in (4).

Next, we substitute $\lambda u + G(u) = v$ and consider the associated nonlinear equation

$$\lambda u + G(u) = v \tag{10}$$

for a given v in a neighborhood of $\bar{v} = \lambda \bar{u} + G(\bar{u})$. This substitution will be used for the transformation of (P) into a purely control-constrained problem. By the implicit function theorem, we show under a suitable regularity assumption that (10) admits a unique solution in a neighborhood of the optimal solution $\bar{u}$ for all given $v \in L^2(\Omega)$ in a neighborhood of $\bar{v}$. To this aim, we introduce an auxiliary operator $T : L^2(\Omega) \times L^2(\Omega) \to L^2(\Omega)$ by $T(u, v) = \lambda u + G(u) - v$. Associated with T is a mapping $K : v \mapsto u$ that is implicitly defined by $T(K(v), v) = 0$. To apply the implicit function theorem, we need that

$$\frac{\partial T}{\partial u}(\bar{u}, \bar{v})u = \lambda u + G'(\bar{u})u,$$

is invertible, where $\bar{v} = \lambda \bar{u} + G(\bar{u})$. Due to Lemma 1, $G'(\bar{u})$ is continuous from $L^2(\Omega)$ to $H^1(\Omega) \cap C(\bar{\Omega})$. Let us consider $G'(\bar{u})$ with range in $L^2(\Omega)$ and denote this operator by $\boldsymbol{G}$. Because of the compact embedding of $H^1(\Omega)$ in $L^2(\Omega)$, $\boldsymbol{G}$ is compact, and hence $\boldsymbol{G}$ represents a Fredholm operator that has only countably many eigenvalues accumulating at 0. Here and in the following, $I : L^2(\Omega) \to L^2(\Omega)$ denotes the identity.

We rely on the following REGULARITY ASSUMPTION:

(R) The prescribed $\lambda \neq 0$ is not an eigenvalue of $-\boldsymbol{G}$, i.e. the equation $\lambda u + G'(\bar{u})u = 0$ admits only the trivial solution.

Note that this is fulfilled for all $\lambda > 0$. From the theory of Fredholm operators, it is known that the equation

$$\frac{\partial T}{\partial u}(\bar{u}, \bar{v})u = \lambda u + G'(\bar{u})u = f$$

is uniquely solvable for given $f \in L^2(\Omega)$, provided that (R) is satisfied. Thus, $\frac{\partial T}{\partial u}(\bar{u}, \bar{v})$ is continuously invertible by the Banach theorem, and hence the implicit function theorem gives the existence of open balls $B_{r_1}(\bar{u})$, $B_{\rho_1}(\bar{v})$ in $L^2(\Omega)$ such that for all $v \in B_{\rho_1}(\bar{v})$, there is exactly one $u \in B_{r_1}(\bar{u})$ with $T(u, v) = 0$. Therefore, by the definition of T, equation (10) has exactly one solution $u \in B_{r_1}(\bar{u})$ for all $v \in B_{\rho_1}(\bar{v})$. Notice that K is of class C^2 since T is twice continuously Fréchet differentiable in $L^2(\Omega)$ with respect to u.

Lemma 3. *The first- and second-order derivatives of $K : L^2(\Omega) \to L^2(\Omega)$ are given by*

$$K'(v) = \Big(\lambda I + G'(K(v))\Big)^{-1}, \tag{11}$$

$$K''(v)[v_1, v_2] = -\Big(\lambda I + G'(K(v))\Big)^{-1} G''(K(v))[K'(v)v_1, K'(v)v_2]. \tag{12}$$

Proof. As K is implicitly defined by $T(K(v), v) = 0$, the equation $\lambda K(v) + G(K(v)) = v$ holds true for all v in a neighborhood of $\bar{v}$. Differentiating on both sides yields

$$\lambda K'(v) + G'(K(v))K'(v) = I, \tag{13}$$

which implies (11). Next, we apply both sides of (13) to v_1 and differentiate in the direction v_2. One obtains

$$\lambda K''(v)[v_1, v_2] + G''(K(v))[K'(v)v_1, K'(v)v_2] + G'(K(v))K''(v)[v_1, v_2] = 0.$$

Resolving for $K''(v)[v_1, v_2]$ immediately gives (12). □

With these results at hand, we can convert (P), at least locally around $\bar{u}$, into an optimization problem in the variable v by substituting $\lambda u + G(u) = v$. For the objective functional, we obtain

$$J(y, u) = f(u) = f(K(v)) =: F(v),$$

where F is defined at least on $B_{\rho_1}(\bar{v})$. Local optimality of $\bar{u}$ implies the existence of an open ball $B_{r_2}(\bar{u})$ in $L^2(\Omega)$ such that $f(\bar{u}) \leq f(u)$ for all $u \in B_{r_2}(\bar{u})$ with $y_a(x) \leq \lambda u(x) + y(x) \leq y_b(x)$. This yields

$$F(\bar{v}) \leq F(v) \tag{14}$$

for all $v \in L^2(\Omega)$ satisfying $y_a(x) \leq v(x) \leq y_b(x)$ a.e. in Ω and $\|v - \bar{v}\|_{L^2(\Omega)} < \rho_2$ with a sufficiently small $\rho_2 > 0$. This ρ_2 is taken so small so that $\rho_2 \leq \rho_1$ and $u = K(v) \in B_{r_2}(\bar{u})$. Thus, $\bar{v}$ is the optimal solution of

$$\text{(PV)} \quad \text{minimize } F(v) \text{ subject to } v \in V_{ad},\ v \in B_{\rho_2}(\bar{v}),$$

with an admissible set defined by

$$V_{ad} := \{v \in L^2(\Omega) \,|\, y_a(x) \leq v(x) \leq y_b(x) \text{ a.e. in } \Omega\}.$$

Now, we are able to derive the following standard result.

Lemma 4. *Assume that* (R) *is fulfilled. Then the variational inequality*

$$F'(\bar{v})(v - \bar{v}) \geq 0 \tag{15}$$

holds true for all $v \in V_{ad}$.

Proof. Since V_{ad} is convex, we have for arbitrary $v \in V_{ad}$ that $v_t = \bar{v} + t(v - \bar{v}) \in V_{ad}\ \forall t \in [0, 1]$. Moreover, we find $\|v_t - \bar{v}\|_{L^2(\Omega)} < \rho_2$ if t is sufficiently small. Thus, (14) yields $[F(\bar{v} + t(v - \bar{v})) - F(\bar{v})]/t \geq 0$. Since f and K are Fréchet differentiable, the same holds for F. Thus, passing to the limit $t \downarrow 0$ implies (15). □

By the Riesz theorem, the functional $F'(\bar{v}) \in L^2(\Omega)^*$ can be identified with a function from $L^2(\Omega)$. Let us denote this function by μ, i.e.

$$F'(\bar{v})v = \int_\Omega \mu(x)\, v(x)\, dx. \tag{16}$$

Furthermore, we define nonnegative functions μ_a, $\mu_b \in L^2(\Omega)$ by

$$\begin{aligned} \mu_a(x) &= \mu(x)_+ = \frac{1}{2}(\mu(x) + |\mu(x)|), \\ \mu_b(x) &= \mu(x)_- = \frac{1}{2}(-\mu(x) + |\mu(x)|). \end{aligned} \tag{17}$$

Then, $\mu(x) = \mu_a(x) - \mu_b(x)$ and identifying $F'(\bar{v})$ with μ implies

$$F'(\bar{v}) + \mu_b - \mu_a = 0. \tag{18}$$

We show that the functions μ_a, μ_b, that have been defined by (17), are Lagrange multipliers for the control-state constraints (2). To see this, let us first set up the optimality system that should be satisfied at $(\bar{y}, \bar{u})$. We derive it in a formal way by the following Lagrange function $\mathcal{L} : Y \times L^2(\Omega) \times H^1(\Omega) \times L^2(\Omega)^2 \to \mathbb{R}$:

$$\begin{aligned} \mathcal{L}(y,u,p,\omega) = J(y,u) - \int_\Omega \nabla y \cdot \nabla p\, dx - \int_\Omega d(y)\, p\, dx - \int_\Gamma y\, p\, ds + \int_\Omega u\, p\, dx \\ + \int_\Omega \big(\mu_b(\lambda u + y - y_b) + \mu_a(y_a - \lambda u - y)\big)\, dx \end{aligned} \tag{19}$$

with $\omega := (\mu_a, \mu_b)$. Note that the last integral is well defined because of $\mu_a, \mu_b \in L^2(\Omega)$. The optimality system consists of $\partial\mathcal{L}/\partial y = 0$, $\partial\mathcal{L}/\partial u = 0$ and the complementary slackness conditions. We show that this is the expected optimality system for $(\bar{y}, \bar{u})$ following from the variational inequality (15) for $\bar{v}$. Straightforward computations give that $\partial\mathcal{L}/\partial y(\bar{y}, \bar{u}, p, \omega)y = 0$ for all $y \in H^1(\Omega)$ is equivalent to the adjoint equation

$$\begin{aligned} -\Delta p + d'(\bar{y})\, p &= \bar{y} - y_d + \mu_b - \mu_a \quad \text{in } \Omega \\ \partial_\nu p + p &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{20}$$

Analogously, $\partial\mathcal{L}/\partial u(\bar{y}, \bar{u}, p, \omega)u = 0$ for all $u \in L^2(\Omega)$ corresponds to

$$\kappa\, \bar{u} + p + \lambda(\mu_b - \mu_a) = 0. \tag{21}$$

In the following, we will show that (20) and (21), together with the complementary slackness condition

$$(\mu_a\,,\, y_a - \lambda\, \bar{u} - \bar{y})_{L^2(\Omega)} = (\mu_b\,,\, \lambda\, \bar{u} + \bar{y} - y_b)_{L^2(\Omega)} = 0, \tag{22}$$

indeed follow from the variational inequality (15).

Theorem 3. *If $\bar{u}$ is locally optimal with associated state $\bar{y}$, then there exist nonnegative Lagrange multipliers $\mu_a \in L^\infty(\Omega)$ and $\mu_b \in L^\infty(\Omega)$ and an associated adjoint state $p \in H^1(\Omega) \cap C(\bar{\Omega})$ such that the adjoint equation* (20)*, the condition* (21)*, and the complementary slackness conditions* (22) *are satisfied.*

Proof. We show that μ_a, μ_b defined by (17) do this. Moreover, we verify $\mu_a, \mu_b \in L^\infty(\Omega)$. To this end, we first have to transfer all expressions in terms of v to such in terms of (y, u).

(i) Adjoint equation and condition (21): We start with equation (18) where we express F' in terms of f and u. We recall $F(v) = f(K(v))$. By the chain rule, it holds $F'(\bar{v})v = f'(K(\bar{v}))K'(\bar{v})v$. Hence, (18) is equivalent to

$$f'(K(\bar{v}))K'(\bar{v})v + (\mu_b - \mu_a\,,\, v)_{L^2(\Omega)} = 0 \quad \forall\, v \in L^2(\Omega).$$

By substituting $u = K'(\bar{v})v$ and $\bar{u} = K(\bar{v})$, one obtains

$$f'(\bar{u})u + (\mu_b - \mu_a\,,\, K'(\bar{v})^{-1}u)_{L^2(\Omega)} = 0.$$

Moreover, we insert expression (11) for $K'(\bar{v})$ and arrive at

$$f'(\bar{u})u + \Big(\mu_b - \mu_a\,,\, \big(\lambda I + G'(\bar{u})\big)u\Big)_{L^2(\Omega)} = 0. \tag{23}$$

Lemma 2, equation (7), shows that the first derivative of f is given by

$$f'(\bar{u})u = (\kappa\,\bar{u} + q_1\,,\, u)_{L^2(\Omega)}, \tag{24}$$

where $q = q_1$ represents the solution of (8), with $y = \bar{y}$ in the right hand side. Due to Remark 3, we have $q_1 \in Y$ because of $\bar{y} \in Y \subset L^2(\Omega)$. For the second term in (23), we find

$$\Big(\mu_b - \mu_a\,,\, \big(\lambda I + G'(\bar{u})\big)u\Big)_{L^2(\Omega)} = \lambda(\mu_b - \mu_a\,,\, u)_{L^2(\Omega)} + (\mu_b - \mu_a\,,\, w)_{L^2(\Omega)}, \tag{25}$$

with $w = G'(\bar{u})u$, i.e., w is the solution of the linearized equation (4) with $y := \bar{y}$ and $h := u$. Arguing as in the proof of Lemma 2, we find

$$(\mu_b - \mu_a\,,\, w)_{L^2(\Omega)} = (q_2\,,\, u)_{L^2(\Omega)}, \tag{26}$$

where q_2 solves the adjoint equation

$$\begin{aligned} -\Delta\, q_2 + d'(\bar{y})\, q_2 &= \mu_b - \mu_a \quad \text{in } \Omega \\ \partial_\nu q_2 + q_2 &= 0 \qquad\quad \text{on } \Gamma. \end{aligned} \tag{27}$$

Again, this equation has the same structure as (4). From $(\mu_b - \mu_a) \in L^2(\Omega)$, we deduce $q_2 \in Y$ (cf. Remark 3). Inserting (26), (25) and (24) in (23) yields

$$(\kappa\,\bar{u} + q_1 + q_2 + \lambda(\mu_b - \mu_a)\,,\, u)_{L^2(\Omega)} = 0. \tag{28}$$

It is clear that $p = q_1 + q_2$ solves the adjoint equation (20). Therefore, since v and hence u are arbitrary, (28) is equivalent with (21). Moreover, (21) implies

$$\mu_b - \mu_a = -\frac{1}{\lambda}\,(\kappa\,\bar{u} + p) \tag{29}$$

with $p \in Y \subset C(\bar{\Omega})$ and $\bar{u} \in L^\infty(\Omega)$ due to Remark 2. Thus, since $\mu_a(x) \cdot \mu_b(x) = 0$ by definition (17), it follows that $\mu_a, \mu_b \in L^\infty(\Omega)$, because the right-hand side of (29) is bounded and measurable.

(ii) Complementary slackness conditions: The variational inequality (15) and equation (16) give

$$F'(\bar{v})(v - \bar{v}) = \int_\Omega (\mu_a - \mu_b)(v - \bar{v})\,dx \geq 0$$

for all $v \in V_{ad}$ and thus

$$(\mu_a - \mu_b\,,\,\bar{v})_{L^2(\Omega)} = \min_{v \in V_{ad}} (\mu_a - \mu_b\,,\,v)_{L^2(\Omega)} = (\mu_a\,,\,y_a)_{L^2(\Omega)} - (\mu_b\,,\,y_b)_{L^2(\Omega)},$$

since $\mu_a(x) \cdot \mu_b(x) = 0$ and $\mu_a(x), \mu_b(x) \geq 0$ by definition (17). Therefore, if $\mu_a(x) > 0$, we have $\bar{v}(x) = y_a(x)$, while $\mu_b(x) > 0$ implies $\bar{v}(x) = y_b(x)$. This immediately yields

$$(\mu_a\,,\,y_a - \bar{v})_{L^2(\Omega)} + (\mu_b\,,\,\bar{v} - y_b)_{L^2(\Omega)} = 0. \tag{30}$$

However, because of $\mu_a(x), \mu_b(x) \geq 0$ and $\bar{v} \in V_{ad}$, both addends on the right side of (30) are nonpositive and thus we arrive at

$$(\mu_a\,,\,y_a - \bar{v})_{L^2(\Omega)} = (\mu_b\,,\,\bar{v} - y_b)_{L^2(\Omega)} = 0.$$

Together with $\bar{v} = \lambda\,\bar{u} + G(\bar{u}) = \lambda\,\bar{u} + \bar{y}$, this implies (22). □

4 Second-Order Sufficient Conditions

As in case of first-oder conditions in Section 3, the proof of second-order sufficient conditions for (P) is based on the results for the auxiliary problem (PV), which is an optimization problem with simple box-constraints. For problems of such type, the theory of second-order conditions is well-known. To formulate these conditions for (PV), we introduce the *strongly active set* as follows:

Definition 1. *Let $\tau > 0$ be given. Then the strongly active set A_τ is defined by $A_\tau := \{x \in \Omega \,|\, \mu_a(x) + \mu_b(x) \geq \tau\}$.*

Notice that, according to (17), μ_a and μ_b cannot be jointly positive. Moreover, the corresponding τ-critical cone with respect to v is defined in a standard way by

$$\hat{C}_\tau := \left\{ v \in L^2(\Omega) \;\middle|\; \begin{array}{l} v(x) = 0 \;,\ \text{a.e. in } A_\tau \\ v(x) \geq 0 \;,\ \text{where } \bar{v}(x) = y_a(x) \text{ and } x \notin A_\tau \\ v(x) \leq 0 \;,\ \text{where } \bar{v}(x) = y_b(x) \text{ and } x \notin A_\tau \end{array} \right\}, \tag{31}$$

with $\bar{v} = \lambda\,\bar{u} + \bar{y}$ as defined above. With these definitions at hand, one can prove by standard arguments the following theorem covering the local optimality of $\bar{v}$, cf. eg. [5].

Theorem 4. *Suppose that $\bar{v}$ is feasible for* (PV) *and satisfies the variational inequality* (15). *Assume further that the coercivity condition*

$$F''(\bar{v})v^2 \geq \tilde{\delta}\,\|v\|_{L^2(\Omega)} \quad \forall\, v \in \hat{C}_\tau \tag{32}$$

is satisfied with some $\tilde{\delta} > 0$. Then there exist $\tilde{\varepsilon} > 0$ and $\tilde{\sigma} > 0$ such that

$$F(v) \geq F(\bar{v}) + \tilde{\sigma}\,\|v - \bar{v}\|^2_{L^2(\Omega)} \tag{33}$$

for all $v \in V_{ad}$ with $\|v - \bar{v}\|_{L^\infty(\Omega)} \leq \tilde{\varepsilon}$.

Due to (33), (15) and (32) yield local optimality of $\bar{v}$ for (PV) and hence, (32) is a second-order sufficient optimality condition. It remains to transfer this condition to the original terms y and u. For this reason, we need the following lemma on $F''(\bar{v})$.

Lemma 5. *Assume that* (R) *is fulfilled. Then F is twice continuously Fréchet differentiable at $\bar{v}$ and its second derivative is given by*

$$F''(\bar{v})v^2 = \mathcal{L}''_{(y,u)}(\bar{y}, \bar{u}, p, \mu)(y, u)^2. \tag{34}$$

Proof. Thanks to $F(v) = f(K(v))$ and (R), it is clear that F is twice continuously Fréchet differentiable in a neighborhood of $\bar{v}$. The chain rule implies

$$F''(v)[v_1, v_2] = f''(K(v))[K'(v)v_1, K'(v)v_2] + f'(K(v))K''(v)[v_1, v_2]. \tag{35}$$

We substitue $v = \bar{v}$ and thus $K(\bar{v}) = \bar{u}$. Moreover, we set $v_1 = v_2 = v$, and $K'(v)v_1 = K'(v)v_2 = K'(\bar{v})v = u$. Hence, (35) is equivalent to

$$F''(\bar{v})v^2 = f''(\bar{u})u^2 + f'(\bar{u})K''(\bar{v})v^2.$$

In view of (23), we have for the second addend

$$f'(\bar{u})K''(\bar{v})v^2 = -\Big(\mu_b - \mu_a\,,\ \big(\lambda\, I + G'(\bar{u})\big)K''(\bar{v})v^2\Big)_{L^2(\Omega)}.$$

Together with the expression for $K''(\bar{v})$ in (12), we arrive at

$$\begin{aligned} F''(\bar{v})v^2 &= f''(\bar{u})u^2 + \big(\mu_b - \mu_a\,,\, G''(\bar{u})[K'(\bar{v})v, K'(\bar{v})v]\big)_{L^2(\Omega)} \\ &= f''(\bar{u})u^2 + \big(\mu_b - \mu_a\,,\, G''(\bar{u})u^2\big)_{L^2(\Omega)}\,. \end{aligned} \tag{36}$$

Since $z = G''(\bar{u})u^2$ solves equation (5), similar arguments as in the proof of Lemma 2 give

$$(\mu_b - \mu_a\,,\, z)_{L^2(\Omega)} = -(d''(\bar{y})y^2\,,\, q_2)_{L^2(\Omega)},$$

where q_2 is the solution of (27) and $y = G'(\bar{u})u$, i.e. y represents the solution of the linearized equation (4). Thus, together with (9) for the second derivative of f (see Lemma 2), (36) is transformed into

$$F''(\bar{v})v^2 = \|y\|^2_{L^2(\Omega)} + \kappa\,\|u\|^2_{L^2(\Omega)} - \int_\Omega d''(\bar{y})\,y^2\,(q_1 + q_2)\,dx,$$

where q_1 again denotes the solution of (8) with $y = \bar{y}$ in the right side. As in the proof of Theorem 3, we have $p = q_1 + q_2$ and hence we obtain

$$\begin{aligned} F''(\bar{v})v^2 &= \|y\|^2_{L^2(\Omega)} + \kappa\,\|u\|^2_{L^2(\Omega)} - \int_\Omega d''(\bar{y})\,y^2\,p\,dx \\ &= J''_{(y,u)}(\bar{y},\bar{u})(y,u)^2 - \int_\Omega d''(\bar{y})\,y^2\,p\,dx = \mathcal{L}''_{(y,u)}(\bar{y},\bar{u},p,\mu)(y,u)^2, \end{aligned}$$

according to the definition of $\mathcal{L}$ in (19). □

Based on (31), we define the τ-critical cone for the original problem (P), denoted by C_τ as follows:

Definition 2. *(Critical cone) Let $\hat{C}_\tau$ be defined as in* (31)*. The critical cone associated to* (P) *is given by*

$$C_\tau := \{(y,u) \in Y \times L^2(\Omega) \mid y = G'(\bar{u})u \;\; \textit{and} \;\; \lambda\,u + y \in \hat{C}_\tau\}.$$

Now, we are able to state second-order sufficient conditions for (P).

$$\text{(SSC)} \quad \left\{ \begin{array}{l} \textit{Let } \delta > 0 \textit{ exist such that} \\ \qquad \mathcal{L}''(\bar{y},\bar{u},p,\omega)(y,u)^2 \geq \delta\,\|u\|^2_{L^2(\Omega)} \quad \textit{for all } (y,u) \in C_\tau. \end{array} \right.$$

We show that (SSC) is indeed sufficient for local optimality of $\bar{u}$.

Theorem 5. *Let $(\bar{y},\bar{u})$ satisfy the first-order necessary optimality conditions for Problem* (P) *and assume that condition* (SSC) *is fulfilled with some $\delta > 0$, $\tau > 0$. Then there exist $\varepsilon > 0$ and $\sigma > 0$ such that*

$$J(y,u) \geq J(\bar{y},\bar{u}) + \sigma\,\|u - \bar{u}\|^2_{L^2(\Omega)} \tag{37}$$

for all $(y,u) \in Y \times L^2(\Omega)$ with $y = G(u)$, $y_a(x) \leq \lambda\,u(x) + y(x) \leq y_b(x)$, and $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$.

Proof. First, we choose an arbitrary pair $(\eta, h) \in C_\tau$ and define $v := \lambda h + \eta$. Notice that $\eta = G'(\bar{u})h$ holds according to the definition of C_τ. Due to Lemma 5, one obtains

$$F''(\bar{v})v^2 = L''_{(y,u)}(\bar{y}, \bar{u}, p, \mu)(\eta, h)^2 \geq \delta \|h\|^2_{L^2(\Omega)}, \tag{38}$$

where we used condition (SSC) for the last estimate. Due to $h = (\lambda I + G'(\bar{u}))^{-1}v$, (38) is equivalent to

$$\begin{aligned} F''(\bar{v})v^2 &\geq \delta \|(\lambda I + G'(\bar{u}))^{-1}v\|^2_{L^2(\Omega)} \\ &\geq \delta \left(\frac{1}{\|\lambda I + G'(\bar{u})\|_{\mathcal{L}(L^2(\Omega))}} \|v\|_{L^2(\Omega)} \right)^2 \\ &\geq \delta \|\lambda I + G'(\bar{u})\|^{-2}_{\mathcal{L}(L^2(\Omega))} \|v\|^2_{L^2(\Omega)} \\ &= \tilde{\delta} \|v\|^2_{L^2(\Omega)}, \end{aligned} \tag{39}$$

with $\tilde{\delta} > 0$. Because of $(\eta, h) \in C_\tau$, clearly $v \in \hat{C}_\tau$ holds true. Moreover, thanks to (R), every $v \in \hat{C}_\tau$ can be expressed by some $(\eta, h) \in C_\tau$, and hence (39) holds true for all $v \in \hat{C}_\tau$. In this way, F'' satisfies a coercivity condition and thus, Theorem 4 yields

$$F(v) \geq F(\bar{v}) + \tilde{\sigma} \|v - \bar{v}\|^2_{L^2(\Omega)} \tag{40}$$

for all $v \in V_{ad}$ with $\|v - \bar{v}\|_{L^\infty(\Omega)} \leq \tilde{\varepsilon}$. In particular, we may take

$$v = \lambda u + G(u),$$

where u is taken arbitrary with $y_a(x) \leq \lambda u(x) + G(u)(x) \leq y_b(x)$ and $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$ such that $\|v - \bar{v}\|_{L^\infty(\Omega)} \leq \tilde{\varepsilon}$ and $\|v - \bar{v}\|_{L^2(\Omega)} \leq \rho_1$. Notice that, because of (R), to every $v \in V_{ad}$ with $\|v - \bar{v}\|_{L^2(\Omega)} \leq \rho_1$ a function u exists with $u = K(v)$ and $\|u - \bar{u}\|_{L^2(\Omega)} \leq r_1$. On the other hand, the continuity of the mapping $\lambda I + G$ from $L^\infty(\Omega)$ to $L^\infty(\Omega)$ ensures that $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$ implies $\|v - \bar{v}\|_{L^\infty} \leq r$. If we take ε sufficiently small, then it follows that $r \leq \tilde{\varepsilon}$ and $\|v - \bar{v}\|_{L^2(\Omega)} \leq c\|v - \bar{v}\|_{L^\infty(\Omega)} \leq \rho_1$. Hence, for all u with $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$, there exists a v with $\lambda u + G(u) = v$ and with $\|v - \bar{v}\|_{L^\infty(\Omega)} \leq \tilde{\varepsilon}$. Then, with $F(v) = f(u)$ and $F(\bar{v}) = f(\bar{u})$, (40) gives

$$f(u) \geq f(\bar{u}) + \tilde{\sigma} \|\lambda u + G(u) - (\lambda \bar{u} + G(\bar{u}))\|^2_{L^2(\Omega)} \tag{41}$$

for all u with $\lambda u + G(u) \in V_{ad}$ and $\|u - \bar{u}\|_{L^\infty(\Omega)} \leq \varepsilon$. This already implies the local optimality of $\bar{u}$. It remains to show the quadratic growth condition (37). A Taylor expansion for the last term in (41) yields

$$\lambda u + G(u) - (\lambda \bar{u} + G(\bar{u})) = \lambda(u - \bar{u}) + G'(\bar{u})(u - \bar{u}) + r_1^G(\bar{u}, u - \bar{u}),$$

and, since G is continuously Fréchet differentiable from $L^2(\Omega)$ to Y (see Lemma 1), the remainder term satisfies

$$\frac{\|r_1^G\|_{L^2(\Omega)}}{\|u-\bar{u}\|_{L^2(\Omega)}} \to 0, \quad \text{as } \|u-\bar{u}\|_{L^2(\Omega)} \to 0. \tag{42}$$

Therefore, we obtain

$$\begin{aligned}
&\|\lambda u + G(u) - (\lambda \bar{u} + G(\bar{u}))\|_{L^2(\Omega)} \\
&\quad = \|(\lambda I + G'(\bar{u}))(u-\bar{u}) + r_1^G\|_{L^2(\Omega)} \\
&\quad \geq \|(\lambda I + G'(\bar{u}))(u-\bar{u})\|_{L^2(\Omega)} - \|r_1^G\|_{L^2(\Omega)} \\
&\quad \geq \left(\frac{1}{\|(\lambda I + G'(\bar{u}))^{-1}\|_{\mathcal{L}(L^2(\Omega))}} - \frac{\|r_1^G\|_{L^2(\Omega)}}{\|u-\bar{u}\|_{L^2(\Omega)}} \right) \|u-\bar{u}\|_{L^2(\Omega)} \\
&\quad \geq \tilde{c}\, \|u-\bar{u}\|_{L^2(\Omega)}.
\end{aligned}$$

Since $(\lambda I + G'(\bar{u}))$ is continuously invertible because of (R), (42) yields $\tilde{c} > 0$ if $\|u-\bar{u}\|_{L^2(\Omega)}$ is sufficiently small. Thus (41) implies

$$f(u) \geq f(\bar{u}) + \tilde{\sigma}\, \tilde{c}^2\, \|u-\bar{u}\|^2_{L^2(\Omega)} = f(\bar{u}) + \sigma \|u-\bar{u}\|^2_{L^2(\Omega)}. \qquad \square$$

Remark 4. Clearly, due to (37), $\bar{u}$ is a strict optimal solution.

5 Numerical Tests

For our numerical tests, we consider an optimal control problem that differs slightly from (P), as already mentioned in Remark 1. Instead of (1), the state equation is now given by

$$\begin{aligned}
-\Delta y(x) + y(x) + d(y(x)) &= u(x) \quad \text{in } \Omega \\
\partial_\nu y(x) &= 0 \quad \text{on } \Gamma.
\end{aligned} \tag{43}$$

One can easily verify that the theory presented above is also valid with the new state equation (43).

We investigated two examples with different nonlinearities $d(y)$. In both cases, the desired state was given by

$$y_d(x_1, x_2) = 8\, \sin(\pi\, x_1)\, \sin(\pi\, x_2) - 4$$

and the bounds were fixed at $y_a(x_1, x_2) \equiv -1$ and $y_b(x_1, x_2) \equiv 1$. The Tikhonov regularization parameter was set to $\kappa = 0.5 \cdot 10^{-5}$. Moreover, to approximate a purely state constrained problem, we fixed $\lambda = 0.5 \cdot 10^{-5}$. In the first example, the nonlinearity was defined by

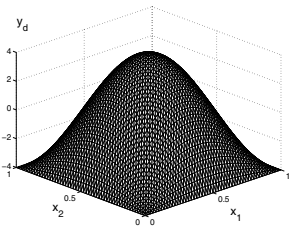


Fig. 1. Desired state y_d.

$$d(y) = y^3, \tag{44}$$

whereas we took

$$d(y) = e^{5\,y} \tag{45}$$

in the second one. Thus, the assumptions on d mentioned in Section 1, are fulfilled in both cases.

The optimization problems were solved numerically by a SQP method that is described in detail for instance in [8] or [9]. To solve the arising linear quadratic problems, a primal-dual active set strategy was applied, see for instance [1] or [3]. We used a conforming finite element method with linear ansatz functions to solve the state equation and the adjoint equation. For all computations, uniform meshes were used. The number of intervals in one dimension, denoted by N, is related to the mesh-size, i.e. the diameter of the triangles, by $h = \sqrt{2}\,\mathrm{N}^{-1}$. The following figures show the numerical solution for the first example. This computation was performed with a mesh size N=50. Here and in the following, the numerical solutions are denoted by the subscript h.

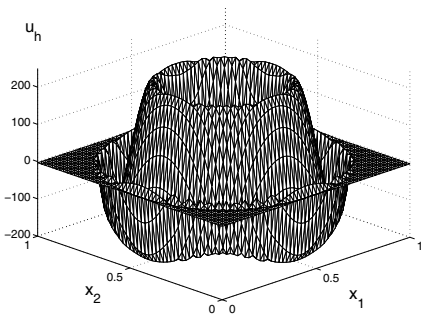


Fig. 2. Control u_h in the first example.

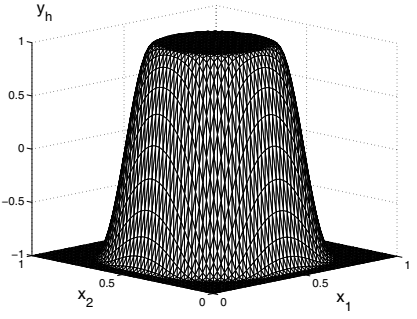


Fig. 3. State y_h in the first example.

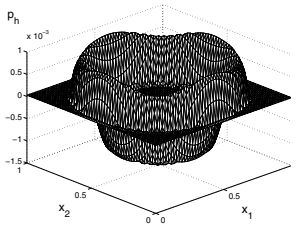


Fig. 4. Adjoint state p_h.

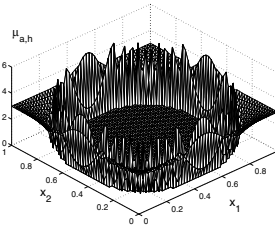


Fig. 5. Lagrange multiplier $\mu_{a,h}$.

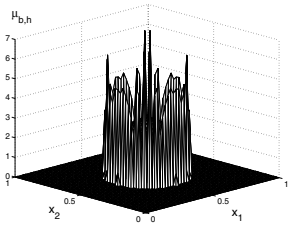


Fig. 6. Lagrange multiplier $\mu_{b,h}$.

As one can see in the Figures 5 and 6, the Lagrange multipliers tend to be irregular on the boundaries of the active sets. This might indicate that the Lagrange multipliers associated with the state constraints for $\lambda = 0$ should be measures. This verifies the known theory, see for instance Casas [4] or Bergounioux and Kunisch [2]. However, in view of (21) with $p = G'(\bar{u})^*(G(\bar{u}) - y_d + \mu)$, the equation for $\mu = \mu_a - \mu_b$ is given by

$$\lambda\,\mu + G'(\bar{u})^*\mu = G'(\bar{u})^*(y_d - G(\bar{u})) - \kappa\,\bar{u}$$

with a compact operator $G'(\bar{u})^* : L^2(\Omega) \to L^2(\Omega)$. This equation is ill-posed for $\lambda = 0$. Therefore, as $\lambda = 0.5 \cdot 10^{-5}$ is chosen quite small, we are faced with the characteristic difficulties of ill-posed problems. In view of this, the computed Lagrange multipliers are certainly overlaid by rounding errors that are difficult to quantify.

To describe the convergence behaviour of the algorithm, the values of the discrete objective functional $J_h = 1/2\,\|y_h - y_d\|^2_{L^2(\Omega)} + \kappa/2\,\|u_h\|^2_{L^2(\Omega)}$ are displayed in the following Tables 1–3 for each step of the SQP-iteration, denoted by $\mathrm{it}_{\mathrm{SQP}}$. As a further convergence indicator, the error in the semilinear state equation is approximated by

$$e_y = \frac{\|G_h^{-1}(y_h) - u_h\|_{L^2(\Omega)}}{\|y_h\|_{L^2(\Omega)}},$$

where G_h denotes the discrete control-to-state operator $G_h : u_h \mapsto y_h$. Thus, e_y quantifies the relative error of the discrete analogon of $-\Delta y + c\,y + d(y) - u$, i.e. the error in the semilinear state equation. Similarly the error in the adjoint equation is measured by

$$e_p = \frac{\|\left(G_h'(y_h)^{-1}\right)^* p_h - (y_h - y_d + \mu_{b,h} - \mu_{a,h})\|_{L^2(\Omega)}}{\|p_h\|_{L^2(\Omega)}},$$

where $\left(G_h'(y_h)^{-1}\right)^*$ is associated with $-\Delta p + c\,p + d'(y)\,p$. Furthermore, the error in the necessary condition (21) is approximated by

$$e_{opt} = \|\kappa\,u_h + p_h + \lambda\,(\mu_{b,h} - \mu_{a,h})\|_{L^2(\Omega)}.$$

The difference between two consecutive iterates, quantified by

$$\delta = \frac{1}{3}\left(\frac{\|u_h^{(n)} - u_h^{(n+1)}\|_{L^2(\Omega)}}{\|u_h^{(n+1)}\|_{L^2(\Omega)}} + \frac{\|y_h^{(n)} - y_h^{(n+1)}\|_{L^2(\Omega)}}{\|y_h^{(n+1)}\|_{L^2(\Omega)}} + \frac{\|p_h^{(n)} - p_h^{(n+1)}\|_{L^2(\Omega)}}{\|p_h^{(n+1)}\|_{L^2(\Omega)}}\right)$$

was used for the termination condition of the SQP method. More precisely, the iteration stopped if $\delta < 10^{-2}$. The following table shows the convergence behavior in the first example for a mesh size of N=50. In addition to the values of J_h and the error approximations described above, the number of active set iterations denoted by $\mathrm{it}_{\mathrm{AS}}$ is shown in the last column.

Table 1. Example 1 with N=50

it_{SQP}	J_h	e_{opt}	e_y	e_p	δ	$\#\text{it}_{\text{AS}}$
0	3.1099e+00	1.0000e+00	3.5361e-03	9.1101e-04	-	
1	1.3793e+00	4.1930e-20	3.4860e-04	3.7443e-01	5.7686e+02	13
2	1.3757e+00	3.5225e-20	3.7393e-04	4.4817e-05	2.5864e-01	6
3	1.3757e+00	3.3836e-20	3.6347e-04	2.1590e-11	3.3737e-04	1

We observe that e_p is much smaller than e_y. A possible explanation for this fact could be that the adjoint equation represents a linear PDE in contrast to the semilinear state equation.

Table 2 illustrates the convergence behaviour in the first example for N=100. As one can see, the error in the approximation of the PDEs is reduced significantly. However, the value of the discrete objective functional is not decreased noticeably.

Table 2. Example 1 with N=100

it_{SQP}	J_h	e_{opt}	e_y	e_p	δ	$\#\text{it}_{\text{AS}}$
0	3.1112e+00	1.0000e+00	8.9151e-04	2.3143e-04	-	-
1	1.3800e+00	4.0038e-20	8.8727e-05	9.3948e-02	5.6869e+02	23
2	1.3757e+00	3.3583e-20	9.5252e-05	1.2688e-05	2.6991e-01	8
3	1.3757e+00	3.3876e-20	9.2619e-05	6.4219e-12	3.3493e-04	1

Figures 7–11 show the numerical solution of the second example for N=50. Again, the Lagrange multipliers are comparatively irregular on the borders of the active sets.

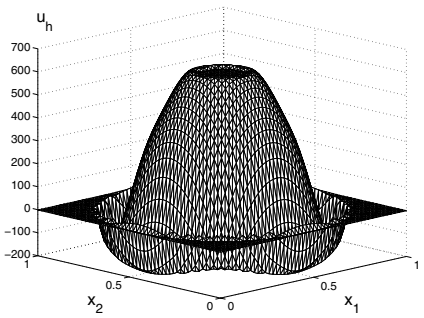


Fig. 7. Control u_h in the second example.

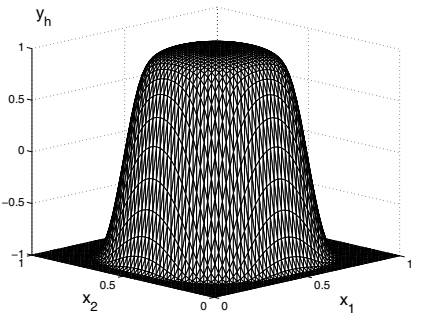


Fig. 8. State y_h in the second example.

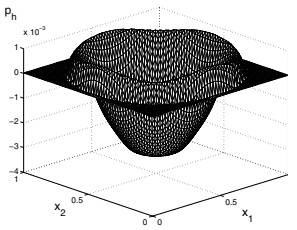


Fig. 9. Adjoint state p_h.

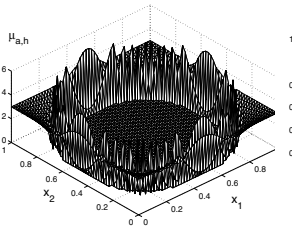


Fig. 10. Lagrange multiplier $\mu_{a,h}$.

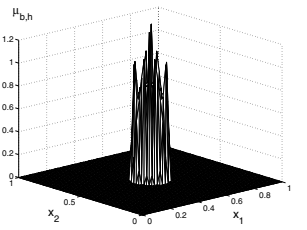


Fig. 11. Lagrange multiplier $\mu_{b,h}$.

The convergence behavior of the algorithm in this example is illustrated in Table 3. The nonlinearity $d(y) = e^{5\,y}$ of this example is much steeper than $d(y) = y^3$. Therefore, the number of SQP-iterations is larger than for $d(y) = y^3$.

Table 3. Example 2 with N=50

it_{SQP}	J_h	e_{opt}	e_y	e_p	δ	$\#\text{it}_{\text{AS}}$
0	3.1099e+00	1.0000e+00	2.9450e-03	3.4595e-03	-	-
1	3.2742e+00	1.9729e-20	1.9334e-02	3.7889e+00	1.2591e+03	1
2	1.3780e+00	3.6935e-20	2.9660e-02	9.8747e-02	1.1220e+00	13
3	1.5610e+00	8.9187e-20	8.8222e-02	8.0814e-02	5.8821e-01	14
4	1.4711e+00	7.3050e-20	5.4490e-02	6.1055e-03	2.0554e-01	10
5	1.5523e+00	8.6599e-20	8.5751e-02	1.2353e-02	1.8589e-01	10
6	1.5102e+00	8.5926e-20	6.9141e-02	1.2245e-03	8.9151e-02	7
7	1.5449e+00	8.2400e-20	8.2864e-02	1.7494e-03	7.2685e-02	8
8	1.5203e+00	8.3514e-20	7.2910e-02	6.1815e-04	5.2567e-02	5
9	1.5392e+00	8.9509e-20	8.0637e-02	4.7578e-04	3.9750e-02	5
10	1.5248e+00	8.7443e-20	7.4701e-02	2.4842e-04	3.1121e-02	5
11	1.5357e+00	9.0720e-20	7.9241e-02	1.6924e-04	2.3635e-02	4
12	1.5275e+00	8.2449e-20	7.5798e-02	9.1957e-05	1.8277e-02	5
13	1.5337e+00	8.5332e-20	7.8403e-02	5.9436e-05	1.3928e-02	3
14	1.5291e+00	8.6704e-20	7.6457e-02	3.2479e-05	1.0600e-02	3
15	1.5325e+00	8.4110e-20	7.7909e-02	1.9987e-05	8.0250e-03	3

References

1. M. Bergounioux, K. Ito, and K. Kunisch. Primal-dual strategy for constrained optimal control problems. SIAM J. Control Optim., 37:1176–1194, 1999.
2. M. Bergounioux and K. Kunisch. On the structure of the Lagrange multiplier for state-constrained optimal control problems. Systems and Control Letters, 48:16–176, 2002.

3. M. Bergounioux and K. Kunisch. Primal-dual active set strategy for state-constrained optimal control problems. Computational Optimization and Applications, 22:193–224, 2002.
4. E. Casas. Control of an elliptic problem with pointwise state constraints. SIAM J. Control Optim., 4:1309–1322, 1986.
5. E. Casas and M. Mateos. Second order sufficient optimality conditions for semilinear elliptic control problems with finitely many state constraints. SIAM J. Control Optim., 40:1431–1454, 2002.
6. E. Casas and F. Tröltzsch. Second order necessary and sufficient optimality conditions for optimization problems and applications to control theory. SIAM J. Control Optim., 13:406–431, 2002.
7. E. Casas, F. Tröltzsch, and A. Unger. Second order sufficient optimality conditions for some state-constrained control problems of semilinear elliptic equations. SIAM J. Control and Optimization, 38(5):1369–1391, 2000.
8. K. Ito and K. Kunisch. Augmented Lagrangian-sqp-methods for nonlinear optimal control problems of tracking type. SIAM J. Control Optim., 34(3):874–891, 1996.
9. F. Tröltzsch. An SQP method for the optimal control of a nonlinear heat equation. Control and Cybernetics, 23(1/2):267–288, 1994.
10. F. Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control–state constraints. SIAM J. Optimization, to appear.

On Abstract Control Problems with Non-Smooth Data*

Zsolt Páles

Institute of Mathematics, University of Debrecen, 4029 Debrecen, Pf. 12, Hungary.
pales@math.klte.hu **

Summary. The aim of this paper is to extend and generalize the known necessary optimality conditions to non-smooth as well as higher-order setting concerning the optimization problem (which is called an abstract control problem)

$$\begin{cases} \text{Minimize } f_0(x,u) \text{ with respect to} \\ f_1(x,u) \le 0, \dots, f_k(x,u) \le 0, F(x,u) = 0, \ (x,u) \in D \times U, \end{cases}$$

where D is an open set of a Banach space X, U is a nonempty set, and the data fulfill a certain convexity condition which can often be verified in the context of ordinary optimal control problems.

1 Introduction

We consider abstract control problems of the form

$$\begin{cases} \text{Minimize } f_0(x,u) \text{ with respect to} \\ f_1(x,u) \le 0, \dots, f_k(x,u) \le 0, F(x,u) = 0, \ (x,u) \in D \times U, \end{cases} \qquad (\mathcal{CP})$$

where D is an open set of a Banach space X, U is a nonempty set (called *control set*), $f_0, \dots, f_k : D \to \mathbb{R}$, and F is a mapping of the product set $D \times U$ into another Banach space Y (the equation $F(x,u) = 0$ is called *control system*). Optimization problems of this form are also called *mixed problems* and were considered for the first time by Pshenichnyi [17] and Pshenichnyi and Nenakhov [13, 18]. These problems were also dealt with in the first chapter of the book by Ioffe and Tihomirov [10] where also further references can be found. Due to the hidden convexity properties of optimal control problems, the results concerning abstract control problems can successfully be applied to develop the well-known necessary conditions for optimality, such as the

* Dedicated to the 70th birthday of Professor V. M. Tihomirov

** The research was supported by the Hungarian Research Fund (OTKA) Grant T038072.

Pontryagin maximum principle. The detailed description of this approach can be found also in [10].

A pair (x, u) is called *admissible (feasible) for* $(\mathcal{CP})$ if

$$(x, u) \in D \times U, \ f_1(x, u) \leq 0, \ \ldots, \ f_k(x, u) \leq 0, \text{ and } F(x, u) = 0.$$

An admissible pair (x_*, u_*) is said to be a *(strong) local minimum of the problem* $(\mathcal{CP})$ if there exists a neighborhood $V \subset D$ of x_* such that

$$f_0(x, u) \geq f_0(x_*, u_*) \quad \text{for all admissible pairs } (x, u) \text{ with } x \in V.$$

Assuming certain smoothness and convexity assumptions, first-order necessary conditions can be derived for the optimality of a pair (x_*, u_*). We restate here Theorem 1.1.3 of the book [10] for the reader's convenience. Recall that the Lagrangian function $\mathcal{L} : D \times U \times \mathbb{R}^{k+1} \times Y^* \to \mathbb{R}$ of problem $(\mathcal{CP})$ is defined by

$$\mathcal{L}(x, u, \lambda_0, \ldots, \lambda_k, y^*) := \sum_{i=0}^{k} \lambda_i f_i(x, u) + (y^* \circ F)(x, u). \tag{1}$$

As usual, Y^* stands for the topological dual space of Y.

Theorem 1. *Let* $f_0, f_1, \ldots, f_k : D \times U \to \mathbb{R}$, $F : D \times U \to Y$, *and* $(x_*, u_*) \in D \times U$. *Assume that*

(i) For every $u \in U$, *the mapping* $x \mapsto \big(f_0(x, u), \ldots, f_k(x, u), F(x, u)\big)$ *is continuously Fréchet differentiable at the point* x_*;

(ii) The range of the continuous linear operator $F_x(x_*, u_*) : X \to Y$ *has finite codimension in* Y;

(iii) For every $x \in D$, *the mapping* $u \mapsto \big(f_0(x, u), \ldots, f_k(x, u), F(x, u)\big)$ *satisfies the following convexity assumption: for every* $u_1, u_2 \in U$ *and* $\alpha \in [0, 1]$, *there exists an element* $u \in U$ *such that*

$$f_i(x, u) \leq \alpha f_i(x, u_1) + (1 - \alpha) f_i(x, u_2) \qquad \big(i \in \{0, \ldots, k\}\big),$$
$$F(x, u) = \alpha F(x, u_1) + (1 - \alpha) F(x, u_2).$$

If $(x_*, u_*) \in D \times U$ *is a local minimum of the problem* $(\mathcal{CP})$, *then there exist Lagrange multipliers* $\lambda_0, \ldots, \lambda_k \geq 0$ *and* $y^* \in Y^*$, *not all zero, such that*

$$\lambda_1 f_1(x_*, u_*) = 0, \quad \ldots, \quad \lambda_k f_k(x_*, u_*) = 0, \tag{2}$$

$$\lambda_0 f_{0x}(x_*, u_*) + \cdots + \lambda_k f_{kx}(x_*, u_*) + y^* \circ [F_x(x_*, u_*)] = 0, \tag{3}$$

and the minimum principle

$$\min_{u \in U} \mathcal{L}(x_*, u, \lambda_0, \ldots, \lambda_k, y^*) = \mathcal{L}(x_*, u_*, \lambda_0, \ldots, \lambda_k, y^*) \tag{4}$$

holds.

Here and in the sequel, the subscript x denotes derivative/differentiation with respect to the x variable. Obviously, (3) can also be written as

$$\mathcal{L}_x(x_*, u_*, \lambda_0, \ldots, \lambda_k, y^*) = 0.$$

Adopting the setting of Theorem 1, the idea behind our approach is as follows. First, for all fixed $m \in \mathbb{N}$ and $\mathbf{u} = (u_1, \ldots, u_m) \in U^m$, a new problem (denoted by $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$ below) is constructed which will be an ordinary programming problem. Due to the convexity assumption (iii) of Theorem 1, it will turn out that the local optimality of the point (x_*, u_*) in $(\mathcal{CP})$ yields the local optimality of $(x_*, 0, \ldots, 0) \in D \times \mathbb{R}^m$ in $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$. Applying the standard Lagrange principle to the problem $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$, the existence of multipliers $\lambda_0, \ldots, \lambda_k \geq 0$ and $y^* \in Y^*$, not all zero, can be obtained that satisfy (2), (3) and

$$\min_{u \in \{u_1, \ldots, u_m\}} \mathcal{L}(x_*, u, \lambda_0, \ldots, \lambda_k, y^*) = \mathcal{L}(x_*, u_*, \lambda_0, \ldots, \lambda_k, y^*) \qquad (5)$$

In the final step, applying a compactness argument and using the finite codimensionality of the range of the Fréchet derivative $F_x(x_*, u_*)$, one can derive the existence of multipliers that also satisfy (4) instead of (5).

In our main result, involving the notion of strict prederivative introduced by Ioffe in [8] for F, Clarke's subgradient for $f_0, \ldots, f_k$, and applying the nonsmooth implicit function theorem of [14] and the Lagrange principle developed in [15], we generalize the above result to the case of not necessarily smooth functions $f_0, \ldots, f_k$ and F and we obtain first as well as higher-order necessary conditions for optimality.

2 Strict Prederivatives and a Lagrange Principle

In what follows, $L(X, Y)$ denotes the space of continuous linear operators from X to Y. This space is equipped with its usual operator norm. For the analysis of locally Lipschitz operators, Ioffe [8] introduced the following notion:

A set $\mathcal{A} \subset L(X, Y)$ is called a *strict (Fréchet) prederivative* for $F : D \to Y$ at x_* if, for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$F(x_1) - F(x_2) \in \mathcal{A}(x_1 - x_2) + \varepsilon \|x_1 - x_2\| B_Y \qquad (x_1, x_2 \in \delta B_X + x_*).$$

Here B_X and B_Y denote the closed unit balls of X and Y, respectively, and $\mathcal{A}(x) := \{Ax : A \in \mathcal{A}\}$. Later we shall also use the notation $y^* \circ \mathcal{A} := \{y^* \circ A : A \in \mathcal{A}\}$.

Generally, (and unfortunately) there is no natural candidate for a strict prederivative. However, it can be proved that F is strictly Fréchet differentiable at x_* if and only if $\mathcal{A} = \{F'(x_*)\}$ is a strict prederivative of F at x_*. In the case when X and Y are finite dimensional spaces, then *Clarke's generalized Jacobian* $\partial F(x_*)$ ([4]) is a strict prederivative of F at x_* ([8]). If F is locally Lipschitz at x_*, then one can easily check that the set of all continuous

linear operators with norm less than or equal to L (where L is the Lipschitz modulus of F around x_*) is a strict prederivative of F at x_*. Conversely, the existence of a (norm-)bounded strict prederivative obviously implies that the function F is locally Lipschitz at x_*.

A sequence $\xi = (x_n, t_n)_{n\in\mathbb{N}}$ in $X \times \mathbb{R}_+$ such that $x_n \to x_*$, $t_n \to 0+$, will be termed an *approximation of* x_* in the sequel. Given an approximation $\xi = (x_n, t_n)_{n\in\mathbb{N}}$ of x_*, a normed-space valued function φ defined on D is said to be *differentiable with respect to* ξ if the limit

$$D_\xi\varphi(x_*) := \lim_{n\to\infty} \frac{\varphi(x_n) - \varphi(x_*)}{t_n}$$

exists, which is called the *derivative of* φ *with respect to* ξ. In order to derive the the standard first-order and second-order necessary conditions from our results below, one has to take approximations of the form $\xi = (x_*, t_n)_{n\in\mathbb{N}}$ and $\xi = \left(x_* + \sqrt{t_n}d, t_n\right)_{n\in\mathbb{N}}$, respectively. For approximations of the form $\xi = (x_*, t_n)_{n\in\mathbb{N}}$, one obviously has the differentiability of any φ with respect to ξ and $D_\xi\varphi(x_*) = 0$, hence all the conditions and statements involving such a particular ξ are trivially satisfied in the sequel. For approximations of the form $\xi = \left(x_* + \sqrt{t_n}d, t_n\right)_{n\in\mathbb{N}}$, the differentiability of φ with respect to ξ holds if φ is first-order smooth, $\varphi'(x^*)d = 0$ and the second-order directional derivative of φ in direction d exists. Thus, in this setting, all conditions and statements involving ξ reduce to second-order regularity and necessary conditions. More generally, the use of approximations of the form $\xi = \left(x_* + \sqrt[k]{t_n}d_1 + \cdots + \sqrt[k]{t_n^{k-1}}d_{k-1}, t_n\right)_{n\in\mathbb{N}}$, where $d_1, \ldots, d_{k-1} \in X$ are fixed vectors, leads to necessary conditions of order k.

In relation to the notions of strict prederivative and derivative with respect to approximations and also applying the approach of Dubovitskii and Milyutin [5, 6, 7], the following nonsmooth Lagrange Principle has been stated in [15]:

Theorem 2. *Let* $f_0, f_1, \ldots, f_k : D \to \mathbb{R}$, $F : D \to Y$, *and* $x_* \in D$. *Assume that*

(i) $f_0, f_1, \ldots, f_k$ *are locally Lipschitz at* x_*;
(ii) There exists a compact convex set $\mathcal{A} \subset L(X, Y)$ *which is a strict prederivative of* F *at* x_*;
(iii) Each $A \in \mathcal{A}$ *has a closed range.*

If $x_* \in D$ *is a local minimum of the problem*

$$\begin{cases} \text{Minimize } f_0(x) \text{ with respect to} \\ f_1(x), \ldots, f_k(x) \le 0,\ F(x) = 0,\ x \in D \end{cases} \tag{$\mathcal{P}$}$$

Then, for all approximations $\xi = (x_n, t_n)_{n\in\mathbb{N}}$ *of* x_* *such that* $f_0, \ldots, f_k, F$ *are differentiable with respect to* ξ, *there are multipliers* $\lambda_0, \lambda_1, \ldots, \lambda_k \ge 0$ *and* $y^* \in Y^*$, *not all zero, such that*

$$\lambda_1 f_1(x_*) = 0, \quad \ldots, \quad \lambda_k f_k(x_*) = 0, \tag{6}$$

$$0 \in \lambda_0 \partial f_0(x_*) + \cdots + \lambda_k \partial f_k(x_*) + y^* \circ \mathcal{A}, \tag{7}$$

$$\lambda_0 D_\xi f_0(x_*) + \cdots + \lambda_k D_\xi f_k(x_*) + y^*\big(D_\xi F(x_*)\big) \geq 0. \tag{8}$$

Here ∂f denotes Clarke's subgradient ([4]):

$$\partial f(x_*) := \{x^* \in X^* | \langle x^*, v\rangle \leq f^\circ(x_*, v)\ \forall v \in X\}.$$

In the case when ξ is the trivial approximation $(x_*, t_n)_{n\in\mathbb{N}}$, the notation $\partial f(x_*)$ could also stand for the so-called Michel-Penot subgradient (cf. [12]).

The above result incorporates and generalizes most of the multiplier rules obtained in the papers [1, 2, 3, 9, 11, 16, 19].

3 Main Result

Our main result is contained in the following theorem which offers a generalization of Theorem 1.1.3 in [10]. It takes the same convexity assumptions as condition (iii) of Theorem 1 but all the other regularity assumptions are considerably weaker than those of Theorem 1. This result also incorporates the notion of approximation, therefore higher-order necessary conditions of optimality also easily follow from it.

Theorem 3. *Let $f_0, f_1, \ldots, f_k : D \times U \to \mathbb{R}$, $F : D \times U \to Y$, and $(x_*, u_*) \in D \times U$. Assume that*

(i) For every $u \in U$, the mapping $x \mapsto \big(f_0(x,u), \ldots, f_k(x,u), F(x,u)\big)$ is locally Lipschitz at x_*;

(ii) The partial subgradients $\partial_x f_0(x_*, u_*), \ldots, \partial_x f_1(x_*, u_*) \subset X^*$ are norm-compact sets of linear functionals;

(iii) There is compact convex set $\mathcal{A} \subset L(X,Y)$ which is a strict prederivative of the map $x \mapsto F(x, u_*)$ at x_*;

(iv) The range of each $A \in \mathcal{A}$ has finite codimension in Y;

(v) For every $x \in D$, the mapping $u \mapsto \big(f_0(x,u), \ldots, f_k(x,u), F(x,u)\big)$ satisfies the following convexity assumption: for every $u_1, u_2 \in U$ and $\alpha \in [0,1]$, there exists an element $u \in U$ such that

$$f_i(x,u) \leq \alpha f_i(x,u_1) + (1-\alpha) f_i(x,u_2) \qquad \big(i \in \{0,\ldots,k\}\big),$$
$$F(x,u) = \alpha F(x,u_1) + (1-\alpha) F(x,u_2).$$

If the pair (x_*, u_*) is a local minimum of the problem $(\mathcal{CP})$, then, for all approximations ξ of x_* such that $x \mapsto \big(f_0(x,u_*), \ldots, f_k(x,u_*), F(x,u_*)\big)$ is a differentiable map with respect to ξ, there exist multipliers $\lambda_0, \ldots, \lambda_k \geq 0$ and $y^* \in Y^*$, not all zero, such that

$$\lambda_1 f_1(x_*, u_*) = 0, \ \ldots, \ \lambda_k f_k(x_*, u_*) = 0, \tag{9}$$

$$0 \in \lambda_0 \partial_x f_0(x_*, u_*) + \cdots + \lambda_k \partial_x f_k(x_*, u_*) + y^* \circ \mathcal{A}, \tag{10}$$

$$\lambda_0 D_\xi f_0(x_*, u_*) + \cdots + \lambda_k D_\xi f_k(x_*, u_*) + y^*\big(D_\xi F(x_*, u_*)\big) \geq 0, \tag{11}$$

and, with the notation (1), the minimum principle

$$\min_{u \in U} \mathcal{L}(x_*, u, \lambda_0, \ldots, \lambda_k, y^*) = \mathcal{L}(x_*, u_*, \lambda_0, \ldots, \lambda_k, y^*) \tag{12}$$

holds.

Proof. For convenience, we split the proof of the theorem in five steps.

Step 1. We construct a family of optimization problems, each one satisfying the regularity assumptions of Theorem 2. Let $m \in \mathbb{N}$ and $u_1, \ldots, u_m$ be arbitrarily fixed elements of U and denote the m-tuple $(u_1, \ldots, u_m)$ by $\mathbf{u}$. Given a function φ defined on $D \times U$, introduce the transformed function $\mathcal{T}_{(u_1,\ldots,u_m)}\varphi = \mathcal{T}_{\mathbf{u}}\varphi$ defined on $D \times \mathbb{R}^m$ as follows:

$$\begin{aligned}&\mathcal{T}_{\mathbf{u}}\varphi(x, \alpha_1, \ldots, \alpha_m)\\ &\quad := (1 - \alpha_1 - \cdots - \alpha_m)\varphi(x, u_*) + \alpha_1 \varphi(x, u_1) + \cdots + \alpha_m \varphi(x, u_m).\end{aligned}$$

Now, given $\mathbf{u} = (u_1, \ldots, u_m) \in U^m$, we consider the following optimization problem

$$\left\{\begin{array}{c}\text{Minimize } \mathcal{T}_{\mathbf{u}} f_0(x, \alpha_1, \ldots, \alpha_m) \text{ with respect to}\\ x \in D, \quad \alpha_1, \ldots, \alpha_m \geq 0,\\ \mathcal{T}_{\mathbf{u}} f_i(x, \alpha_1, \ldots, \alpha_m) \leq 0 \quad \big(i \in \{1, \ldots, k\}\big),\\ \mathcal{T}_{\mathbf{u}} F(x, \alpha_1, \ldots, \alpha_m) = 0,\end{array}\right. \tag{$\mathcal{T}_{\mathbf{u}}\mathcal{CP}$}$$

which is called the $\mathcal{T}_{\mathbf{u}}$ transformation of the original problem $(\mathcal{CP})$. The advantage of dealing with $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$ is that it does not contain a control variable, i.e., it is an ordinary optimization problem with $(m + k)$ scalar inequalities and a Banach space valued equality constraint.

Step 2. We claim that, for all fixed $\mathbf{u} = (u_1, \ldots, u_m) \in U^m$, the point $(x_*, 0, \ldots, 0) \in U \times \mathbb{R}^m$ is a local minimum of the transformed problem $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$. Assume, on the contrary, that $(x_*, 0, \ldots, 0)$ is not a local minimum of $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$. Then there exist sequences $x_n \to x_*$ and $(\alpha_{1n}, \ldots, \alpha_{mn}) \to (0, \ldots, 0)$ such that, for all $n \in \mathbb{N}$, $\alpha_{1n} \geq 0, \quad \ldots, \quad \alpha_{mn} \geq 0$, and

$$\begin{gathered}\mathcal{T}_{\mathbf{u}} f_0(x_n, \alpha_{1n}, \ldots, \alpha_{mn}) < \mathcal{T}_{\mathbf{u}} f_0(x_*, 0, \ldots, 0) = f_0(x_*, u_*),\\ \mathcal{T}_{\mathbf{u}} f_i(x_n, \alpha_{1n}, \ldots, \alpha_{mn}) \leq 0, \qquad \big(i \in \{1, \ldots, k\}\big),\\ \mathcal{T}_{\mathbf{u}} F(x_n, \alpha_{1n}, \ldots, \alpha_{mn}) = 0.\end{gathered}$$

For large n, we have that $1 - \alpha_{1n} - \cdots - \alpha_{mn} \geq 0$, therefore, the left hand sides of the above inequalities and equation are convex combinations of $\varphi(x_n, u_*)$, $\varphi(x_n, u_1), \ldots, \varphi(x_n, u_m)$, where $\varphi \in \{f_0, \ldots, f_k, F\}$. Thus, by assumption (v) of the theorem, there exits $v_n \in U$ such that

$$f_i(x_n, v_n) \le \mathcal{T}_{\mathbf{u}} f_i(x_n, \alpha_{1n}, \dots, \alpha_{mn}), \qquad \big(i \in \{0, \dots, k\}\big),$$
$$F(x_n, v_n) = \mathcal{T}_{\mathbf{u}} F(x_n, \alpha_{1n}, \dots, \alpha_{mn}).$$

Hence

$$f_1(x_n, v_n) \le 0, \quad \dots, \quad f_k(x_n, v_n) \le 0, \quad F(x_n, v_n) = 0,$$

i.e., (x_n, v_n) is an admissible solution of $(\mathcal{CP})$. If n is large enough, then the local optimality of (x_*, u_*) yields that $f_0(x_n, v_n) \ge f(x_*, u_*)$, which contradicts

$$f_0(x_n, v_n) \le \mathcal{T}_{\mathbf{u}} f_0(x_n, \alpha_{1n}, \dots, \alpha_{mn}) < \mathcal{T}_{\mathbf{u}} f_0(x_*, 0, \dots, 0) = f_0(x_*, u_*)$$

and proves our claim.

Step 3. Now we check that the data of $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$ satisfy the assumptions of Theorem 2. Observe that the functions $\mathcal{T}_{\mathbf{u}} f_0, \dots, \mathcal{T}_{\mathbf{u}} f_k$ are locally Lipschitz at the point $(x_*, 0, \dots, 0) \in D \times \mathbb{R}^k$. It is also not difficult to see that the operators $\mathcal{T}_{\mathbf{u}} A : X \times \mathbb{R}^m \to Y$ defined by

$$\mathcal{T}_{\mathbf{u}} A(x, t_1, \dots, t_m) := A(x) + \sum_{j=1}^{m} \big(F(x_*, u_j) - F(x_*, u_*)\big) t_j,$$

where $A \in \mathcal{A}$, form a compact convex strict prederivative for $\mathcal{T}_{\mathbf{u}} F$ at the point $(x_*, 0, \dots, 0)$. The range of A is finite codimensional, therefore (by Banach's open mapping theorem) it is closed. Thus the range of $\mathcal{T}_{\mathbf{u}} A$ will also be closed (since it is of the form $A(X) + S$, where S is a finite dimensional subspace of Y). If $\xi = (x_n, t_n)_{n \in \mathbb{N}}$ is an approximation of the point x_* such that the map $x \mapsto \big(f_0(x, u_*), \dots, f_k(x, u_*), F(x, u_*)\big)$ is differentiable with respect to ξ, then the sequence $\eta = \mathcal{T}_{\mathbf{u}} \xi = ((x_n, 0, \dots, 0), t_n)_{n \in \mathbb{N}}$ is obviously an approximation of $(x_*, 0, \dots, 0) \in D \times \mathbb{R}^m$ and

$$D_\eta \mathcal{T}_{\mathbf{u}} f_i = D_\xi f_i(x_*, u_*) \qquad \big(i \in \{0, \dots, k\}\big), \qquad D_\eta \mathcal{T}_{\mathbf{u}} F = D_\xi F(x_*, u_*).$$

Step 4. We now write down the optimality conditions for $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$. The Lagrange functional of the problem $(\mathcal{T}_{\mathbf{u}}\mathcal{CP})$ is defined by

$$\mathcal{T}_{\mathbf{u}} \mathcal{L}(x, \alpha_1, \dots, \alpha_m; \mu_1, \dots, \mu_m, \lambda_0, \lambda_1, \dots, \lambda_k, y^*)$$
$$:= -\sum_{j=1}^{m} \mu_j \alpha_j + \sum_{i=0}^{k} \lambda_i \mathcal{T}_{\mathbf{u}} f_i(x, \alpha_1, \dots, \alpha_m) + y^* \big(\mathcal{T}_{\mathbf{u}} F(x, \alpha_1, \dots, \alpha_m)\big).$$

Applying Theorem 2, we get that there exist multipliers $\mu_1, \dots, \mu_m \ge 0$, $\lambda_0, \lambda_1, \dots, \lambda_k \ge 0$ and $y^* \in Y^*$, not all zero, such that

$$\lambda_i \mathcal{T}_{\mathbf{u}} f_i(x_*, 0, \dots, 0) = \lambda_i f_i(x_*, u_*) = 0 \qquad \big(i \in \{1, \dots, k\}\big), \tag{13}$$

$$0 \in \sum_{i=0}^{k} \lambda_i \partial_x f_i(x_*, u_*) + y^* \circ \mathcal{A}, \tag{14}$$

$$-\mu_j + \sum_{i=0}^{k} \lambda_i \big(f_i(x_*, u_j) - f_i(x_*, u_*)\big) + y^* \big(F(x_*, u_j) - F(x_*, u_*)\big) = 0 \quad (15)$$

for all $j \in \{1, \dots, m\}$, and

$$\sum_{i=0}^{k} \lambda_i D_\xi f_i(x_*, u_*) + y^* \big(D_\xi F(x_*, u_*)\big) \geq 0. \quad (16)$$

Due to (15), if $\lambda_0, \dots, \lambda_k$ and y^* were zero, then $\mu_1, \dots, \mu_m$ would also be zero, a contradiction. Therefore, $\lambda_0, \dots, \lambda_k$ and y^* cannot be simultaneously zero. By the homogeneity, we can assume that

$$|\lambda_0| + \cdots + |\lambda_k| + \|y^*\| = 1. \quad (17)$$

In view of the nonnegativity of the multipliers $\mu_1, \dots, \mu_m$, it follows from (15) that, for all $u \in \{u_1, \dots, u_m\}$,

$$\mathcal{L}(x_*, u, \lambda_0, \dots, \lambda_k, y^*) \geq \mathcal{L}(x_*, u_*, \lambda_0, \dots, \lambda_k, y^*) \quad (18)$$

Therefore, from what we have proved, it follows that, for arbitrary finite subset $V = \{u_1, \dots, u_m\} \subset U$, there exist multipliers $\lambda_0, \dots, \lambda_k \geq 0$ and $y^* \in Y^*$ such that (13), (14), (16), (17), and (18) hold for $u \in V$.

Step 5. We now come back to the original problem $(\mathcal{CP})$, and show that the multipliers can be chosen so that they satisfy (13), (14), (16), (17) and (18) for all $u \in U$. Define the set

$$\Lambda \subset \partial_x f_0(x_*, u_*) \times \cdots \times \partial_x f_k(x_*, u_*) \times \mathcal{A} \times \mathbb{R}_+^{k+1} \times Y^*$$

by

$$\Lambda := \Big\{ (x_0^*, \dots, x_k^*, A, \lambda_0, \dots, \lambda_k, y^*) \Big| \sum_{i=0}^{k} \lambda_i x_i^* + y^* \circ A = 0, \sum_{i=0}^{k} |\lambda_i| + \|y^*\| = 1 \Big\}.$$

Utilizing the norm-compactness of the subgradients $\partial_x f_i(x_*, u_*)$ and $\mathcal{A}$, and that the image spaces of the elements of $\mathcal{A}$ are of finite codimension, we can establish the norm (sequential) compactness of Λ in $X^{k+1} \times L(X, Y) \times \mathbb{R}^{k+1} \times Y^*$. Let

$$\big((x_{0n}^*, \dots, x_{kn}^*, A_n, \lambda_{0n}, \dots, \lambda_{kn}, y_n^*)\big)_{n \in \mathbb{N}} \quad (19)$$

be an arbitrary sequence in Λ. Since $\partial_x f_0(x_*, u_*), \dots, \partial_x f_k(x_*, u_*)$, and $\mathcal{A}$ are norm-compact sets and $|\lambda_{0n}| + \cdots + |\lambda_{kn}| \leq 1$, we may assume that the sequence

$$\big((x_{0n}^*, \dots, x_{kn}^*, A_n, \lambda_{0n}, \dots, \lambda_{kn})\big)_{n \in \mathbb{N}} \quad (20)$$

strongly converges to an element

$$(x_{00}^*, \dots, x_{k0}^*, A_0, \lambda_{00}, \dots, \lambda_{k0}).$$

Since y_n^* belongs to the closed unit ball B_{Y^*}, hence we may also assume that y_n^* converges to an element y_0^* in the weak* topology. Thus,

$$\begin{aligned}\langle y_n^*, A_0 x\rangle &= \langle y_n^*, A_n x\rangle + \langle y_n^*, (A_0 - A_n)x\rangle \\ &= -\sum_{i=0}^{k} \lambda_{in}\langle x_{in}^*, x\rangle + \langle y_n^*, (A_0 - A_n)x\rangle,\end{aligned}$$

whence, taking the limit $n \to \infty$, we get that

$$\langle y_0^*, A_0 x\rangle = -\sum_{i=0}^{k} \lambda_{i0}\langle x_{i0}^*, x\rangle. \tag{21}$$

Finally we show that y_n^* tends to y_0^* also strongly. The operator A_0 being of finite codimension, there exist elements $y_1, \dots, y_\ell \in Y$ such that

$$(x, t_1, \dots, t_\ell) \mapsto A_0 x + t_1 y_1 + \cdots + t_\ell y_\ell$$

is surjective bounded linear operator from $X \times \mathbb{R}^\ell$ to Y. By the open mapping theorem, there exists $r > 0$ such that the image of the ball

$$\left\{(x, t_1, \dots, t_\ell) \in X \times \mathbb{R}^\ell \,\middle|\, \max(\|x\|, |t_1|, \dots, |t_\ell|) \le r\right\}$$

contains the closed unit ball B_Y of Y. Now,

$$\begin{aligned}\|y_n^* - y_0^*\| &= \sup_{y \in B_Y} |\langle y_n^* - y_0^*, y\rangle| \\ &\le \sup_{\|x\|, |t_1|, \dots, |t_\ell| \le r} |\langle y_n^* - y_0^*, A_0 x + t_1 y_1 + \cdots + t_\ell y_\ell\rangle| \\ &\le \sup_{\|x\| \le r} |\langle y_n^* - y_0^*, A_0 x\rangle| + r\sum_{i=1}^{\ell} |\langle y_n^* - y_0^*, y_i\rangle| \\ &\le \sup_{\|x\| \le r} \left|\langle y_n^*, (A_0 - A_n)x\rangle + \langle y_n^*, A_n x\rangle - \langle y_0^*, A_0 x\rangle\right| \\ &\qquad + r\sum_{i=1}^{\ell} |\langle y_n^* - y_0^*, y_i\rangle| \\ &\le \sup_{\|x\| \le r} \left|\langle y_n^*, (A_0 - A_n)x\rangle\right| + \left|\sum_{i=0}^{k} \left(\lambda_{in}\langle x_{in}^*, x\rangle - \lambda_{i0}\langle x_{i0}^*, x\rangle\right)\right| \\ &\qquad + r\sum_{i=1}^{\ell} |\langle y_n^* - y_0^*, y_i\rangle| \\ &\le r\Big(\|A_0 - A_n\| + \sum_{i=0}^{k} \left\|\lambda_{in} x_{in}^* - \lambda_{i0} x_{i0}^*\right\| + \sum_{i=1}^{\ell} |\langle y_n^* - y_0^*, y_i\rangle|\Big).\end{aligned}$$

Hence, using the norm-convergence of the sequence (20) and the weak* convergence of y_n^*, it follows that $\|y_n^* - y_0^*\| \to 0$ if $n \to \infty$. Taking the limit $n \to \infty$ in

$$|\lambda_{0n}| + \cdots + |\lambda_{kn}| + \|y_n^*\| = 1 \qquad (n \in \mathbb{N}),$$

it follows that

$$|\lambda_{00}| + \cdots + |\lambda_{k0}| + \|y_0^*\| = 1.$$

This equation together with (21) yields that

$$(x_{00}^*, \ldots, x_{k0}^*, A_0, \lambda_{00}, \ldots, \lambda_{k0}, y_0^*) \in \Lambda,$$

i.e., the sequence (19) has a strong limit in Λ, proving the sequential compactness of Λ.

Having clarified this technical point, consider now, for fixed $u \in U$, the set $\Lambda(u)$ the set of those elements $(x_0^*, \ldots, x_k^*, A, \lambda_0, \ldots, \lambda_k, y^*)$ of Λ that also satisfy (13), (16) and (18). Then $\Lambda(u)$ is a closed subset of Λ, furthermore, as we have proved, the system $\{\Lambda(u) | u \in U\}$ is centered, i.e., for each finite system $\{u_1, \ldots, u_k\} \subset U$, we have that

$$\bigcap_{i=1}^{k} \Lambda(u_i) \neq \emptyset.$$

Therefore, $\bigcap_{u \in U} \Lambda(u) \neq \emptyset$, whence it follows that there exist elements $x_0^* \in \partial_x f_0(x_*, u_*)$, $\ldots$, $x_k^* \in \partial_x f_k(x_*, u_*)$, $A \in \mathcal{A}$, and multipliers $\lambda_0, \ldots, \lambda_k \geq 0$, $y^* \in Y^*$ such that (16), (17), (18) for all $u \in U$ hold, furthermore,

$$\sum_{i=0}^{k} \lambda_i x_i^* + y^* \circ A = 0.$$

The latter equation clearly implies (14). The proof of the theorem is now complete. □

Remark 1. The validity of the first step does not require conditions (i)-(iv) of Theorem 3, it applies only the convexity condition (v). Thus, the transformed problem $(\mathcal{T}_{\mathrm{u}}\mathcal{CP})$ could also be investigated using other multiplier rules known from the literature. This may lead to other variants of Theorem 3. The assumption on the norm-compactness of $\partial_x f_0(x_*, u_*)$, $\ldots$, $\partial_x f_k(x_*, u_*)$ seems to be too strong. It is not clear if these conditions could be weakened. It is also of interest to clarify that what kind of other (possibly nonconvex) subdifferentials could be used? When applying the result of Theorem 3 to control problems, the natural candidates for strict prederivatives concerning the control equation (differential equation) and other equality constraints have to be obtained and described.

Remark 2. The particular case of Theorem 3 when condition (ii), (iii), and (iv) are replaced by

(ii') The mapping $x \mapsto (f_0(x, u_*), \ldots, f_k(x, u_*))$ is strictly differentiable at x_* in the sense of Gâteaux;

(iii') The mapping $x \mapsto F(x, u_*)$ is strictly differentiable at x_* in the sense of Fréchet;

(iv') The range of the linear operator $F_x(x_*, u_*) : X \to Y$ is of finite codimension in Y,

respectively, results (under weaker regularity assumptions) the conclusions of Theorem 1 and the inequality (11). For the proof, observe that condition (ii') implies that, for all $i \in \{0, \ldots, k\}$ the Clarke's subgradient $\partial_x f_i(x_*, u_*)$ equals $\{f_{ix}(x_*, u_*)\}$, whence condition (ii) of Theorem 3 follows. Furthermore, due to (iii'), the set $\mathcal{A} = \{F_x(x_*, u_*)\}$ is a compact convex strict prederivative for F at (x_*, u_*), whence condition (iii) of Theorem 3 is also satisfied. In the case when ξ is a first-order approximation, i.e., when ξ is of the form $(x_*, t_n)_{n\in\mathbb{N}}$ then the differentiability of $x \mapsto \big(f_0(x, u_*), \ldots, f_k(x, u_*), F(x, u_*)\big)$ with respect to ξ automatically holds and the left hand side of (11) vanishes, i.e., then the statement of Theorem 3 is equivalent to that of Theorem 1. If one takes an approximation of the form $(x_* + \sqrt{t_n}d, t_n)_{n\in\mathbb{N}}$, where d is a fixed vector of X, then one requires second-order differentiability assumptions of the data and one can deduce a second-order necessary condition from (11).

Acknowledgments. The author is indebted to the anonymous referee for the unusually detailed report and the valuable suggestions and comments.

References

1. A. Ben-Tal. Second-order and related extremality conditions in nonlinear programming. J. Optim. Theory Appl. 31(2):143–165, 1980.
2. A. Ben-Tal. Second order theory of extremum problems. In: V. Fiacco V and K. Kortanek (eds), Extremal methods and systems analysis (Internat. Sympos., Univ. Texas, Austin, 1977), Vol. 174 of Lecture Notes in Econom. and Math. Systems, pp. 336–356. Springer, Berlin, 1980.
3. A. Ben-Tal and J. Zowe. A unified theory of first and second order conditions for extremum problems in topological vector spaces. Math. Programming Stud. 19:39–76, 1982.
4. F.H. Clarke. Optimization and Nonsmooth Analysis. John Wiley & Sons Inc., New York, 1983.
5. A.J. Dubovitskii and A.A. Milyutin. Extremum problems with constraints. Dokl. Akad. Nauk SSSR 149:759–762, 1963 (in Russian).
6. A.J. Dubovitskii and A.A. Milyutin. Extremal problems with constraints. Ž. Vyčisl. Mat. i Mat. Fiz. 5:395–453, 1965 (in Russian).
7. A.J. Dubovitskii and A.A. Milyutin. Second variations in extremal problems with constraints. Dokl. Akad. Nauk SSSR 160:18–21, 1965 (in Russian).
8. A.D. Ioffe. Nonsmooth analysis: differential calculus of nondifferentiable mappings. Trans. Amer. Math. Soc. 266(1):1–56, 1991.
9. A.D. Ioffe. On some recent developments in the theory of second order optimality conditions. In: S. Dolecki (ed) Optimization (Varetz, 1988), Vol. 1405 of Lecture Notes in Math., pp. 55–68. Springer, Berlin, 1989.

10. A.D. Ioffe and V.M. Tihomirov. Theory of Extremal Problems. North-Holland, Amsterdam, 1979.
11. E.E. Levitin, A.A. Milyutin, and N.P. Osmolovskii. Higher order conditions for local minima in problems with constraints. Uspekhi Mat. Nauk 33(6(204)):85–148, 1978.
12. P. Michel and J.P. Penot. Calcul sous-differential pour des fonctions Lipschitziennes et non-Lipschitziennes. C. R. Acad. Sci. Paris, Ser. I Math. 298:269–272, 1971.
13. E.I. Nenakhov and B.I. Pshenichnyĭ. Necessary conditions for the extremum in problems of nonlinear programming. Prikl. Mat. i Programmirovanie (4):40–49, 1971.
14. Zs. Páles. Inverse and implicit function theorems for nonsmooth maps in Banach spaces. J. Math. Anal. Appl. 209(1):202–220, 1997.
15. Zs. Páles. Optimum problems with nonsmooth equality constraints. Nonlinear Anal., Theory, Methods, Appl., 2005, accepted.
16. Zs. Páles and V. Zeidan. Nonsmooth optimum problems with constraints. SIAM J. Control Optim. 32(5):1476–1502, 1994.
17. B.N. Pshenichnyĭ. Necessary conditions for the extremum in problems of partially convex programming. Kibernetika (Kiev) (2):90–93, 1969.
18. B.N. Pshenichnyĭ and E.I. Nenakhov. Necessary conditions for the minimum in problems with operator constraints. Kibernetika (Kiev) (3):35–46, 1971.
19. Q.J. Zhu. Necessary conditions for constrained optimization in smooth Banach spaces and applications. SIAM J. Optim. 12:1032–1047, 2002.

Sufficiency Conditions for Infinite Horizon Optimal Control Problems

Sabine Pickenhain[1] and Valeriya Lykina[2]

[1] Brandenburg University of Technology Cottbus, Germany.
sabine@math.tu-cottbus.de
[2] Brandenburg University of Technology Cottbus, Germany.
lykina@math.tu-cottbus.de

Summary. In this paper we formulate and use the duality concept of Klötzler (1977) for infinite horizon optimal control problems. The main idea is choosing weighted Sobolev and weighted L_p spaces as the state and control spaces, respectively. Different criteria of optimality are known for specific problems, e.g. the overtaking criterion of von Weizsäcker (1965), the catching up criterion of Gale (1967) and the sporadically catching up criterion of Halkin (1974). Corresponding to these criteria we develop the duality theory and prove sufficient conditions for local optimality. Here we use some remarkable properties of weighted spaces. An example is presented where the solution is obtained in the framework of these weighted spaces, but which does not belong to standard Sobolev spaces.

1 Introduction

It is well known that in problems of economic growth we have to deal with infinite horizon optimal control problems. The range of applications of such type of problems is large, starting with famous Ramsey accumulation model up to diverse problems in continuum mechanics. Numerous advertising models and renewable resources models go back to control problems with infinite horizon as well [8, 16]. Applications of infinite horizon problems in continuum mechanics were studied by Leizarowitz and Mizel [12], and Zaslavski [19].
The usual maximum principle cannot easily be adjusted to the case of infinite horizon problems as it was first demonstrated in an example of Halkin [9]. Since the usual transversality condition does not hold anymore, some authors have investigated particular situations where ad-hoc transversality conditions are necessary for optimality. Such transversality conditions were obtained by Aseev and Kryazhimskiy [1], Michel [14] and Smirnov [17]. The simplest way to solve optimal control problems with infinite horizon is to find a solution on a finite interval and try to extend the solution onto the whole half-axis. But there is no guarantee for the extended solution to be optimal on an infinite interval. For that reason the proof of optimality is very important and is usually

based on sufficient conditions. A lot of work has been done in the last decades to prove necessary conditions for problems in the calculus of variations, see e.g. [4, 5], and optimal control, see e.g. [6]. Results concerning sufficiency conditions were derived via Fenchel-Rockafellar duality by Rockafellar [15], Aubin and Clarke [2], Magill [13], and Benveniste and Scheinkman [3]. In our paper we use the duality concept of Klötzler [10] and a special choice of state and control spaces to obtain sufficiency conditions. Considering the exponential factor $e^{-\rho t}$ as a density function we propose to choose weighted Sobolev and weighted L_p-spaces as state and control spaces, respectively, defined in the second section. Here we include a brief review of important aspects concerning differences between Lebesgue and improper Riemann integrals, which can influence optimality on an infinite interval. According to [8] and [6], there are several optimality criteria for considered class of problems and they are introduced in section 3. The fourth section is devoted to the development of the duality theory taking some properties of weighted spaces into account. A localized problem and the corresponding dual problem are formulated in section 5. The last section includes sufficiency conditions, which are proved via linear approach in the dual problem. An example illustrating existence of optimal solution with respect to weighted spaces-while no solution in usual Sobolev spaces exists- is presented as well.

2 Problem Formulation

We deal with problems of the following type: Minimize the functional

$$J(x,u) = \int_0^\infty f(t,x(t),u(t))\nu(t)\,dt \tag{1}$$

with respect to all

$$(x,u) \in W^{1,n}_{p,\nu}(0,\infty) \times L^r_{p,\nu}(0,\infty) \tag{2}$$

fulfilling the

$$\begin{aligned} \textit{State equation} \quad & \dot{x}(t) = g(t,x(t),u(t)) \text{ a.e. on } (0,\infty), && (3)\\ \textit{Control restriction} \quad & u(t) \in U \text{ a.e. on } (0,\infty), && (4)\\ \textit{Initial condition} \quad & x(0) = x_0. && (5) \end{aligned}$$

Here U is a nonempty compact set in $\mathbb{R}^r$. The spaces $W^{1,n}_{p,\nu}(0,\infty)$ and $L^r_{p,\nu}(0,\infty)$ will be defined below.

2.1 Weighted Sobolev Spaces

We consider weighted Sobolev spaces $W^{1,n}_{p,\nu}(\Omega)$ as subspaces of weighted $L^n_{p,\nu}(\Omega)$ spaces of those absolutely continuous functions x for which both x and its derivative $\dot{x}$ lie in $L^n_{p,\nu}(\Omega)$, see [11].

Let $\Omega = [0, \infty)$ and let $\mathcal{M}^n = \mathcal{M}(\Omega; \mathbb{R}^n)$ denote the space of Lebesgue measurable functions defined on Ω with values in $\mathbb{R}^n$. The function $\nu : \Omega \to \mathbb{R}_+ \backslash \{0\}$ is a density function if $\nu \in \mathcal{M}$ and

$$\int_\Omega \nu(t)dt < \infty.$$

Let $\nu \in C(\Omega)$, $0 < \nu(t) < \infty$ be given, then we define the space $L^n_{p,\nu}(\Omega)$ by

$$L^n_{p,\nu}(\Omega) = \{x \in \mathcal{M}^n | \, \|x\|^p_p := \int_\Omega |x(t)|^p \nu(t)\, dt < \infty\}, \quad (\text{when } p \geq 2)$$

$$L^n_{\infty,\nu}(\Omega) = \{x \in \mathcal{M}^n | \, \|x\|_\infty := \operatorname{ess\,sup}_{t\in\Omega} |x(t)|\nu(t) < \infty\} \quad (\text{when } p = \infty)$$

and the weighted Sobolev space by

$$W^{1,n}_{p,\nu}(\Omega) = \{x \in \mathcal{M}^n | \, x \in L^n_{p,\nu}(\Omega), \dot{x} \in L^n_{p,\nu}(\Omega)\} \quad (p = \infty).$$

Here $\dot{x}$ is the distributional derivative of x in the sense of [18, p.49]. This space, equipped with the norm

$$\|x\|^p_{W^{1,n}_{p,\nu}} = \int_\Omega \{|x(t)| + |\dot{x}(t)|\}^p \nu(t)dt,$$

is a Banach space. For later use we also define the space

$$L^{n\times n}_{\infty,\nu^{-1}}(\Omega) = \left\{ Q \in \mathcal{M}^{n\times n} \;\middle|\; \|Q\|_\infty := \max_{i,j} \left(\operatorname{ess\,sup}_{t\in\Omega} \frac{|Q_{i,j}(t)|}{\nu(t)} \right) < \infty \right\}.$$

For $x \in L^n_{p,\nu}(\Omega)$ and $y \in L^n_{q,\nu^{1-q}}(\Omega)$ the scalar product $\ll x, y \gg$ in $L^n_2(\Omega)$ defines a continuous bilinear form, since

$$\begin{aligned} |\ll x, y \gg| &\leq \int_0^\infty |x(t)|\nu^{1/p}(t)|y(t)|\nu^{-1+1/q}(t)dt \\ &\leq \|x\|_{L^n_{p,\nu}(\Omega)} \|y\|_{L^n_{q,\nu^{1-q}}(\Omega)} \end{aligned}$$

holds true. For the special case $p = 2$ one has $[L^n_{2,\nu}(\Omega)]^* = L^n_{2,\nu}(\Omega)$ due to the Riesz representation theorem. Therefore, we obtain the following relation between the scalar products in $L^n_{2,\nu}(\Omega)$ and $L^n_2(\Omega)$: For $x \in L^n_{2,\nu}(\Omega)$ and $y \in L^n_{2,\nu^{-1}}(\Omega)$ there exists $\hat{y} \in L^n_{2,\nu}(\Omega)$ such that

$$\begin{aligned} \langle x, \hat{y} \rangle_{L^n_{2,\nu}(\Omega)} &= \ll x, y \gg_{L^n_2(\Omega)} \\ \hat{y} &= y/\nu. \end{aligned} \tag{6}$$

Equation (6) is essentially used to formulate the duality theory in the sense of Klötzler in the following sections.

Remark. It is well known, see [7], that the inclusion $L^n_{p,\nu}(\Omega) \subseteq L^n_{q,\nu}(\Omega)$ holds true for all $p \geq q$, i.e. there is a $C \in \mathbb{R}_+$ such that

$$\|x\|_{L^n_{q,\nu}} \leq C \, \|x\|_{L^n_{p,\nu}} . \tag{7}$$

Note that here and in the proofs of other sections we abbreviate $L^n_{p,\nu}(\Omega)$ by $L^n_{p,\nu}$ in the indices.

Now some aspects concerning the integral in (1) should be mentioned. We assume that the function f in (1) is continuously differentiable and allow both Lebesgue and improper Riemann integrals to appear in (1). The main difference between the Lebesgue and improper Riemann integrals is that one of them may not exist while the other one is convergent. In the case

$$\int_0^\infty |f(t, x(t), u(t))| \, \nu(t) dt < \infty, \tag{8}$$

both Lebesgue and Riemann improper integrals exist and coincide [7] and we have

$$\begin{aligned} \int_0^\infty f(t, x(t), u(t)) \nu(t) dt &= \lim_{T \to \infty} \int_0^T f(t, x(t), u(t)) d\nu(t) dt \\ &= \lim_{T \to \infty} J_T(x(t), u(t)). \end{aligned} \tag{9}$$

But it can happen, that the integral in (8), i.e. Lebesgue integral, does not exist and at the same time the Riemann integral is conditionally convergent.

3 Global Optimality Criteria

In the case of infinite horizon optimal control problems the standard optimality notion should be newly defined. Namely, there are several new optimality criteria [8], which are also suitable in the case of a divergent integral in (1). We introduce global optimality criteria for the case when the integral in (1) is understood in the Lebesgue sense.

Definition 1. *Suppose that the integral in (1) exists. Furthermore, denote the problem (1)-(5) by* (P_∞). *Let* (x^*, u^*) *be an admissible pair of* (P_∞). *For any other arbitrary admissible pair* (x, u) *and for* $T \geq 0$, *let*

$$\Delta(T) = \int_0^T f(t, x(t), u(t)) \nu(t) \, dt - \int_0^T f(t, x^*(t), u^*(t)) \nu(t) \, dt.$$

Then the pair (x^*, u^*) *is called optimal for* (P_∞) *in the sense of*

1. *criterion L1, if for any admissible pair* (x,u) *we have* $\lim_{T\to\infty} \Delta(T) \geq 0$;
2. *criterion L2, if for any admissible pair* (x,u) *there exists a moment* τ *such that for all* $T \geq \tau$ *we have* $\Delta(T) \geq 0$ *(overtaking criterion of von Weizsäcker (1965)).*

Optimality in the sense of L1 coincides with usual optimality, while L2-optimality is stronger than the first one. The definition of local optimality will be introduced later.

Remark. In the case of Riemann improper integral in (1) there are some other optimality criteria which are defined in [8].

4 Duality Theory

Before formulating the duality theory for infinite horizon optimal control problems, we prove:

Lemma 1. *Let* (x^*,u^*) *be an admissible pair of* (P_∞) *and* $S : \Omega \times \mathbb{R}^n \to \mathbb{R}$ *be a function of the form*

$$S(t,\xi) = a(t) + y(t)^T(\xi - x^*(t)) + 1/2(\xi - x^*(t))^T Q(t)(\xi - x^*(t)), \quad (10)$$

with $a \in W_1^1(\Omega)$, $y \in W^{1,n}_{q,\nu^{1-q}}(\Omega)$, *and* $Q \in W^{1,n\times n}_{\infty,\nu^{-1}}(\Omega)$ *symmetric. Assume also that* $p \geq q$. *Then, for any* $x \in W^{1,n}_{p,\nu}(\Omega)$ *with* $x(0) = x_0$, *one has:*

$$\lim_{T\to\infty} S(T,x(T)) = 0, \quad (11)$$

$$\int_0^\infty \frac{d}{dt} S(t,x(t))dt = -S(0,x_0), \quad (12)$$

Proof. Observe that

$$\tilde{S} := \int_0^\infty |S(t,x(t))|\,dt \leq \int_0^\infty |a(t)|\,dt + \int_0^\infty \left|y(t)^T(x(t) - x^*(t))\right| dt + \frac{1}{2}\int_0^\infty \left|(x(t) - x^*(t))^T Q(t)(x(t) - x^*(t))\right| dt.$$

Applying Hölder's inequality we obtain

$$\tilde{S} \leq \|a\|_{W_1^1} + \left(\int_0^\infty |y(t)|^q \nu^{1-q}(t)dt\right)^{1/q} \cdot \left(\int_0^\infty |x(t) - x^*(t)|^p \nu(t)dt\right)^{1/p}$$

$$+\frac{1}{2}\left(\int_0^\infty |(x(t)-x^*(t))|^p\,\nu(t)dt\right)^{1/p}\cdot\left(\int_0^\infty |Q(t)(x(t)-x^*(t))|^q\,\nu^{1-q}(t)dt\right)^{1/q}.$$

This yields

$$\tilde{S}\le \|a\|_{W_1^1}+\|y\|_{L^n_{q,\nu^{1-q}}}\|x-x^*\|_{L^n_{p,\nu}}+\frac{1}{2}\|x-x^*\|_{L^n_{p,\nu}}\|Q(x-x^*)\|_{L^n_{q,\nu^{1-q}}}$$

$$\le \|a\|_{W_1^1}+\|y\|_{L^n_{q,\nu^{1-q}}}\|x-x^*\|_{L^n_{p,\nu}}+C\|x-x^*\|^2_{L^n_{p,\nu}}\|Q\|_{L^{n\times n}_{\infty,\nu^{-1}}}<\infty.$$

The last estimate is true because of

$$\begin{aligned}\|Q(x-x^*)\|^q_{L^n_{q,\nu^{1-q}}}&=\int_0^\infty |Q(t)(x-x^*)(t)|^q\,\nu^{1-q}(t)dt\\&\le 2^q\int_0^\infty\left(\max_{i,j}\operatorname*{ess\,sup}_{t\ge 0}\frac{|Q_{i,j}(t)|}{\nu(t)}\right)^q|(x-x^*)(t)|^q\,\nu(t)dt\\&\le 2^q\|Q\|^q_{L^{n\times n}_{\infty,\nu^{-1}}}\cdot\|(x-x^*)\|^q_{L^n_{q,\nu}}\\&\le\left(2C\|Q\|_{L^{n\times n}_{\infty,\nu^{-1}}}\cdot\|(x-x^*)\|_{L^n_{p,\nu}}\right)^q.\end{aligned}$$

To estimate $\|(x-x^*)\|_{L^n_{q,\nu}}$ we applied (7). The convergence of $\int_0^\infty |S(t,x(t))|\,dt$ yields (11), since

$$\begin{aligned}\lim_{T\to\infty}\int_0^T S(t,x(t))dt&=\lim_{T\to\infty}\left(\int_0^{T-1}S(t,x(t))dt+\int_{T-1}^T S(t,x(t))dt\right)\\&=\lim_{T\to\infty}\int_0^T S(t,x(t))dt+\lim_{\tau\to\infty}S(\tau,x(\tau)),\end{aligned}$$

where τ is an element in $[T-1,T]$. Condition (12) can now easily be derived applying (11). □

We introduce the Hamiltonian as

$$\mathcal{H}(t,\xi,\eta)=\sup_{v\in U}H(t,\xi,v,\eta),\tag{13}$$

where

$$H(t,\xi,v,\eta)=-f(t,\xi,v)+\frac{1}{\nu(t)}<\eta,g(t,\xi,v)>$$

represents the Pontrjagin function. Furthermore, we define the set

$$Y = \left\{ S : \Omega \times \mathbb{R}^n \to \mathbb{R} \;\middle|\; \begin{array}{c} S(t,\xi) = a(t) + y(t)^T(\xi - x^*(t)) \\ \quad + \frac{1}{2}(\xi - x^*(t))^T Q(t)(\xi - x^*(t)) \\ \\ a \in W_1^1,\ y \in W^{1,n}_{\infty,\nu^{1-q}},\ Q \in W^{1,n\times n}_{\infty,\nu^{-1}} \\ Q - \text{symmetric} \\ \\ \frac{1}{\nu(t)}\partial_t S(t,\xi) + \mathcal{H}(t,\xi,\partial_\xi S(t,\xi)) \le 0 \\ \forall (t,\xi) \in \Omega \times \mathbb{R}^n \end{array} \right\}.$$

Using the dual problem formalism described in [10] we construct a problem (D_∞) and prove:

Theorem 1. *Let a problem (P_∞) be given. Then for the problem*

$$(D_\infty) \quad \textit{maximize } g_\infty(S) := -S(0,x_0) \textit{ with respect to } S \in Y,$$

the weak duality relation

$$\inf(P_\infty) \ge \sup(D_\infty) \tag{14}$$

holds.

Proof. Let (x,u) be admissible for (P_∞) and S be admissible for (D_∞), i.e. $S \in Y$. Then we have

$$\begin{aligned}
J(x,u) &= \int_0^\infty f(t,x(t),u(t))\nu(t)dt \\
&= \int_0^\infty \left(-H(t,x(t),u(t),\partial_\xi S(t,x(t)))\right)\nu(t)dt \\
&+ \int_0^\infty \left(\frac{\partial_\xi S(t,x(t))}{\nu(t)} g(t,x(t),u(t))\right)\nu(t)dt \\
&= \int_0^\infty \left(-H(t,x(t),u(t),\partial_\xi S(t,x(t))) - \frac{\partial_t S(t,x(t))}{\nu(t)}\right)\nu(t)dt \\
&+ \int_0^\infty \left(\frac{\partial_t S(t,x(t))}{\nu(t)} + \frac{\partial_\xi S(t,x(t))}{\nu(t)}\dot{x}(t)\right)\nu(t)dt \\
&\ge -\int_0^\infty \left(\mathcal{H}(t,x(t),\partial_\xi S(t,x(t))) + \frac{\partial_t S(t,x(t))}{\nu(t)}\right)\nu(t)dt \\
&+ \int_0^\infty \left(\partial_t S(t,x(t)) + \partial_\xi S(t,x(t))\dot{x}(t)\right)dt \\
&\ge -\int_0^\infty \sup_{\xi\in\mathbb{R}^n}\left\{\left(\mathcal{H}(t,\xi,\partial_\xi S(t,\xi)) + \frac{\partial_t S(t,\xi)}{\nu(t)}\right)\right\}\nu(t)dt \\
&+ \int_0^\infty \left(\partial_t S(t,x(t)) + \partial_\xi S(t,x(t))\dot{x}(t)\right)dt.
\end{aligned}$$

This shows that

$$J(x,u) \geq \int_0^\infty \frac{d}{dt} S(t,x(t))dt = \lim_{T\to\infty} \int_0^T \frac{d}{dt} S(t,x(t))dt$$
$$= \lim_{T\to\infty} S(T,x(T)) - S(0,x(0)) = -S(0,x_0),$$

completing the proof in this way. □

Remark. As we can see, the proper decision variable in the dual problem (D_∞) is (a,y,Q), but we use $S \in Y$ for simplicity.

The next two corollaries provide sufficiency conditions for global optimality in the sense of criterion L1 and criterion L2, respectively.

Corollary 1. *An admissible pair (x^*,u^*) is a global minimizer of (P_∞) in the sense of criterion L1, if there exists an admissible S^* for (D_∞), such that the following conditions are fulfilled for almost all $t > 0$:*

$$(M)\quad \mathcal{H}(t,x^*(t),\partial_\xi S^*(t,x^*(t))) = H(t,x^*(t),u^*(t),\partial_\xi S^*(t,x^*(t))),$$

$$(HJ)\qquad \tfrac{1}{\nu(t)}\partial_t S^*(t,x^*(t)) + \mathcal{H}(t,x^*(t),\partial_\xi S(t,x^*(t))) = 0.$$

Proof. This follows immediately from Theorem 1. □

Remark. The boundary condition

$$(B)\qquad \lim_{T\to\infty} S^*(T,x^*(T)) = 0 \tag{15}$$

is automatically satisfied due to Lemma 1.

Corollary 2. *An admissible pair (x^*,u^*) is a global minimizer of (P_∞) in the sense of criterion L2, if there exists a family $\{(S_T^*)\}_{T\geq\tau} \subset Y$, for a sufficiently large τ, such that the following conditions are fulfilled for almost all $t \in (0,T)$:*

$$(M_T)\ \mathcal{H}(t,x^*(t),\partial_\xi(S_T^*)(t,x^*(t))) = H(t,x^*(t),u^*(t),\partial_\xi(S_T^*)(t,x^*(t))),$$

$$(HJ_T)\quad \tfrac{1}{\nu(t)}\partial_t(S_T^*)(t,x^*(t)) + \mathcal{H}(t,x^*(t),\partial_\xi(S_T^*)(t,x^*(t))) = 0,$$

$$(B_T)\qquad \inf_{\xi\in\mathbb{R}^n} S_T^*(T,\xi) = S_T^*(T,x^*(T)). \tag{16}$$

Proof: According to criterion L2, we obtain the following inequalities for all $T \geq \tau$ and $S_T^* \in Y$:

$$
\begin{aligned}
J_T(x,u) &= \int_0^T f(t,x(t),u(t))\nu(t)dt \\
&= \int_0^T \left(-H(t,x(t),u(t),\partial_\xi S_T^*(t,x(t))) - \frac{\partial_t S_T^*(t,x(t))}{\nu(t)}\right)\nu(t)dt \\
&+ \int_0^T \left(\frac{\partial_t S_T^*(t,x(t))}{\nu(t)} + \frac{\partial_\xi S_T^*(t,x(t))}{\nu(t)}\dot{x}(t)\right)\nu(t)dt \\
&\geq -\int_0^T \left(\mathcal{H}(t,x(t),\partial_\xi S_T^*(t,x(t))) + \frac{\partial_t S_T^*(t,x(t))}{\nu(t)}\right)\nu(t)dt \qquad (17) \\
&+ \int_0^T \left(\partial_t S_T^*(t,x(t)) + \partial_\xi S_T^*(t,x(t))\dot{x}(t)\right)dt \\
&\geq -\int_0^T \sup_{\xi\in\mathbb{R}^n}\left\{\left(\mathcal{H}(t,\xi,\partial_\xi S_T^*(t,\xi)) + \frac{\partial_t S_T^*(t,\xi)}{\nu(t)}\right)\right\}\nu(t)dt \\
&+ S_T^*(T,x(T)) - S_T^*(0,x(0)) \\
&\geq \inf_{\xi\in\mathbb{R}^n} S_T^*(T,\xi) - S_T^*(0,x_0).
\end{aligned}
$$

All inequalities in (17) become equalities if the conditions (M_T), (HJ_T) and (B_T) are satisfied for the pair (x^*,u^*). This means that for all $T \geq \tau$ the strong duality relation for problems with finite horizon, see [10],

$$
J_T(x^*,u^*) = \inf_{\xi\in\mathbb{R}^n}\{S_T^*(T,\xi) - S_T^*(0,x_0)\} \qquad (18)
$$

holds. Having in mind the definition of criterion L2, we can easily see that the pair is the optimal solution of the problem (P_∞) in the sense of criterion L2. □

Remark. It follows from (16) that the transversality condition $y_T(T) = 0$ has to be satisfied for all $T \geq \tau$.

5 Formulation of the Local Problem and Local Optimality Criteria

In this section we discuss local optimality. Evidently every function from $W_{p,\nu}^{1,n}$ is absolutely continuous. For that reason the imbedding of the weighted Sobolev space into the space of continuous functions allows us to formulate the notion of strong local optimality as follows.

Definition 2. *An admissible pair (x^*,u^*) of (P_∞) is strong local optimal in the sense of criterion L1, if there is a function $\delta : \mathbb{R}_+ \to \mathbb{R}_+$ such that $J(x^*,u^*) \leq J(x,u)$ for any admissible pair (x,u) of (P_∞) satisfying $|(x(t) - x^*(t))\nu(t)| < \delta(t)$ for all $t > 0$.*

In this paper we concentrate only on L1 strong local optimality while definition of L2 strong local optimality will be omitted. The problem (P_∞) can now be localized by writing (2) in the form

$$[\,x\,,\,u\,] \in W^{1,n}_{p,\nu}(\Omega) \times L^r_{p,\nu}(\Omega),\;\; x(t) \in \mathcal{K}_{\delta,\nu}(x^*(t)),$$

where

$$\mathcal{K}_{\delta,\nu}(x^*(t)) := \{\xi \in \mathbb{R}^n \,|\, |(\xi - x^*(t))\nu(t)| < \delta(t)\,\}.$$

The localized version of problem (P_∞) will be denoted by $(P_{\infty,\mathrm{loc}})$. Now we define the set

$$Y_{\mathrm{loc}} = \left\{ S \in Y \;\middle|\; \begin{array}{c} \frac{1}{\nu(t)}\partial_t S(t,\xi) + \mathcal{H}(t,\xi,\partial_\xi S(t,\xi)) \le 0 \\ \\ \text{on } \{(t,\xi)\,|t \in \Omega,\, \xi \in \mathcal{K}_{\delta,\nu}(x^*(t))\,\} \end{array} \right\}.$$

Using this notation we now formulate an equivalent version of Theorem 1 for the localized case.

Theorem 2. *Let us consider the problem $(P_{\infty,\mathrm{loc}})$. Then, for problem*

$$(D_{\infty,\mathrm{loc}}) \quad \textit{maximize } g_\infty(S) := -S(0,x_0) \textit{ with respect to } S \in Y_{\mathrm{loc}},$$

the weak duality relation $inf(P_{\infty,\mathrm{loc}}) \ge sup(D_{\infty,\mathrm{loc}})$ holds.

Remark. The corresponding versions of Corollaries 1 and 2 hold true for $(P_{\infty,\mathrm{loc}})$ if S is local admissible in the dual problem, i.e. admissible in $(D_{\infty,\mathrm{loc}})$.

6 Sufficiency Conditions for Local Optimality

6.1 Auxiliary Result

It is important to ascertain that the adjoint variable belongs to the desirable space $W^{1,n}_{q,\nu^{1-q}}(\Omega)$. For this aim we prove

Lemma 2. *Consider an admissible pair (x^*,u^*) of (P_∞). Assume that, for some constant $C \in \mathbb{R}_+$, we have:*

$$|\partial_\xi g(t,x^*(t),u^*(t))| \le C(\|\,x^*\,\|_{L^n_{p,\nu}(\Omega)} + \|u^*\|_{L^r_{p,\nu}(\Omega)}),$$
$$\partial_\xi f(t,x^*(t),u^*(t)) \in L^n_{q,\nu}(\Omega);\;\; \omega(t,0) \in L^n_{q,\nu^{1-q}}(\Omega),$$
$$t \to \Phi(t) = \int_0^t \omega(t,s)\partial_\xi f(s,x^*(s),u^*(s))\nu(s)ds \;\; is\, in\, L^n_{q,\nu^{1-q}}(\Omega).$$

Here $\omega(t,s)$ denotes the Green matrix defined for $t \ge s$ as a solution of the system

$$\frac{d\omega(t,s)}{dt} = -\partial_\xi g(t,x^*(t),u^*(t))\omega(t,s)\,, \quad \omega(s,s) = I. \tag{19}$$

Then the solution y of the adjoint equation

$$\dot{y}(t) = -y(t)^T\partial_\xi g(t,x^*(t),u^*(t)) + \partial_\xi f(t,x^*(t),u^*(t))\nu(t) \tag{20}$$

is an element of the weighted Sobolev space $W^{1,n}_{q,\nu^{1-q}}(\Omega)$.

Proof. The solution of (20) can be written as

$$y(t) = \omega(t,0)y(0) + \Phi(t),$$

where $\omega(t,s)$ is a solution of the system (19). Noticing

$$\|y\|_{W^{1,n}_{q,\nu^{1-q}}} \leq C\left(\|y\|_{L^n_{q,\nu^{1-q}}} + \|\dot{y}\|_{L^n_{q,\nu^{1-q}}}\right),$$

we estimate first the function y itself

$$\begin{aligned}
\|y\|^q_{L^n_{q,\nu^{1-q}}} &= \int_0^\infty |\omega(t,0)y(0) + \Phi(t)|^q \nu^{1-q}(t)dt \\
&\leq 2^q \int_0^\infty |\omega(t,0)y(0)|^q \nu^{1-q}(t)dt + 2^q \int_0^\infty |\Phi(t)|^q \nu^{1-q}(t)dt \\
&= 2^q \|\omega(t,0)y(0)\|^q_{L^n_{q,\nu^{1-q}}} + 2^q \|\Phi\|^q_{L^n_{q,\nu^{1-q}}},
\end{aligned}$$

and then its distributional derivative

$$\begin{aligned}
\|\dot{y}\|^q_{L^n_{q,\nu^{1-q}}} &= \int_0^\infty \left|-y(t)^T \partial_\xi g(t,x^*(t),u^*(t)) + \partial_\xi f(t,x^*(t),u^*(t))\nu(t)\right|^q \nu^{1-q}dt \\
&\leq (2C)^q \left(\| x^* \|_{L^n_{p,\nu}} + \|u^*\|_{L^n_{p,\nu}}\right)^q \int_0^\infty |y(t)|^q \nu^{1-q}(t)dt \\
&+ 2^q \int_0^\infty |\partial_\xi f(t,x^*(t),u^*(t))\nu(t)|^q \nu^{1-q}(t)dt \\
&= \left(2\tilde{C}\|y\|_{L^n_{q,\nu^{1-q}}}\right)^q + 2^q \|\partial_\xi f\|^q_{L^n_{q,\nu}} \\
&\leq \left(4\tilde{C}\right)^q \left(\|\omega(t,0)y(0)\|^q_{L^n_{q,\nu^{1-q}}} + \|\Phi\|^q_{L^n_{q,\nu^{1-q}}}\right) + 2^q \|\partial_\xi f\|^q_{L^n_{q,\nu}}.
\end{aligned}$$

Under the assumptions of this lemma, both $\|y\|_{L^n_{q,\nu^{1-q}}}$ and $\|\dot{y}\|_{L^n_{q,\nu^{1-q}}}$ are finite and we conclude $y \in W^{1,n}_{q,\nu^{1-q}}(\Omega)$. □

6.2 The Main Result on Sufficiency Conditions

We now present the main result of this paper and prove sufficiency conditions for local optimality. We have developed the duality theory via quadratic approach in the dual problem, but we now formulate the following theorem applying the linear approach. To derive analogous sufficiency conditions by means of the quadratic approach we need some a priori assumptions which guarantee that $Q \in L^{n\times n}_{\infty,\nu^{-1}}(\Omega)$ holds. This will be a task of further studies.

Theorem 3. *Let the assumptions of Lemma 2 be satisfied for an admissible pair (x^*, u^*) of $(P_{\infty,\mathrm{loc}})$. Suppose that y solves (20) and fulfills the conditions*

$$\partial^2_{vv} H(t, x^*(t), u^*(t), y(t)) < 0, \tag{21}$$

$$\mathcal{H}(t, x^*(t), y(t)) = H(t, x^*(t), u^*(t), y(t)), \tag{22}$$

$$\partial^2_{\xi\xi} \mathcal{H}(t, x^*(t), y(t)) \text{ is negative} - \text{definite}, \tag{23}$$

almost everywhere on Ω. Then the pair (x^*, u^*) is a strong local minimizer of $(P_{\infty,\mathrm{loc}})$ in the sense of criterion L1.

Proof. In order to verify whether an S defined in (10) is admissible for the problem $(D_{\infty,\mathrm{loc}})$ we define the defect of the Hamilton-Jacobi differential equation as

$$\begin{aligned} \Lambda(t,\xi) &= \frac{1}{\nu(t)} \partial_t S(t,\xi) + \mathcal{H}(t, \xi, \partial_\xi S(t,\xi)) \\ &= \frac{1}{\nu(t)} \left(\dot{a}(t) + \dot{y}(t)^T (\xi - x^*(t)) - y(t)^T \dot{x}^*(t) \right) + \mathcal{H}(t, \xi, \partial_\xi S(t,\xi)). \end{aligned}$$

Choosing $a(t)$ from the Hamilton-Jacobi differential equation

$$\Lambda(t, x^*(t)) = 0 \tag{24}$$

we obtain

$$\dot{a}(t) = (y(t)^T \dot{x}^*(t) - \mathcal{H}(t, x^*(t), y(t)))\nu(t). \tag{25}$$

Substitution of (25) into the expression for $\Lambda(t,\xi)$ yields

$$\Lambda(t,\xi) = \frac{1}{\nu(t)} \dot{y}(t)^T (\xi - x^*(t)) + \mathcal{H}(t, \xi, \partial_\xi S(t,\xi)) - \mathcal{H}(t, x^*(t), y(t)).$$

It can easily be seen that S belongs to Y_{loc} if $x^*(t)$ maximizes $\Lambda(t,\xi)$ in a $\delta(t)$-neighborhood of $x^*(t)$ for all $t > 0$. For that reason we consider a parametric optimization problem

$$(P_t) \quad \text{maximize } \Lambda(t,\xi) \text{ with respect to } \xi \in \mathcal{K}_{\delta,\nu}(x^*(t)).$$

The first order necessary condition

$$\partial_\xi \Lambda(t, x^*(t)) = \frac{1}{\nu(t)} \dot{y}(t) + \partial_\xi \mathcal{H}(t, x^*(t), y(t)) = 0 \tag{26}$$

together with the second order sufficiency condition represented by (23) guarantee that $x^*(t)$ solves the problem (P_t) for all $t > 0$. Due to condition (21), $\partial_\xi \mathcal{H}(t, x^*(t), y(t))$ consists of only one point and the canonical equation (26) can be rewritten in form (20). Since the conditions of Lemma 2 are satisfied we obtain $y \in W^{1,n}_{q,\nu^{1-q}}(\Omega)$. It means that S has the form (10) with $Q(t) \equiv 0$ and Lemma 1 can be applied to get the condition (15) of the generalized maximum principle for the criterion L1 satisfied. The maximum condition stated in (22) and two other conditions of Corollary 1 are satisfied, what allows us to deduce that the pair (x^*, u^*) is a strong local minimizer of $(P_{\infty,\mathrm{loc}})$ in the sense of criterion L1. □

6.3 An Example

We consider the Production-Inventory Model [16]

$$\min_{u\geq 0}\left\{J(x,u)=\int_0^\infty e^{-\rho t}\left(\frac{h}{2}(x-\hat{x})^2+\frac{c}{2}(u-\hat{u})^2\right)dt\right\} \tag{27}$$

$$\dot{x}(t)=u(t)-v_0,\quad x(0)=x_0, \tag{28}$$

where $h>0$, $c>0$, $\rho\geq 0$, $\hat{x}$, $\hat{u}$ denote the inventory holding cost coefficient, the production cost coefficient, the constant discount rate, the inventory goal level and the production goal level, respectively. The state equation expresses that the inventory x at time t is increased by the production rate $u(t)$ and decreased by the constant sales rate v_0.

We find the Pontrjagin function

$$H(t,\xi,v,\eta)=-\frac{h}{2}(\xi-\hat{x})^2-\frac{c}{2}(v-\hat{u})^2+\frac{\eta}{e^{-\rho t}}(v-v_0) \tag{29}$$

and verify the condition (21) of Theorem 3

$$\partial^2_{vv}H(t,\xi,v,\eta)=-c<0\Longrightarrow\partial^2_{vv}H(t,x^*(t),u^*(t),y(t))<0,$$

which is evidently satisfied. Furthermore, obtaining the control from maximum condition (22)

$$u^*(t)=\max\left\{\hat{u}+\frac{\eta}{c}e^{\rho t},0\right\},$$

we assume $\hat{u}$ to be large enough that the production rate always gives a nonnegative value:

$$u^*(t)=\hat{u}+\frac{\eta}{c}e^{\rho t}. \tag{30}$$

Substitution of u^* into (29) yields the Hamilton function

$$\mathcal{H}(t,\xi,\eta)=-\frac{h}{2}(\xi-\hat{x})^2+\frac{\eta^2}{2c}e^{2\rho t}+\eta e^{\rho t}(\hat{u}-v_0).$$

Relation (20) together with the state equation define the canonical system

$$\begin{cases}\dot{y}(t)=-y(t)^T\underbrace{\partial_\xi g(t,x^*(t),u^*(t))}_{=0}+\partial_\xi f(t,x^*(t),u^*(t))\nu(t)\\ \dot{x}^*(t)=\hat{u}+\frac{y(t)}{c}e^{\rho t}-v_0,\quad x^*(0)=x_0,\end{cases}$$

which can be rewritten as follows

$$\begin{cases}\dot{y}(t)=h(x^*(t)-\hat{x})e^{-\rho t}\\ \dot{x}^*(t)=\hat{u}+\frac{y(t)}{c}e^{\rho t}-v_0,\quad x^*(0)=x_0.\end{cases} \tag{31}$$

By differentiating the first equation and replacing the expression for $\dot{y}(t)$ by $h(x^*(t) - \hat{x})e^{-\rho t}$, one gets

$$\begin{aligned}\ddot{y}(t) &= h\left(x^*(t)e^{-\rho t} - \rho(x^*(t) - \hat{x})e^{-\rho t}\right) \\ &= h\left(\hat{u} + \frac{y(t)}{c}e^{\rho t} - v_0\right)e^{-\rho t} - \rho\dot{y}(t).\end{aligned}$$

The equation for the adjoint variable becomes

$$\ddot{y}(t) + \rho\dot{y}(t) - \frac{h}{c}y = h(\hat{u} - v_0)e^{-\rho t}.$$

By using the notation

$$k_1 = \frac{-\rho + \sqrt{\rho^2 + 4h/c}}{2}, \quad k_2 = \frac{-\rho - \sqrt{\rho^2 + 4h/c}}{2},$$

one can write $y(t) = c_1 e^{k_1 t} + c_2 e^{k_2 t} + (v_0 - \hat{u})ce^{-\rho t}$. Then we obtain the state function from the first equation of (31):

$$x^*(t) = \frac{1}{h}\left\{c_1 k_1 e^{-k_2 t} + c_2 k_2 e^{-k_1 t}\right\} - (v_0 - \hat{u})\frac{c\rho}{h} + \hat{x}.$$

The initial condition from (28) yields

$$c_2 = \frac{(x_0 - \hat{x})h - c_1 k_1 - c(\hat{u} - v)\rho}{k_2}.$$

In order to satisfy $y \in W^{1,n}_{q,\nu^{1-q}}$ it is necessary to set $\lim\limits_{t\to\infty} y(t) = 0$. This can occur only if $c_1 = 0$ holds true, otherwise $y(t)$ tends to infinity. The complete solution of (31) is stated below:

$$\begin{aligned}y(t) &= c_2 e^{k_2 t} + (v_0 - \hat{u})ce^{-\rho t}, \\ x^*(t) &= \frac{1}{h}\left\{c_2 k_2 e^{-k_1 t}\right\} - (v_0 - \hat{u})\frac{c\rho}{h} + \hat{x}.\end{aligned}$$

Using (30) we derive the control function

$$u^*(t) = \frac{c_2}{c}e^{-k_1 t} + v_0.$$

We now investigate the question concerning the spaces the solution belongs to. The function $x^*(t)$ does not belong to any usual Sobolev space W_p, since the constant $|(v_0 - \hat{u})c\rho|^p$ is not integrable over the infinite interval. The same holds true for the control function u^* as it includes a constant as well. We try to figure out whether these functions belong to some weighted Sobolev space $W^{1,n}_{p,\nu}$ and weighted $L^r_{p,\nu}$ space, respectively. Moreover, we will show that for all ω and p satisfying

$$\frac{\omega}{p} < \rho \tag{32}$$

we have $x^* \in W^{1,n}_{p,\nu}$, $y \in W^{1,n}_{q,\nu^{1-q}}$, and $u^* \in L^r_{p,\nu}$, where $\nu(t) = e^{-\omega t}$, $\omega > 0$. For that purpose we estimate as follows

$$\|x^*\|^p_{L^n_{p,\nu}} = \int_0^\infty \left| \frac{1}{h}\left(c_2 k_2 e^{-k_1 t} + (\hat{u} - v_0)c\right) + \hat{x} \right|^p e^{-\omega t} dt$$

$$\leq 2^p \int_0^\infty \left| \frac{c_2 k_2}{h} \right|^p e^{(-pk_1 - \omega)t} dt + 2^p \int_0^\infty \left| \frac{(\hat{u} - v_0)c}{h} + \hat{x} \right|^p e^{-\omega t} dt < \infty.$$

The first integral on the right-hand-side converges due to the positivity of k_1, and it allows us to say $x^* \in L^n_{p,\nu}$. Almost the same estimate for $\|\dot{x}^*\|^p_{L^n_{p,\nu}}$ implies $x^* \in W^{1,n}_{p,\nu}$. We repeat the whole procedure for $\|u^*\|^p_{L^r_{p,\nu}}$ and obtain

$$\|u^*\|^p_{L^r_{p,\nu}} = \int_0^\infty \left| \frac{c_2}{c} e^{-k_1 t} + v_0 \right|^p e^{-\omega t} dt$$

$$\leq 2 \left| \frac{c_2}{c} \right|^p \int_0^\infty e^{(-pk_1 - \omega)t} dt + (2v_0)^p \int_0^\infty e^{-\omega t} dt < \infty.$$

By (32), one derives $u^* \in L^r_{p,\nu}$. Setting $q = \frac{p}{p-1}$ and $1 - q = -\frac{1}{p-1}$, one gets

$$\|y\|^q_{L^n_{q,\nu^{1-q}}} = \int_0^\infty \left| c_2 e^{k_2 t} + (\hat{u} - v_0)c e^{-\rho t} \right|^{\frac{p}{p-1}} e^{\frac{\omega}{p-1} t} dt$$

$$\leq |2c_2|^{\frac{p}{p-1}} \int_0^\infty e^{(\frac{p}{p-1}k_2 + \frac{\omega}{p-1})t} dt + |2(v_0 - \hat{u})c|^{\frac{p}{p-1}} \int_0^\infty e^{\frac{-\rho p + \omega}{p-1} t} dt < \infty.$$

Repeating the same procedure for $\|\dot{y}\|^q_{L^n_{q,\nu^{1-q}}}$ we prove $y \in W^{1,n}_{q,\nu^{1-q}}$. The inclusion $y \in W^{1,n}_{q,\nu^{1-q}}$ is necessary in order to justify the application of Theorem 3. Now it remains to verify (23), but this can be easily done because $\partial^2_{\xi\xi}\mathcal{H}(t, x^*(t), y(t)) = -h < 0$. As a consequence, all the conditions of Theorem 3 are satisfied and we can conclude that the pair (x^*, u^*) is a strong local minimizer of the problem (27)-(28) in the sense of the criterion L1.

References

1. S.M. Aseev and A.V. Kryazhimskiy. The Pontryagin maximum Principle and transversality conditions for a class of optimal control problems with infinite time horizons. Interim Report IR-03-013, IIASI, Laxenburg, Austria, 2003.

2. J.P. Aubin and F.H. Clarke. Shadow prices and duality for a class of optimal control problems. SIAM J. Conrol Optim., 17(5):567-586, 1979.
3. L.M. Benveniste and J.A. Scheinkman. Duality theory for dynamic optimization models of economics: The continuous time case. J. Econ. Theory 27:1-29, 1982.
4. J. Blot and P. Michel. First-order conditions for infinite-horizon variational problems. J. Optim. Theory Appl., 88(2):339-364, 1996.
5. J. Blot and N. Hayek. Second-Order Necessary Conditions for the Infinite-Horizon Variational Problems. Math. Oper. Res., 21(4):979-990, 1996.
6. D.A. Carlson, A.B. Haurie, and A. Leizarowitz. Infinite Horizon Optimal Control. Springer-Verlag, New York, 1991.
7. J. Elstrodt. Maß und Integrationstheorie. Springer-Verlag, Heidelberg, 1996.
8. G. Feichtinger and R.F. Hartl. Optimale Kontrolle Ökonomischer Prozesse. Walter de Gruyter, Berlin-New York, 1986.
9. H. Halkin. Necessary conditions for optimal control problems with infinite horizons. Econometrica, 42(2):267-272, 1974.
10. R. Klötzler. On a general conception of duality in optimal control. Proceed. Equadiff 4, Prague, 1977.
11. A. Kufner. Weighted Sobolev Spaces. John Wiley and Sons, New York,1985.
12. A. Leizarowitz and V.J. Mizel. One dimensional infinite-horizon variational problems arising in continuum mechanics. Archive for Rational Mechanics and Analysis, 106(2):161-194, 1988.
13. M.J.P. Magill. Pricing optimal horizon problems. J. Math. Anal. Appl., 88:398-421, 1982.
14. P. Michel. On the transversality condition in infinite horizon optimal problems, Econometrica, 50(4):975-985, 1982.
15. R.T. Rockafellar. Convex processes and hamilton dynamical systems. In: Convex Analysis and Mathematical Economics. (Proc. Sympos., Univ. of Tilburg, 1978). Lecture Notes in Economics and Mathematical Systems, 168, Springer-Verlag, Berlin 1979, pp. 122-136.
16. S.P. Sethi and G.L. Thompson. Optimal Control Theory, Kluwer Academic Publishers, Boston, 2000.
17. G.V. Smirnov. Transversality condition for infinite horizon problems, J. Optim. Theory Appl., 88(3):671-688, 1996.
18. K. Yosida. Functional Analysis, Springer-Verlag, New York, 1974.
19. A.J. Zaslavski. The existence of periodic minimal energy configurations for one-dimensional infinite horizon variational problems arising in continuum mechanics. J. Math. Anal. Appl. 194:459-476, 1995.

On Nonconvex Relaxation Properties of Multidimensional Control Problems

Marcus Wagner

Cottbus University of Technology, Department of Mathematics, Karl-Marx-Str. 17, P. O. B. 10 13 44, D-03013 Cottbus, Germany. wagner@math.tu-cottbus.de

Summary. We provide two examples concerning the relaxation properties of a model problem in multidimensional control: $\int_\Omega f(Jx(t))dt \to \inf!$, $\Omega \subset \mathbb{R}^m$, $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)$, $Jx(t) \in K \subset \mathbb{R}^{nm}$ a. e. where $n \geq 2$, $m \geq 2$, $Jx(t)$ is the Jacobian of x, and K is a convex body. The first example justifies the use of quasiconvex functions with infinite values in the relaxation process. In the second one, we examinate the relaxation properties of a restricted quasiconvex envelope function f^* introduced by Dacorogna/Marcellini.

1 Introduction

1.1 The Model Problem

In the present paper, we investigate the relaxation properties of the following optimization problem:

$$(P) \quad \begin{cases} F(x) = \int_\Omega f(Jx(t))dt \longrightarrow \inf!\,, \;\; x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)\,, \\ Jx(t) = \left(\dfrac{\partial x_i}{\partial t_j}(t)\right)_{i,j} \in K \subset \mathbb{R}^{nm} \;\text{ a.e. on } \Omega\,. \end{cases}$$

In what follows, let us assume that $n \geq 1$, $m \geq 2$, $\Omega \subset \mathbb{R}^m$ is the closure of a bounded Lipschitz domain (in strong sense), $K \subset \mathbb{R}^{nm}$ is a convex body with $0 \in \operatorname{int}(K)$ and $f : K \to \mathbb{R}$ is a continuous function.

1.2 Outline and Aim of the Paper

We consider (P) as a model problem within the class of so-called Dieudonné-Rashevsky type problems. These are multidimensional control problems involving a system of first-order partial differential equations

$$Jx(t) = G(t, x(t), u(t))$$

together with general boundary conditions, phase and control restrictions (see [6, 17, 18, 19, 20, 26]). Problems of this type arise in the description of torsion of prismatic bars in the elastic case (St.-Venant's torsion and warping torsion [21, pp. 8–20]) as well as in the elastic-plastic case ([24, p. 531 f.], [25]). Another instance are optimization problems for convex bodies under geometrical restrictions, e. g. maximization of the surface for given width and diameter. These lead again to Dieudonné-Rashevsky type problems for support functions in spherical coordinates ([1], [2, p. 149 f.]). In general, variational problems with "convexity constraints" ([5, 15, 16]) allow a reformulation as Dieudonné-Rashevsky type problems by use of the fact that the convexity of a Lipschitz function $x \in W^{1,\infty}(\Omega, \mathbb{R})$ can be characterized by the variational inequality $\langle \nabla x(s) - \nabla x(t), s - t \rangle \geq 0$ for a. e. $s, t \in \operatorname{int}(\Omega)$.

In our model problem (P), the differential equations and control restrictions are reduced to $Jx(t) = u(t)$, $u(t) \in K$ a. e. on Ω. Thus we can formally omit the control variables u while the control restrictions turn into restrictions for the Jacobians: $Jx(t) \in K$ a. e. on Ω. In order to guarantee the existence of (global) minimizers, we must study – as in the multidimensional calculus of variations – the relaxation of the problem (P). So we have to look for a function $f^{\#} : K \to \mathbb{R}$ with the following properties:

(a) $f^{\#}(v) \leq f(v)$ $\forall v \in K$ what implies $F^{\#}(x) = \int_\Omega f^{\#}(Jx(t))dt \leq \int_\Omega f(Jx(t))dt = F(x)$ for all admissible functions x of (P).

(b) For all sequences of admissible functions $\{x^N\}$ with $x^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^n)} \hat{x}$ and $Jx^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^{nm})} J\hat{x}$, it holds: $F^{\#}(\hat{x}) \leq \liminf_{N\to\infty} F^{\#}(x^N)$.

(c) The minimal value of (P) (which is finite under the assumptions above) coincides with the minimal value of the problem $(P)^{\#}$ given by

$$(P)^{\#} \quad \begin{cases} F^{\#}(x) = \int_\Omega f^{\#}(Jx(t))dt \longrightarrow \text{inf!}\,, \quad x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)\,, \\ \\ Jx(t) \in K \text{ a. e. on } \Omega\,. \end{cases}$$

If a function $f^{\#} : K \to \mathbb{R}$ has the properties (a)–(c) then from a given minimizing sequence $\{x^N\}$ of (P) one can extract a subsequence $\{x^{N'}\}$ converging together with their derivatives weakly* (in the sense of $L^\infty(\Omega, \mathbb{R}^n)$ resp. $L^\infty(\Omega, \mathbb{R}^{nm})$) to a global minimizer $\hat{x}$ of $(P)^{\#}$. In multidimensional calculus of variations, the relaxation $f^{\#}$ is well-known; for $n = 1$ one gets (with some appropriate assumptions) the convex envelope f^c and for $n \geq 2$ the quasiconvex envelope f^{qc} of f (see [7, p. 228 ff., Theorem 2.1]). For multidimensional control problems, however, only the scalar case $n = 1$ was investigated (cf. [10, p. 327, Corollary 2.17., together with p. 334, Proposition 3.4. and p. 335 f., Proposition 3.6.]). In [18, 19, 20], first-order necessary conditions for problems (P) were proved in the case where the integrand f may be replaced by its convex envelope f^c, but for every $n \geq 2$ one can easily find problems (P) where the minimal value is changed when f is replaced by f^c (see [17, pp. 20 – 23, Example 2]). With this observation in mind, we conjecture that

– analogously to the vectorial case in the calculus of variations – the function $f^{\#}$ is rather quasiconvex than convex.

In order to closer determine the properties of the function $f^{\#}$ in the case $n \geq 2$, we present in this paper two examples. Our first example shows that it will not suffice, in general, to take some finite extension of $f\big|_K$ to the whole space $\mathbb{R}^{nm}$ and then to form the quasiconvex envelope of this extension. One has rather to extend f with $+\infty$ to $\mathbb{R}^{nm} \setminus K$, and so one is forced to investigate quasiconvex functions with values in $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. An appealing idea is then to define $f^{\#}$ by adding a restriction to the representation formula for the quasiconvex envelope

$$f^{qc}(v) = \inf \Big\{ \frac{1}{|\Omega|} \int_\Omega f(v + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n), \\ v + Jx(t) \in \mathbb{R}^{nm} \text{ a. e. on } \Omega \Big\} \tag{1}$$

(see [7, p. 201, Theorem 1.1., (4)]), and to take the infimum in (1) only over functions x with $v + Jx(t) \in K$ a. e. on Ω. The corresponding envelope function f^* was introduced in [8] (already in [13] in a special case) (see Definition 6 below). In our second example, however, we present a situation where this envelope f^* is missing the relaxation property (b). In a forthcoming paper the author will show that it is not f^* but its lower semicontinuous envelope which satisfies all relaxation properties (a)–(c).

The paper is organized as follows. Section 2 is devoted to convexity and quasiconvexity. In particular, we introduce quasiconvex functions taking values in $\overline{\mathbb{R}}$ and the envelope f^*. The announced examples will follow then in Sections 3 and 4.

1.3 Notation

Throughout the paper, we assume that the effective domain of a function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is always nonempty, i.e., $\operatorname{dom} f = \{v \in \mathbb{R}^{nm} \,|\, f(v) < +\infty\} \neq \emptyset$.

Definition 1. *Given a convex body $K \subset \mathbb{R}^{nm}$ with $\mathbf{0} \in \operatorname{int}(K)$, we say that a function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ belongs to the function class $\mathcal{F}_K$ if $f\big|_K \in C^0(K, \mathbb{R})$ and $f\big|_{\mathbb{R}^{nm} \setminus K} \equiv +\infty$.*

Further notations: $W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ – the space of Lipschitz n-vector functions with boundary values zero on Ω, $L^\infty(\Omega, \mathbb{R}^{nm})$ – the space of measurable, essentially bounded nm-vector functions, $C^0(K, \mathbb{R})$ – the space of continuous functions on K. $f\big|_A$ will denote the restriction of f to A; the abbreviation "a. e." is always related to the m-dimensional Lebesgue measure.

2 Quasiconvex Functions

2.1 Generalized Notions of Convexity

Definition 2. *A function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is said to be convex if Jensen's inequality is fulfilled for every $v', v'' \in \mathbb{R}^{nm}$:*

$$f(\lambda' v' + \lambda'' v'') \leq \lambda' f(v') + \lambda'' f(v'') \quad \forall\, \lambda', \lambda'' \geq 0,\ \lambda' + \lambda'' = 1. \tag{2}$$

A function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is said to be rank one convex if Jensen's inequality (2) is fulfilled in any rank one direction, i.e. for all $v', v'' \in \mathbb{R}^{nm}$ (considered as (n,m)-matrices) with $\operatorname{rank}(v' - v'') \leq 1$.

For functions $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$, we have the implication f convex $\Longrightarrow$ f rank one convex for all $n \geq 1$, $m \geq 1$ ([7, p. 102, Theorem 1.1., (i)]); if $n = 1$ or $m = 1$ then both notions are equivalent ([7, p.102, Theorem 1.1., (ii)]).

Definition 3. *Let $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ be a function bounded from below. The convex envelope $f^c : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is defined by*

$$f^c(v) = \sup\left\{ g(v) \,\middle|\, g : \mathbb{R}^{nm} \to \overline{\mathbb{R}} \text{ convex, } g \leq f \text{ on } \mathbb{R}^{nm} \right\}.$$

The rank one convex envelope $f^{rc} : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is defined by

$$f^{rc}(v) = \sup\left\{ g(v) \,\middle|\, g : \mathbb{R}^{nm} \to \overline{\mathbb{R}} \text{ rank one convex, } g \leq f \text{ on } \mathbb{R}^{nm} \right\}.$$

For any function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ bounded from below, the inequalities $f^c(v) \leq f^{rc}(v) \leq f(v)$ hold for all $v \in \mathbb{R}^{nm}$.

Definition 4. *i) (cf. [7, p.99, Definition ii]). A finite-valued function $f : \mathbb{R}^{nm} \to \mathbb{R}$ is said to be quasiconvex if f is Borel measurable, integrable on every compact subset of $\mathbb{R}^{nm}$ and satisfies Morrey's integral inequality for all $v \in \mathbb{R}^{nm}$:*

$$f(v) \leq \frac{1}{|\Omega|} \int_\Omega f(v + Jx(t))\, dt \quad \forall x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n); \tag{3}$$

or equivalently

$$f(v) = \inf\Big\{ \frac{1}{|\Omega|} \int_\Omega f(v + Jx(t)) dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n), \\ v + Jx(t) \in \mathbb{R}^{nm} \text{ a. e. on } \Omega \Big\}. \tag{4}$$

Here $\Omega \subset \mathbb{R}^m$ is the closure of a bounded Lipschitz domain (in strong sense).
ii) Let $f : \mathbb{R}^{nm} \to \mathbb{R}$ be a function bounded from below. The quasiconvex envelope $f^{qc} : \mathbb{R}^{nm} \to \mathbb{R}$ is defined by

$$f^{qc}(v) = \sup\left\{ g(v) \,\middle|\, g : \mathbb{R}^{nm} \to \mathbb{R} \text{ quasiconvex, } g \leq f \text{ on } \mathbb{R}^{nm} \right\}.$$

For functions $f : \mathbb{R}^{nm} \to \mathbb{R}$ one has the implications: f convex $\Longrightarrow$ f quasiconvex $\Longrightarrow$ f rank one convex for all $n \geq 1$, $m \geq 1$; if $n = 1$ or $m = 1$ then all these notions are equivalent (see [7, p. 102, Theorem 1.1., (ii)] for details). We will give the extension of Definition 4 (i) to functions with values in $\overline{\mathbb{R}}$ in Definition 5 below. For any function $f : \mathbb{R}^{nm} \to \mathbb{R}$ bounded from below, the different envelopes satisfy the inequalities $f^c(v) \leq f^{qc}(v) \leq f^{rc}(v) \leq f(v)$ for all $v \in \mathbb{R}^{nm}$.

Theorem 1. *i)([7, p. 201, Theorem 1.1., (1)]) Given a function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ bounded from below. Then for all $v \in \mathbb{R}^{nm}$, the convex envelope f^c admits the representation*

$$f^c(v) = \inf\Big\{\sum\nolimits_{s=1}^{nm+1} \lambda_s f(v_s) \,\Big|\, \sum_s \lambda_s = 1, \sum_s \lambda_s v_s = v,\ 0 \leq \lambda_s \leq 1,\ v_s \in \mathbb{R}^{nm},\ 1 \leq s \leq nm+1\Big\}.$$

ii) Given $f \in \mathcal{F}_K$ and a k-dimensional face Φ of K, $0 \leq k \leq nm$. Then for all $v \in \Phi$, f^c admits the representation

$$f^c(v) = \inf\Big\{\sum\nolimits_{s=1}^{k+1} \lambda_s f(v_s) \,\Big|\, \sum_s \lambda_s = 1, \sum_s \lambda_s v_s = v,\ 0 \leq \lambda_s \leq 1,\ v_s \in \Phi,\ 1 \leq s \leq k+1\Big\}.$$

In particular, $f^c(v) = f(v)$ for all $v \in \operatorname{ext}(K)$ and $f^c(v) = +\infty$ for all $v \in \mathbb{R}^{nm} \setminus K$. f^c is lower semicontinuous on the whole space $\mathbb{R}^{nm}$ and continuous on $\operatorname{int}(K)$.
iii) Consider a lower semicontinuous function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ bounded from below (in particular, for $f \in \mathcal{F}_K$ these assumptions are satisfied). Then in the definition of f^c, the supremum can be restricted to affine functions:

$$f^c(v) = \sup\big\{g(v) \,\big|\, g : \mathbb{R}^{nm} \to \mathbb{R} \ \textit{affine},\ g \leq f \ \textit{on}\ \mathbb{R}^{nm}\big\}.$$

Proof. (ii) is an immediate consequence of (i). (iii) follows from [12, p. 163, Conclusion 1]. □

2.2 Quasiconvex Functions which are Allowed to Take the Value $+\infty$.

Definition 5. *A function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ with the following properties is called quasiconvex:*

a) $\emptyset \neq \operatorname{dom} f \subseteq \mathbb{R}^{nm}$ is a Borel set,
b) $f|_{\operatorname{dom} f}$ is Borel measurable and integrable on every compact subset of $\operatorname{dom} f$,
c) f satisfies Morrey's integral inequality (3) for all $v \in \mathbb{R}^{nm}$.

Within every equivalence class $Jx \in L^\infty(\Omega, \mathbb{R}^{nm})$ we find some Borel measurable representative u (by [4, p. 406, Theorem 5], there exists a representative of second Baire class which is Borel measurable by [4, p. 403, Theorem 4]). Then from conditions (a) and (b) it results that the compositions $f(v+u(\cdot))$ and $\chi_{\mathrm{dom} f}(v+u(\cdot))$ are Borel measurable and essentially bounded and thus integrable functions. Note that it is not allowed to change the values of the integrand f even on a Lebesgue null set of $\mathbb{R}^{nm}$.

We make use of the convention that the integral $\int_A (+\infty)\, dt$ takes the values zero or $+\infty$ if A is either a m-dimensional Lebesgue null set or has positive measure.

If a finite, measurable, locally bounded function $f : \mathbb{R}^{nm} \to \mathbb{R}$ satisfies Morrey's integral inequality then f is rank one convex and by [7, p. 29, Theorem 2.3., 2] continuous on the whole space $\mathbb{R}^{nm}$. Conversely, any finite, continuous function f is measurable and locally bounded. Consequently, when defining quasiconvexity for finite functions $f : \mathbb{R}^{nm} \to \mathbb{R}$ we can assume a priori that f is continuous.

Large parts of the theory of quasiconvexity were formulated and proved under the assumption that the functions take its values only in $\mathbb{R}$. Allowing the value $+\infty$, we have to check the validity of all corresponding assertions from new.

Theorem 2. *Given a convex body $K \subset \mathbb{R}^{nm}$ with $\mathbf{0} \in \mathrm{int}(K)$ and a function $f : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ with $\mathrm{dom} f = K$. Let $f\big|_K$ be bounded and measurable. Then it follows:*

i) For all $v \in \mathbb{R}^{nm} \setminus K$, Morrey's integral inequality holds in the form $+\infty \leq +\infty$.

ii) f satisfies Morrey's integral inequality in $v \in K$ iff

$$f(v) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f(v + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)\,, \quad v + Jx(t) \in K \text{ a. e. on } \Omega\Big\}\,. \tag{5}$$

iii) Let $\Phi \subseteq K$ be a k-dimensional face of K, $0 \leq k \leq nm$. f satisfies Morrey's integral inequality in $v \in \Phi$ iff

$$f(v) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f(v + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)\,, \quad v + Jx(t) \in \Phi \text{ a. e. on } \Omega\Big\}\,.$$

Proof. (i) Given $v \notin K$ and $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)$. Let us assume that $v + Jx(t) \in K$ a. e. on Ω. Then from Gauss' Theorem ([11, p. 133, Theorem 1, (ii)]) it follows that $v_{ij} = \frac{1}{|\Omega|}\int_\Omega \big(v_{ij} + \frac{\partial x_i}{\partial t_j}(t)\big)dt$. On the other hand, the matrix

$$\Big(\frac{1}{|\Omega|}\int_\Omega \big(v_{ij} + \frac{\partial x_i}{\partial t_j}(t)\big)dt\Big)_{i,j}$$

must belong to $\mathrm{clco}(K) = K$ by convexity of the integral (see [3, Chap. IV-6, p. 204, Corollaire]), and we get a contradiction. So we have $v + Jx(t) \in \mathbb{R}^{nm} \setminus K$ for all t of a set of positive measure and $\int_\Omega f(v + Jx(t))dt = +\infty$ since conditions (a), (b) from Definition 5 are fulfilled. Consequently, Morrey's integral inequality holds in the form $+\infty \leq +\infty$.

(ii) If $v \in K$ and $v + Jx(t) \in \mathbb{R}^{nm} \setminus K$ for all t of a set of positive measure then we have again $\int_\Omega f(v + Jx(t))dt = +\infty$. The corresponding functions can be neglected when the infimum in (4) is formed.

(iii) From Part (ii) we know that f satisfies Morrey's integral inequality in $v \in \Phi \subseteq K$ iff (5) holds. For $\Phi = K$ our assertion is true; so let $\Phi \subset K$. Then there exists a finite sequence $\Phi = \Phi_0 \subset \Phi_1 \subset \ldots \subset \Phi_s = K$ of faces of K such that there are no faces $\widetilde{\Phi}_i$ with $\Phi_i \subsetneq \widetilde{\Phi}_i \subsetneq \Phi_{i+1}$, $0 \leq i \leq s-1$. Then $\mathrm{Dim}(\Phi_0) < \mathrm{Dim}(\Phi_1) < \ldots < \mathrm{Dim}(\Phi_s) = nm$, and by [22, p. 63] every face Φ_i is an exposed face of its successor Φ_{i+1}. Assume now $v \in \Phi$ and $v + Jx(t) \in K$ a. e. on Ω. Then it follows: $v \in \Phi_{s-1} = \Phi_s \cap H_{s-1}$ where $H_{s-1} = \{v \in \mathbb{R}^{nm} \,|\, \langle a_{s-1}, v\rangle = b_{s-1}\}$ is a supporting hyperplane of $\Phi_s = K$. This means $\langle a_{s-1}, v + Jx(t)\rangle \geq b_{s-1}$ resp. $\langle a_{s-1}, Jx(t)\rangle \geq 0$ for a. e. $t \in \Omega$. Applying Gauss' theorem, we get $\langle a_{s-1}, Jx(t)\rangle = 0$ as well as $v + Jx(t) \in \Phi_{s-1}$ for a. e. $t \in \Omega$. Again the face $\Phi_{s-2} = \Phi_{s-1} \cap H_{s-2}$ is exposed with respect to Φ_{s-1} where $H_{s-2} = \{v \in \mathbb{R}^{nm} \,|\, \langle a_{s-2}, v\rangle = b_{s-2}\}$ is a supporting hyperplane of Φ_{s-1}. In general, the face $\Phi_{i-1} = \Phi_i \cap H_{i-1}$ is exposed with respect to Φ_i, $1 \leq i \leq s-1$, and repeating the same conclusions as above one arrives at $v + Jx(t) \in \Phi_0 = \Phi$ for a. e. $t \in \Omega$. □

2.3 The Restricted Quasiconvex Envelope f^*.

Definition 6. *Given a convex body $K \subset \mathbb{R}^{nm}$ with $\mathbf{0} \in \mathrm{int}(K)$ and a function $f \in \mathcal{F}_K$. Then we define the function $f^* : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ by*

$$f^*(v) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f(v + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n) \\ v + Jx(t) \in K \text{ a. e. on } \Omega\Big\}.$$

The function f^* was introduced in [13, p. 356], in the special case that K is a closed ball centered in the origin, and later in [8, p. 27, Theorem 7.2], with respect to an arbitrary convex body K. In both cases, it was assumed that $f\big|_K \in C_0(K, \mathbb{R})$. In [8] it is shown that at least for all $v \in \mathrm{int}(K)$, f^* is continuous and Morrey's integral inequality holds. Note that f^* is defined as the pointwise infimum of the uncountable family $\{f_x \,|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)\}$ where $f_x : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is defined by

$$v \longmapsto f_x(v) = \frac{1}{|\Omega|}\int_\Omega f(Jx(t) + v)dt.$$

Theorem 3. *Given a function $f \in \mathcal{F}_K$ and a k-dimensional face Φ of K, $0 \leq k \leq nm$. Then f^* can be represented as follows:*

i) For all $v \in \Phi$,

$$f^*(v) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f(v + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n) \\ v + Jx(t) \in \Phi \text{ a. e. on } \Omega\Big\}. \tag{6}$$

ii) In particular, $f^*(v) = f(v)$ for all $v \in \operatorname{ext}(K)$.
iii) For all $v \in \mathbb{R}^{nm} \setminus K$, one has $f^*(v) = +\infty$.

Proof. Part (i) is an immediate consequence of Theorem 2 (iii). Now, let $v \in \operatorname{ext}(K)$. From part (i) we know that only the null function $x \equiv \mathbf{0}$ is feasible when building the infimum for $f^*(v)$. So, it follows that $f^*(v) = f(v)$. Finally, from the proof of Theorem 2 (ii) we know that for $v \in \mathbb{R}^{nm} \setminus K$ the infimum in Definition 6 is taken over an empty set. So its value is $+\infty$. □

Theorem 4. *For $f \in \mathcal{F}_K$, one has:*

i) If $g : \mathbb{R}^{nm} \to \overline{\mathbb{R}}$ is quasiconvex and $g \le f$ on $\mathbb{R}^{nm}$, then $g \le f^*$ on $\mathbb{R}^{nm}$.
ii) $f^c(v) \le f^*(v) \le f(v)\ \forall v \in \mathbb{R}^{nm}$.

Proof. (i) Choose a quasiconvex function g with $g(v) \le f(v)\ \forall v \in \mathbb{R}^{nm}$ but $f^*(v_0) < g(v_0)$ for some $v_0 \in K$. Then there exists a function $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ with $v_0 + Jx(t) \in K$ for a. e. $t \in \Omega$ and

$$f^*(v_0) \le \frac{1}{|\Omega|}\int_\Omega f(v_0 + Jx(t))dt < g(v_0). \tag{7}$$

This leads to a contradiction since from quasiconvexity of g, with the function x from above it follows

$$g(v_0) \le \frac{1}{|\Omega|}\int_\Omega g(v_0 + Jx(t))\,dt \le \frac{1}{|\Omega|}\int_\Omega f(v_0 + Jx(t))dt\,.$$

(ii) By Theorem 1 (iii), f^c can be expressed as the pointwise supremum of the affine functions $g \le f$ only. These are quasiconvex functions, and for arbitrary $v \in \mathbb{R}^{nm}$ and $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^n)$ it holds:

$$g(v) \le \frac{1}{|\Omega|}\int_\Omega g(v + Jx(t))dt \le \frac{1}{|\Omega|}\int_\Omega f^c(v + Jx(t))dt \quad \Longrightarrow$$
$$f^c(v) = \sup\{g(v) \,|\, g \le f,\ g \text{ affine}\} \le \frac{1}{|\Omega|}\int_\Omega f^c(v + Jx(t))dt,$$

and f^c satisfies Morrey's integral inequality. As a lower semicontinuous function with $\operatorname{dom} f^c = K$, f^c fulfills also conditions a) and b) from Definition 5. Then the inequality $f^c(v) \le f^*(v)$ follows from Part (i), and $f^*(v) \le f(v)$ will follow since $x \equiv \mathbf{0}$ is always feasible when forming the infimum in Definition 6 at $v \in K$. □

3 First Example: Finite or Infinite Extension of f to $\mathbb{R}^{nm} \setminus K$?

In the following example, we consider functions of variables $v = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2}$. We will treat the four-dimensional variables as $(2,2)$-matrices but use the norm $|v| = \left(a^2 + b^2 + c^2 + d^2\right)^{1/2}$.

We define now a convex body $K \subset \mathbb{R}^{2\times 2}$ and a function $f \in \mathcal{F}_K$ in such a way that for any finite, continuous extension $\widetilde{f}: \mathbb{R}^{2\times 2} \to \mathbb{R}$ of $f\big|_K$ to the whole space $\mathbb{R}^{2\times 2}$ there exists some point $v \in K$ with $\widetilde{f}^c(v) \leq \widetilde{f}^{qc}(v) < f^c(v)$. (Thus it is impossible to extend $f^c\big|_K$ to a finite, convex function on $\mathbb{R}^{2\times 2}$, see [23, p. 505, Theorem 1]). However, from Theorem 1 (ii) and the proof of Theorem 4 (ii) we know that f^c is a lower semicontinuous, quasiconvex function. Consequently, none of the functions $\widetilde{f}^{qc}$ is the greatest quasiconvex function below f (the idea for the construction of K can be traced back to [14, p. 698 f.]).

Definition 7. *Given the points* $v_1 = \left(\begin{smallmatrix} -1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$ *and* $v_2 = \left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)$ *and the convex set* $C = \{\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2} \,\big|\, b^2 + c^2 + d^2 \leq 1\}$. *We define* $K_1 = \mathrm{co}\big(\{v_1\} \cup C\big)$, $K_2 = \mathrm{co}\big(\{v_2\} \cup C\big) \subset \mathbb{R}^{2\times 2}$, $K = K_1 \cup K_2$ *and* $f: \mathbb{R}^{2\times 2} \to \overline{\mathbb{R}}$ *by*

$$f(v) = \begin{cases} (a^2-1)^2 & \text{if } v \in K, \\ +\infty & \text{if } v \in \mathbb{R}^{2\times 2} \setminus K. \end{cases}$$

Lemma 1. *i) K is a closed convex set, with $\mathbf{0} \in \mathrm{int}(K)$, which can be represented as*

$$K = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2} \,\big|\, -1 \leq a \leq 1,\ (b + |a|)^2 + c^2 + d^2 \leq (1 - |a|)^2\}.$$

ii) $\mathrm{ext}(K) = \{v_1, v_2\} \cup \big(\mathrm{ext}(C) \setminus \{\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)\}\big)$.

Proof. (i) Denote by C_1 and C_2 the closed convex cones with vertices at v_1 resp. v_2 generated by K_1 and K_2. Thus $K = C_1 \cap C_2$ is a closed convex set. The representation formula results from the fact that every intersection of K with a hyperplane $a = const.$ can be constructed by application of a homothety with centre in v_1 or v_2 to C. For the same reason, K contains some open neighborhood $\{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{R}^{2\times 2} \,\big|\, -\varepsilon < a < \varepsilon,\ b^2 + c^2 + d^2 < \varepsilon\}$ of $\mathbf{0}$.
(ii) The extremal rays of the cones C_1 resp. C_2 are precisely the rays $\overrightarrow{v_1 v_0}$ resp. $\overrightarrow{v_2 v_0}$ with $v_0 \in \mathrm{ext}(C)$. Obviously, $\mathrm{ext}(K)$ contains the points v_1 und v_2 while $\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) = \frac{1}{2} v_1 + \frac{1}{2} v_2 \notin \mathrm{ext}(K)$. Since in every remaining point of $\mathrm{ext}(C)$ precisely two extremal rays of K_1 and K_2 intersect, these points belongs to $\mathrm{ext}(K)$. Obviously, no other points of K are extremal points. □

Theorem 5. *The function f from Definition 7 has the following properties:*

i) $f\big|_K$ belongs to $W^{1,\infty}(K, \mathbb{R})$ and is infinitely differentiable on $\mathrm{int}(K)$.

ii) For all points $\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) \in \operatorname{ext}(C)$ *with* $b \neq (-1)$, *we have* $f^c\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) = 1$ *but* $f^c\left(\begin{smallmatrix} 0 & -1 \\ c & d \end{smallmatrix}\right) = 0$.

iii) For any finite, continuous extension $\widetilde{f} : \mathbb{R}^{2\times 2} \to \mathbb{R}$ *of* $f\big|_K$ *it holds* $\widetilde{f}^{qc}\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) \leq 0$. *Consequently,* $\widetilde{f}^c(v) \leq \widetilde{f}^{qc}(v) < f^c(v)$ *for all points* $v \in \operatorname{ext}(K)$ *with sufficiently small distance to* $\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$.

Proof. (i) $f\big|_K$ is the restriction of a polynomial to K.

(ii) By Lemma 1 (ii), the points $\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) \in \operatorname{ext}(C)$ with $b \neq (-1)$ belong to $\operatorname{ext}(K)$, and we know from Theorem 1 (ii) that $f^c\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) = f\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) = 1$. It is $f^c\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) \geq 0$ since $f \geq 0$; conversely, from Theorem 1 it follows:

$$f^c\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) = f^c\Big(\frac{1}{2}v_1 + \frac{1}{2}v_2\Big) \leq \tfrac{1}{2}f(v_1) + \tfrac{1}{2}f(v_2) = 0.$$

(iii) Let $\widetilde{f} : \mathbb{R}^{2\times 2} \to \mathbb{R}$ be some finite, continuous extension of $f\big|_K$ to the whole space. Since $\operatorname{rank}(v_1 - v_2) = \operatorname{rank}\Big(\left(\begin{smallmatrix} -1 & -1 \\ 0 & 0 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 1 & -1 \\ 0 & 0 \end{smallmatrix}\right)\Big) = \operatorname{rank}\left(\begin{smallmatrix} -2 & 0 \\ 0 & 0 \end{smallmatrix}\right) = 1$, it follows:

$$\begin{aligned} \widetilde{f}^{rc}\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) \leq \tfrac{1}{2}\widetilde{f}^{rc}(v_1) + \tfrac{1}{2}\widetilde{f}^{rc}(v_2) &\leq \tfrac{1}{2}\widetilde{f}(v_1) + \tfrac{1}{2}\widetilde{f}(v_2) \\ &= \tfrac{1}{2}f(v_1) + \tfrac{1}{2}f(v_2) = 0 \end{aligned} \tag{8}$$

and $\widetilde{f}^c\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) \leq \widetilde{f}^{qc}\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) \leq \widetilde{f}^{rc}\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$. The finite quasiconvex function $\widetilde{f}^{qc}$ is continuous in the point $\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$, so there exists a neighborhood U of $\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)$ with $\widetilde{f}^c(v) \leq \widetilde{f}^{qc}(v) < \frac{1}{2}$ for all $v \in U$. However, by part (ii) we have at the same time $f^c(v) = 1$ for all $v \in U \cap \operatorname{ext}(K)$. □

4 Second Example: Calculation of f^* on the Four-Dimensional Cube $K = [-1, 1]^4$

Let $K = [-1, 1]^4 \subset \mathbb{R}^{2\times 2}$. We start with the classification of the faces of K and then calculate the representation of f^* for arbitrary $f \in \mathcal{F}_K$ on the boundary ∂K. Then we are in position to give an example of a function $f \in \mathcal{F}_K$ where f^* is different from the relaxation $f^\#$ defined in Section 1.

4.1 Classification of the Faces of K

The four-dimensional cube K admits

- 8 three-dimensional facets (cubes):

$$W_1 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} 1 & b \\ c & d \end{smallmatrix}\right)\}, \quad W_2 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} -1 & b \\ c & d \end{smallmatrix}\right)\},$$
$$W_3 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & 1 \\ c & d \end{smallmatrix}\right)\}, \quad W_4 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & -1 \\ c & d \end{smallmatrix}\right)\},$$
$$W_5 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & b \\ 1 & d \end{smallmatrix}\right)\}, \quad W_6 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & b \\ -1 & d \end{smallmatrix}\right)\},$$
$$W_7 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & b \\ c & 1 \end{smallmatrix}\right)\}, \quad W_8 = \{v \in \mathbb{R}^{2\times 2} \mid v = \left(\begin{smallmatrix} a & b \\ c & -1 \end{smallmatrix}\right)\};$$

- 24 two-dimensional faces (squares):

$$
\begin{aligned}
Q_1 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & 1 \\ c & d \end{pmatrix}\}, & Q_7 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & 1 \\ c & d \end{pmatrix}\},\\
Q_2 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & -1 \\ c & d \end{pmatrix}\}, & Q_8 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & -1 \\ c & d \end{pmatrix}\},\\
Q_3 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ 1 & d \end{pmatrix}\}, & Q_9 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ 1 & d \end{pmatrix}\},\\
Q_4 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ -1 & d \end{pmatrix}\}, & Q_{10} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ -1 & d \end{pmatrix}\},\\
Q_5 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ c & 1 \end{pmatrix}\}, & Q_{11} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ c & 1 \end{pmatrix}\},\\
Q_6 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ c & -1 \end{pmatrix}\}, & Q_{12} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ c & -1 \end{pmatrix}\},\\
Q_{13} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ 1 & d \end{pmatrix}\}, & Q_{17} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ 1 & d \end{pmatrix}\},\\
Q_{14} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ -1 & d \end{pmatrix}\}, & Q_{18} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ -1 & d \end{pmatrix}\},\\
Q_{15} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ c & 1 \end{pmatrix}\}, & Q_{19} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ c & 1 \end{pmatrix}\},\\
Q_{16} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ c & -1 \end{pmatrix}\}, & Q_{20} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ c & -1 \end{pmatrix}\},\\
Q_{21} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & b \\ 1 & 1 \end{pmatrix}\}, & Q_{23} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & b \\ -1 & 1 \end{pmatrix}\},\\
Q_{22} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ 1 & -1 \end{pmatrix}\}, & Q_{24} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & b \\ -1 & -1 \end{pmatrix}\},
\end{aligned}
$$

- 32 one-dimensional faces (edges):

$$
\begin{aligned}
S_1 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & 1 \\ 1 & d \end{pmatrix}\}, & S_5 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & -1 \\ 1 & d \end{pmatrix}\},\\
S_2 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & 1 \\ -1 & d \end{pmatrix}\}, & S_6 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & -1 \\ -1 & d \end{pmatrix}\},\\
S_3 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & 1 \\ c & 1 \end{pmatrix}\}, & S_7 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & -1 \\ c & 1 \end{pmatrix}\},\\
S_4 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & 1 \\ c & -1 \end{pmatrix}\}, & S_8 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & -1 \\ c & -1 \end{pmatrix}\},\\
S_9 &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ 1 & 1 \end{pmatrix}\}, & S_{11} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ -1 & 1 \end{pmatrix}\},\\
S_{10} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ 1 & -1 \end{pmatrix}\}, & S_{12} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} 1 & b \\ -1 & -1 \end{pmatrix}\},\\
S_{13} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & 1 \\ 1 & d \end{pmatrix}\}, & S_{17} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & -1 \\ 1 & d \end{pmatrix}\},\\
S_{14} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & 1 \\ -1 & d \end{pmatrix}\}, & S_{18} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & -1 \\ -1 & d \end{pmatrix}\},\\
S_{15} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & 1 \\ c & 1 \end{pmatrix}\}, & S_{19} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & -1 \\ c & 1 \end{pmatrix}\},\\
S_{16} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & 1 \\ c & -1 \end{pmatrix}\}, & S_{20} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & -1 \\ c & -1 \end{pmatrix}\},\\
S_{21} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ 1 & 1 \end{pmatrix}\}, & S_{23} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ -1 & 1 \end{pmatrix}\},\\
S_{22} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ 1 & -1 \end{pmatrix}\}, & S_{24} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} -1 & b \\ -1 & -1 \end{pmatrix}\},\\
S_{25} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ 1 & 1 \end{pmatrix}\}, & S_{27} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ -1 & 1 \end{pmatrix}\},\\
S_{26} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ 1 & -1 \end{pmatrix}\}, & S_{28} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & 1 \\ -1 & -1 \end{pmatrix}\},\\
S_{29} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ 1 & 1 \end{pmatrix}\}, & S_{31} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ -1 & 1 \end{pmatrix}\},\\
S_{30} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ 1 & -1 \end{pmatrix}\}, & S_{32} &= \{v \in \mathbb{R}^{2\times 2} \mid v = \begin{pmatrix} a & -1 \\ -1 & -1 \end{pmatrix}\};
\end{aligned}
$$

and 16 zero-dimensional faces (every of them consisting of one extremal point).

4.2 Calculation of f^* on ∂K

Theorem 6. *Consider $K = [-1,1]^4 \subset \mathbb{R}^{2\times 2}$ and a function $f \in \mathcal{F}_K$. Then,*

i) For all $v_0 \in \operatorname{ext}(K)$, we have $f^*(v_0) = f(v_0)$.

ii) Consider the subsets $\mathcal{G}_1 = \{S_1, \ldots, S_{32}\}$ and $\mathcal{G}_{2,3} = \{Q_3, \ldots, Q_6, Q_9, \ldots, Q_{20}\}$ of one- or two-dimensional faces of K. If $\Phi \in \mathcal{G}_1 \cup \mathcal{G}_{2,3}$ then it holds for all $v_0 \in \operatorname{ri}(\Phi)$: $f^*(v_0) = f(v_0)$.

iii) Consider the subsets $\mathcal{G}_{2,1} = \{Q_1, Q_2, Q_7, Q_8\}$ and $\mathcal{G}_{2,2} = \{Q_{21}, \ldots, Q_{24}\}$ of two-dimensional faces of K. If $\Phi \in \mathcal{G}_{2,1} \cup \mathcal{G}_{2,2}$ then it holds for all $v_0 \in \operatorname{ri}(\Phi)$: $f^*(v_0) = \left(f|_\Phi\right)^c(v_0)$.

iv) Consider the subsets $\mathcal{G}_{3,1} = \{W_3, W_4\}$, $\mathcal{G}_{3,2} = \{W_1, W_2\}$, $\mathcal{G}_{3,3} = \{W_7, W_8\}$ and $\mathcal{G}_{3,4} = \{W_5, W_6\}$ of three-dimensional facets of K. Then we have

$$f^*(v_0) = \left(f|_{\Phi \cap \{v \in \mathbb{R}^{2\times 2} \,|\, v = \left(\begin{smallmatrix} a_0 & b \\ c & d \end{smallmatrix}\right)\}}\right)^c(v_0)$$

for $\Phi \in \mathcal{G}_{3,1}$ and $v_0 = \left(\begin{smallmatrix} a_0 & b_0 \\ c_0 & d_0 \end{smallmatrix}\right) \in \operatorname{ri}(\Phi)$. Analogously, if $\Phi \in \mathcal{G}_{3,2}$ and $v_0 = \left(\begin{smallmatrix} a_0 & b_0 \\ c_0 & d_0 \end{smallmatrix}\right) \in \operatorname{ri}(\Phi)$ then it holds:

$$f^*(v_0) = \left(f|_{\Phi \cap \{v \in \mathbb{R}^{2\times 2} \,|\, v = \left(\begin{smallmatrix} a & b_0 \\ c & d \end{smallmatrix}\right)\}}\right)^c(v_0).$$

For $v_0 \in \operatorname{ri}(\Phi)$, $\Phi \in \mathcal{G}_{3,3} \cup \mathcal{G}_{3,4}$, the corresponding representations hold.

Proof. (i) This is a consequence of Theorem 3 (ii).

(ii) Consider a face $\Phi \in \mathcal{G}_{2,3} \cup \mathcal{G}_1$ and a point $v_0 \in \operatorname{ri}(\Phi)$. By Theorem 3 (i) we have

$$f^*(v_0) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f(v_0 + Jx(t))dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^2), \\ v_0 + Jx(t) \in \Phi \text{ a. e. on } \Omega\Big\}.$$

On the other hand, for any function $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^2)$ with $v_0 + Jx(t) \in \Phi$ for a. e. $t \in \Omega$, it follows that

$$Jx(t) \in \left\{\left(\begin{smallmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{smallmatrix}\right) \,\middle|\, v_{1j} = 0\right\} \cap \left\{\left(\begin{smallmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{smallmatrix}\right) \,\middle|\, v_{2k} = 0\right\}$$

with certain indices $j, k \in \{1, 2\}$. Since the Lipschitz functions x_1 and x_2 vanish on $\partial\Omega$, we conclude that $x_1(t) = x_2(t) \equiv \mathbf{0}$. Consequently, when forming $f^*(v_0)$, only the function $x = \mathbf{0}$ is feasible, and we have $f^*(v_0) = f(v_0)$.

(iii) Choose a face $\Phi \in \mathcal{G}_{2,1} \cup \mathcal{G}_{2,2}$, for example $\Phi = Q_1$, together with a point $v_0 = \left(\begin{smallmatrix} 1 & 1 \\ c_0 & d_0 \end{smallmatrix}\right) \in \operatorname{ri}(Q_1)$. We argue from Theorem 3 (i) that only the values of f on Q_1 are involved in the construction of $f^*(v_0)$, and we have $x_1(t) \equiv \mathbf{0}$ as in the previous step:

$$f^*(v_0) = \inf\Big\{\frac{1}{|\Omega|}\int_\Omega f\Big(1, 1, c_0 + \frac{\partial x}{\partial t_1}(t), d_0 + \frac{\partial x}{\partial t_2}(t)\Big)dt \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}), \\ -1 - c_0 \le \frac{\partial x}{\partial t_1}(t) \le 1 - c_0,\ -1 - d_0 \le \frac{\partial x}{\partial t_2}(t) \le 1 - d_0 \text{ a. e. on } \Omega\Big\}. \qquad (9)$$

From Theorem 4 (ii) we know that $f^c(v_0) \leq f^*(v_0)$, and by Theorem 1 (ii) we find for any $\varepsilon > 0$ points $\binom{1\ 1}{c_1\ d_1}$, $\binom{1\ 1}{c_2\ d_2}$, $\binom{1\ 1}{c_3\ d_3} \in Q_1$ (without loss of generality, they can be chosen different from v_0 and in general position) as well as numbers λ_1, λ_2, $\lambda_3 \in (0,1)$ with

$$f^c(v_0) \leq \lambda_1 f(1,1,c_1,d_1) + \lambda_2 f(1,1,c_2,d_2) + \lambda_3 f(1,1,c_3,d_3) \leq f^c(v_0) + \varepsilon,$$

$$\begin{aligned} \lambda_1(c_1 - c_0) + \lambda_2(c_2 - c_0) + \lambda_3(c_3 - c_0) &= 0, \\ \lambda_1(d_1 - d_0) + \lambda_2(d_2 - d_0) + \lambda_3(d_3 - d_0) &= 0, \\ \lambda_1 + \lambda_2 + \lambda_3 &= 1. \end{aligned}$$

Consider now a tetrahedron $C \subset \mathbb{R}^3$ with vertices

$$P_1 = \begin{bmatrix} c_1-c_0 & c_1-c_2 \\ d_1-d_0 & d_1-d_2 \end{bmatrix}^{-1} \cdot \begin{pmatrix} -(d_1-d_2) \\ c_1-c_2 \\ 0 \end{pmatrix}, \quad P_2 = \begin{bmatrix} c_2-c_0 & c_2-c_3 \\ d_2-d_0 & d_2-d_3 \end{bmatrix}^{-1} \cdot \begin{pmatrix} -(d_2-d_3) \\ c_2-c_3 \\ 0 \end{pmatrix},$$

$$P_3 = \begin{bmatrix} c_3-c_0 & c_3-c_1 \\ d_3-d_0 & d_3-d_1 \end{bmatrix}^{-1} \cdot \begin{pmatrix} -(d_3-d_1) \\ c_3-c_1 \\ 0 \end{pmatrix}, \quad P_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

For its lateral faces it holds:

$$F_1 = P_3\,P_1\,P_4 \subset \left\{ \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{11}\,(c_1-c_0) + \mu_{12}\,(d_1-d_0) \end{pmatrix} \;\middle|\; \mu_{11},\ \mu_{12} \in \mathbb{R} \right\} + P_4\,,$$

$$F_2 = P_1\,P_2\,P_4 \subset \left\{ \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{21}\,(c_2-c_0) + \mu_{22}\,(d_2-d_0) \end{pmatrix} \;\middle|\; \mu_{21},\ \mu_{22} \in \mathbb{R} \right\} + P_4\,,$$

$$F_3 = P_2\,P_3\,P_4 \subset \left\{ \begin{pmatrix} \mu_{31} \\ \mu_{32} \\ \mu_{31}\,(c_3-c_0) + \mu_{32}\,(d_3-d_0) \end{pmatrix} \;\middle|\; \mu_{31},\ \mu_{32} \in \mathbb{R} \right\} + P_4\,,$$

and the areas of their projections $F_1' = \mathbf{0}\,P_3\,P_1$, $F_2' = \mathbf{0}\,P_1\,P_2$ and $F_3' = \mathbf{0}\,P_2\,P_3$ onto the base $G = P_1\,P_2\,P_3$ satisfy the relation

$$\left|F_1'\right| : \left|F_2'\right| : \left|F_3'\right| = \lambda_1 : \lambda_2 : \lambda_3\,.$$

Consider now the family $\mathcal{H}$ of all homothetic copies of C with bases in Ω. By Vitali's covering theorem (see [9, p. 231 f., Corollary 10.6]), there exists an at most countable covering of $\operatorname{int}(\Omega)$ with mutually disjoint bases G_1, G_2, ... $\subseteq \operatorname{int}(\Omega)$ of tetrahedrons from $\mathcal{H}$ while $\left|\operatorname{int}(\Omega) \setminus \bigcup_{i=1}^{\infty} G_i\right| = 0$ holds. Identifying over every base G_i the lateral surface of the tetrahedron C_i with the graph of a function x and extending this function by $x(t) = 0$ on the null set $\Omega \setminus \bigcup_{i=1}^{\infty} G_i$, we arrive at a function x which is admissible in the construction of $f^*(v_0)$. Inserting this function into (9), we arrive at

$$\begin{aligned} f^c(v_0) \leq f^*(v_0) &\leq \frac{1}{|\Omega|} \int_{\Omega} f(1,1,c_0 + \frac{\partial x}{\partial t_1}(t), d_0 + \frac{\partial x}{\partial t_2}(t))dt \\ &= \lambda_1 f(1,1,c_1,d_1) + \lambda_2 f(1,1,c_2,d_2) + \lambda_3 f(1,1,c_3,d_3) \leq f^c(v_0) + \varepsilon, \end{aligned}$$

and we find that $f^*(v_0) = f^c(v_0) = \big(f|_{Q_1}\big)^c(v_0)$. For an arbitrary face $\Phi \in \mathcal{G}_{2,1} \cup \mathcal{G}_{2,2}$, one proceeds in a completely analogous manner.

(iv) Choose a facet $\Phi \in \cup_{s=1}^4 \mathcal{G}_{3,s}$, for example $\Phi = W_1$, together with a point $v_0 \in \mathrm{ri}(W_1)$. Then for any function $x \in W_0^{1,\infty}(\Omega, \mathbb{R}^2)$ with $v_0 + Jx(t) \in W_1$ a. e. on Ω it follows $x_1(t) \equiv 0$ as above. Consequently, $f^*(v_0)$ admits the representation

$$f^*(v_0) = \inf\Big\{ \frac{1}{|\Omega|} \int_\Omega f(1, b_0, c_0 + \frac{\partial x}{\partial t_1}(t), d_0 + \frac{\partial x}{\partial t_2}(t))dt \;\Big|\; x \in W_0^{1,\infty}(\Omega, \mathbb{R}),$$
$$\big(1,\ b_0,\ c_0 + \frac{\partial x}{\partial t_1}(t),\ d_0 + \frac{\partial x}{\partial t_2}(t)\big)^T \in W_1 \text{ a. e. on } \Omega \Big\},$$

and the construction of $f^*(v_0)$ will only depend on the values of f on a two-dimensional convex subset of W_1, namely $W_1 \cap \{v \in \mathbb{R}^{2\times 2} \,|\, v = \left(\begin{smallmatrix} a & b_0 \\ c & d \end{smallmatrix}\right)\}$. Now, the proof can be completed as in Part (iii). □

4.3 A Function $f \in \mathcal{F}_K$ with $f^* \neq f^{\#}$.

Definition 8. *Given the four-dimensional cube $K = [-1,1]^4 \subset \mathbb{R}^{2\times 2}$, we define $f : \mathbb{R}^{2\times 2} \to \overline{\mathbb{R}}$ by*

$$f(v) = \begin{cases} a + b + c + \big(1 - d^2\big) & \text{if } v \in K, \\ +\infty & \text{if } v \in \mathbb{R}^{2\times 2} \setminus K. \end{cases}$$

Theorem 7. *Using the notations of Theorem 6, the envelope f^* of the function f from Definition 8 has the following shape:*

$$f^*(v) = \begin{cases} a + b + c + \big(1 - d^2\big) & \text{if } v \in \mathrm{ext}(K) \text{ or } v \in \mathrm{ri}(\Phi), \\ & \Phi \in \mathcal{G}_1 \cup \mathcal{G}_{2,3} \cup \mathcal{G}_{3,4}, \\ a + b + c & \text{if } v \in \mathrm{int}(K) \text{ or } v \in \mathrm{ri}(\Phi), \\ & \Phi \in \mathcal{G}_{2,1} \cup \mathcal{G}_{2,2} \cup \mathcal{G}_{3,1} \cup \mathcal{G}_{3,2} \cup \mathcal{G}_{3,3}, \\ +\infty & \text{if } v \in \mathbb{R}^{2\times 2} \setminus K. \end{cases}$$

Proof. Step 1. Calculation of $f^*(v)$ on the faces of K. By Theorem 6 (i)-(ii), f^* and f coincide on $\mathrm{ext}(K)$ as well as on the relative interior of faces $\Phi \in \mathcal{G}_1 \cup \mathcal{G}_{2,3}$. For $v \in \mathrm{ri}(W_5)$ we have $c = 1$, and with Theorem 6 (iv) we find

$$f^*(v) = \big(f|_{W_5 \cap \{\, v \in \mathbb{R}^{2\times 2} \,|\, d = const.\,\}}\big)^c(v).$$

We must form the convex envelope with respect to the variables a and b only, so we arrive at $f^*(v) = a + b + c + \big(1 - d^2\big)$ again. For $v \in \mathrm{ri}(W_6)$ one argues in the same way. Let now $v \in \mathrm{ri}(\Phi)$ with $\Phi \in \mathcal{G}_{2,1}$. Then by Theorem 6 (iii), one has to form the convex envelope of f with respect to the variables c and d, and it holds

$$\big(c + \big(1 - d^2\big)\big)^c = c + \big(1 - d^2\big)^c = c.$$

For $\Phi \in \mathcal{G}_{3,1} \cup \mathcal{G}_{3,2}$, the same conclusion holds by Theorem 6 (iv). Looking for a point $v \in \mathrm{ri}(\Phi)$, $\Phi \in \mathcal{G}_{2,2}$, one has to form again the convex envelope with respect to a and b while $(1-d^2) = 0$ since $d = \pm 1$. By Theorem 6 (iv), the same result holds for $\Phi \in \mathcal{G}_{3,3}$.

Step 2. Calculation of $f^*(v)$ on $\mathrm{int}(K)$ and $\mathbb{R}^{2\times 2} \setminus K$. We take $v_0 \in \mathrm{int}(K)$. Applying Gauss theorem, we have

$$f^*(v_0) = \inf\Big\{ \frac{1}{|\Omega|} \int_\Omega \Big(a_0 + \frac{\partial x_1}{\partial t_1}(t) + b_0 + \frac{\partial x_1}{\partial t_2}(t) + c_0 + \frac{\partial x_2}{\partial t_1}(t) + \Big(1 - \Big(d_0 + \frac{\partial x_2}{\partial t_2}(t)\Big)^2\Big)\Big) dt_1 dt_2 \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^2),\ v + Jx(t) \in K \text{ a. e. on } \Omega \Big\},$$

that is to say,

$$f^*(v_0) = a_0 + b_0 + c_0 + \inf\Big\{ \frac{1}{|\Omega|} \int_\Omega \Big(1 - \Big(d_0 + \frac{\partial x_2}{\partial t_2}(t)\Big)^2\Big) dt_1 dt_2 \,\Big|\, x \in W_0^{1,\infty}(\Omega, \mathbb{R}^2),\ v + Jx(t) \in K \text{ a. e. on } t \in \Omega \Big\}. \quad (10)$$

If we can find a function $x_2 \in W_0^{1,\infty}(\Omega, \mathbb{R})$ with

$$-1 - c_0 \le \frac{\partial x_2}{\partial t_1}(t) \le 1 - c_0 \text{ and } \frac{\partial x_2}{\partial t_2}(t) \in \{-1 - d_0,\ 1 - d_0\}$$

for almost all $t \in \Omega$, then the infimum in (10) will be taken on with zero (let $x_1(t) \equiv \mathbf{0}$). Consider now a pyramid $C \subset \mathbb{R}^3$ with base $G = P_1 P_2 P_3 P_4 \subset \Omega$ and apex P_5 where the segments $P_1 P_3$ and $P_2 P_4$ are parallel to the t_1- resp. t_2-axis. In the triangle $P_1 P_3 P_5$ let

$$\tan \sphericalangle(P_5 P_1 P_3) = 1 - c_0\,,\ \tan \sphericalangle(P_1 P_3 P_5) = -1 - c_0,$$

and in the triangle $P_2 P_4 P_5$ let

$$\tan \sphericalangle(P_5 P_2 P_4) = 1 - d_0\,,\ \tan \sphericalangle(P_2 P_4 P_5) = -1 - d_0.$$

Starting from C, we can construct a function x_2 with the desired property by the same procedure as in the proof of Theorem 6 (iii). Thus $f^*(v)$ admits the claimed representation on $\mathrm{int}(K)$. By Theorem 3 (iii), we have $f^*(v) = +\infty$ for all $v \in \mathbb{R}^{2\times 2} \setminus K$. □

Obviuosly, the function $f^*\big|_K$ is upper semicontinuous. f^* is not rank one convex since, for example, $f^*(v) = f(v) = a + b + c + (1 - d^2)$ is strongly concave along the edge $S_1 \in \mathcal{G}_1$ (Theorem 7 (i)), while

$$S_1 = \{\lambda \cdot \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right) \} + (1-\lambda) \cdot \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \} \mid 0 \le \lambda \le 1\}$$

is a rank one direction.

Theorem 8. *There are a sequence of functions* $\{x^N\}$ *in* $W_0^{1,\infty}(\Omega,\mathbb{R}^2)$ *and a function* $\hat{x} \in W_0^{1,\infty}(\Omega,\mathbb{R}^2)$ *with* $x^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^2)} \hat{x}$; $Jx^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^4)} J\hat{x}$; $J\hat{x}(t)$, $Jx^N(t) \in K$ *for a. e.* $t \in \Omega$ $\forall N \in \mathbb{N}$ *and*

$$\int_\Omega f^*(J\hat{x}(t))dt > \liminf_{N\to\infty} \int_\Omega f^*(Jx^N(t)dt\,.$$

Proof. We consider two pyramids C', $C'' \subset \mathbb{R}^3$ with base $G' = P_1' P_2' P_3' P_4' \subset \Omega$ and apex P_5' resp. $G'' = P_1'' P_2'' P_3'' P_4'' \subset \Omega$ and apex P_5'', where the segments $P_1' P_3'$ and $P_1'' P_3''$ are parallel to the t_1-axis, and the segments $P_2' P_4'$ and $P_2'' P_4''$ are parallel to the t_2-axis. We define the following angles: in the triangle $P_1' P_3' P_5'$ let

$$\tan \sphericalangle(P_5' P_1' P_3') = 1\,,\ \tan \sphericalangle(P_1' P_3' P_5') = -1\,;$$

in the triangle $P_2' P_4' P_5'$ let

$$\tan \sphericalangle(P_5' P_2' P_4') = 1\,,\ \tan \sphericalangle(P_2' P_4' P_5') = -1\,;$$

in the triangle $P_1'' P_3'' P_5''$ let

$$\tan \sphericalangle(P_5'' P_1'' P_3'') = 1\,,\ \tan \sphericalangle(P_1'' P_3'' P_5'') = -1\,;$$

and in the triangle $P_2'' P_4'' P_5''$ let

$$\tan \sphericalangle(P_5'' P_2'' P_4'') = \tfrac{1}{2}\,,\ \tan \sphericalangle(P_2'' P_4'' P_5'') = -\tfrac{1}{2}\,.$$

As in the proof of Theorem 6 (iii), we construct from C' and C'' functions $\hat{x}_1$, $\hat{x}_2 \in W_0^{1,\infty}(\Omega,\mathbb{R})$. For the vector function $\hat{x} = (\hat{x}_1, \hat{x}_2)^T$ it holds: $J\hat{x}(t) \in S_1 \cup S_2 \cup S_5 \cup S_6 \cup S_{13} \cup S_{14} \cup S_{17} \cup S_{18}$ for a. e. $t \in \Omega$. Further, let us define the functions

$$x^N = \frac{N-1}{N} \cdot \hat{x} \in W_0^{1,\infty}(\Omega,\mathbb{R}^2)$$

with $x^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^2)} \hat{x}$ as well as $Jx^N \overset{*}{\rightharpoonup}_{L^\infty(\Omega,\mathbb{R}^4)} J\hat{x}$ and $Jx^N(t) \in \operatorname{int}(K)$ for a. e. $t \in \Omega$ $\forall N \in \mathbb{N}$. From Theorem 7 we conclude

$$f^*(J\hat{x}(t)) = \frac{\partial \hat{x}_1}{\partial t_1}(t) + \frac{\partial \hat{x}_1}{\partial t_2}(t) + \frac{\partial \hat{x}_2}{\partial t_1}(t) + 1 - \Big(\frac{\partial \hat{x}_2}{\partial t_2}(t)\Big)^2 \text{ a. e. on } \Omega$$

$$\Longrightarrow \int_\Omega f^*(J\hat{x}(t))dt = \frac{3}{4}|\Omega|$$

and

$$f^*(Jx^N(t)) = \frac{\partial \hat{x}_1}{\partial t_1}(t) + \frac{\partial \hat{x}_1}{\partial t_2}(t) + \frac{\partial \hat{x}_2}{\partial t_1}(t) \text{ a. e. on } \Omega$$

$$\Longrightarrow \int_\Omega f^*(Jx^N(t))dt = 0 \quad \forall N \in \mathbb{N}\,.$$

We have found that

$$\int_\Omega f^*(J\hat{x}(t))dt > \liminf_{N\to\infty} \int_\Omega f^*(Jx^N(t))dt,$$

and the envelope f^* of the function f from Definition 8 is not identical with its relaxation $f^{\#}$. □

References

1. J.A. Andrejewa and R. Klötzler. Zur analytischen Lösung geometrischer Optimierungsaufgaben mittels Dualität bei Steuerungsproblemen. Teil I. Z. Angew. Math. Mech., 64:35–44, 1984.
2. J.A. Andrejewa and R. Klötzler. Zur analytischen Lösung geometrischer Optimierungsaufgaben mittels Dualität bei Steuerungsproblemen. Teil II. Z. Angew. Math. Mech., 64:147–153, 1984.
3. N. Bourbaki. Éléments de Mathématique. Livre VI: Intégration, Chap. I–IV. Hermann, Paris, 1952.
4. C. Carathéodory. Vorlesungen über reelle Funktionen. Chelsea; New York, 3rd ed., 1968.
5. G. Carlier and T. Lachand-Robert. Régularité des solutions d'un problème variationnel sous contrainte de convexité. C. R. Acad. Sci. Paris Sér. I Math., 332:79–83, 2001.
6. L. Cesari. Optimization with partial differential equations in Dieudonné-Rashevsky form and conjugate problems. Arch. Rat. Mech. Anal., 33:339–357, 1969.
7. B. Dacorogna. Direct Methods in the Calculus of Variations. Springer, Berlin, 1989.
8. B. Dacorogna and P. Marcellini. General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial case. Acta Mathematica 178:1–37, 1997.
9. B. Dacorogna and P. Marcellini. Implicit Partial Differential Equations. Birkhäuser, Boston, 1999.
10. I. Ekeland and R. Témam. Convex Analysis and Variational Problems. SIAM Philadelphia, 2nd ed., 1999.
11. L.C. Evans and R.F. Gariepy. Measure Theory and Fine Properties of Functions. CRC Press, Boca Raton, 1992.
12. A.D. Ioffe and V.M. Tichomirov. Theorie der Extremalaufgaben. VEB Deutscher Verlag der Wissenschaften, Berlin, 1979.
13. D. Kinderlehrer and P. Pedregal. Characterizations of Young measures generated by gradients. Arch. Rat. Mech. Anal., 115:329–365, 1991.
14. J.B. Kruskal. Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping. Proc. Amer. Math. Soc., 23:697–703, 1969.
15. T. Lachand-Robert and M.A. Peletier. Minimisation de fonctionnelles dans un ensemble de fonctions convexes. C. R. Acad. Sci. Paris Sér. I Math., 325:851–855, 1997.
16. T. Lachand-Robertand M.A. Peletier. An example of non-convex minimization and an application to Newton's problem of the body of least resistance. Ann. Inst. H. Poincaré – Analyse non linéaire 18:179–198, 2001.
17. S. Pickenhain. Beiträge zur Theorie mehrdimensionaler verallgemeinerter Steuerungsprobleme. Habilitationsschrift, Universität Leipzig, 1991.
18. S. Pickenhain and M. Wagner. Critical points in relaxed deposit problems. In: A. Ioffe, S. Reich, and I. Shafrir (Eds.), Calculus of Variations and Optimal Control, Technion 98, Vol. II (Research Notes in Mathematics, Vol. 411). Chapman & Hall, CRC Press; Boca Raton, pp. 217–236, 1999.
19. S. Pickenhain and M. Wagner. Pontryagin principle for state-constrained control problems governed by a first-order PDE system. JOTA 107:297–330, 2000.

20. S. Pickenhain and M. Wagner, Piecewise continuous controls in Dieudonné-Rashevsky type problems. To appear in JOTA, Vol. 127, 2005.
21. E. Sauer. Schub und Torsion bei elastischen prismatischen Balken. Verlag Wilhelm Ernst & Sohn; Berlin - München (Mitteilungen aus dem Institut für Massivbau der TH Darmstadt 29), 1980.
22. R. Schneider. Convex Bodies: The Brunn-Minkowski Theory. Cambridge University Press; Cambridge, 1993.
23. K. Schulz and B. Schwartz. Finite extensions of convex functions. Math. Operationsforschung Statist., Ser. Optimization 10:501–509, 1979.
24. T.W. Ting. Elastic-plastic torsion of convex cylindrical bars. J. Math. Mech., 19:531–551, 1969.
25. T.W. Ting. Elastic-plastic torsion problem III. Arch. Rat. Mech. Anal., 34:228–244, 1969.
26. M. Wagner. Pontryagin's maximum principle for Dieudonné-Rashevsky type problems involving Lipschitz functions. Optimization 46:165–184, 1999.

Existence and Structure of Solutions of Autonomous Discrete Time Optimal Control Problems

Alexander J. Zaslavski[1]

Department of Mathematics, Technion, Haifa, Israel
ajzasl@techunix.technion.ac.il

Summary. In this paper we consider autonomous discrete time optimal control problems. We discuss the reduction to finite cost and the representation formula, the existence of optimal solutions on infinite horizon and their structure, and the structure of optimal solutions on finite intervals.

1 Introduction

The study of optimal control problems defined on infinite intervals has recently been a rapidly growing area of research. See, for example, [4, 5, 10, 11, 20, 30-32, 37, 38] and the references mentioned therein. These problems arise in engineering [1, 37, 38], in models of economic growth [13, 18, 19, 23, 24], in infinite discrete models of solid-state physics related to dislocations in one-dimensional crystals [3, 27] and in the theory of thermodynamical equilibrium for materials [7, 15]. In this paper we consider discrete time autonomous optimal control problems. Sections 2 and 3 are devoted to discrete time control systems on compact metric spaces. In Section 2 we present two fundamental tools in the theory of optimal control on infinite horizon: the reduction to finite cost and the representation formula established in [14]. In Section 3 we present a number of results obtained in [28, 29] which establish the existence of optimal solutions on infinite horizon and describe their structure. The turnpike theorem for infinite dimensional control system with a convex cost function obtained in [34] is discussed in Section 4. A finite dimensional extention of this result for nonconvex cost functions obtained in [35] is presented in Section 5. In Section 6 we discuss the existence of optimal solutions on infinite horizon for noncovex control systems on complete metric spaces which are not necessarily compact. The main result of Section 6 was obtained in [36]. In Section 7-10 we establish a turnpike result for a class of problems in metric spaces which are not necessarily compact.

2 Autonomous Discrete-Time Control Systems on Compact Metric Spaces

In this section we consider the infinite horizon problem of minimizing the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ as N grows to infinity where $\{x_i\}_{i=0}^{\infty}$ is a sequence in a compact metric space K and v is a continuous function on $K \times K$. This provides a convenient setting for the study of various optimization problems [3, 14-17, 27, 30-32].

Let K be a compact metric space, R^n the Euclidean n-dimensional space, $C(K \times K)$ the space of all continuous functions $v\colon K \times K \to R^1$ with the topology of the uniform convergence ($||v|| = \sup\{|v(x,y)|\colon x, y \in K\}$). Let $C(K)$ be the space of all continuous functions $v\colon K \to R^1$ with the topology of uniform convergence ($||v|| = \sup\{|v(x)|\colon x \in K\}$), and $B(K \times K)$ the set of all bounded lower semicontinuous functions $v\colon K \times K \to R^1$.

Consider any $v \in B(K \times K)$. We are interested in the limit behavior as $N \to \infty$ of the expression $\sum_{i=0}^{N-1} v(x_i, x_{i+1})$ where $\{x_i\}_{i=0}^{\infty}$ is an infinite sequence in K which we call a program (or a configuration), and which occasionally will be denoted by a bold face letter $\mathbf{x}$ (similarly $\{y_i\}_{i=0}^{\infty}$ will be denoted by $\mathbf{y}$, etc.) A finite sequence $\{x_i\}_{i=0}^{N} \subset K$ ($N = 0, 1, \ldots$) will be also called a program.

The following notion known as *the overtaking optimality criterion* was introduced in the economic literature [2, 12, 26]. A program $\{x_i\}_{i=0}^{\infty}$ is a (v)-overtaking optimal program if for every program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = x_0$ the following inequality holds:

$$\limsup_{N\to\infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \le 0.$$

A program $\{x_i\}_{i=0}^{\infty}$ is (v)-weakly optimal [6] if for every program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = x_0$ the following inequality holds:

$$\liminf_{N\to\infty} \sum_{i=0}^{N-1} [v(x_i, x_{i+1}) - v(z_i, z_{i+1})] \le 0.$$

A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ is called a (v)-minimal energy configuration [3] if for each pair of integers $n_1, n_2 > n_1$ and each sequence $\{y_i\}_{i=n_1}^{n_2} \subset K$ satisfying $y_i = x_i$, $i = n_1, n_2$ the following inequality holds:

$$\sum_{i=n_1}^{n_2-1} v(x_i, x_{i+1}) \le \sum_{i=n_1}^{n_2-1} v(y_i, y_{i+1}).$$

Of special interest is the minimal long-run average cost growth rate

$$\mu(v) = \inf\Big\{ \liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \text{ is a program} \Big\}.$$

A program $\{z_i\}_{i=0}^{\infty}$ is called a (v)-good program [12] if the sequence

$$\left\{\sum_{i=0}^{N-1}[v(z_i, z_{i+1}) - \mu(v)]\right\}_{N=1}^{\infty} \tag{1}$$

is bounded. It was proved in [14] that for every program $\{z_i\}_{i=0}^{\infty}$ the sequence (1) is either bounded or diverges to infinity and that for every initial value z there is a (v)-good program $\{z_i\}_{i=0}^{\infty}$ satisfying $z_0 = z$.

In [28, 29] we investigated the structure of (v)-good programs and established for a generic $v \in C(K \times K)$, and for every given $x \in K$ the existence of (v)-weakly optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$.

Let $v \in B(K \times K)$. We define

$$a(v) = \sup\{v(x, y)\colon x, y \in K\},\ b(v) = \inf\{v(x, y)\colon x, y \in K\}.$$

The following two results established in [14] are very useful in the study of infinite horizon control problems. We refer to the property described in Theorem 1 as the reduction to finite cost, and to the property described in Theorem 2 as the representation formula.

Theorem 1. *(i) For every program* $\{z_i\}_{i=0}^{\infty}$

$$\sum_{i=0}^{N-1}[v(z_i, z_{i+1}) - \mu(v)] \geq b(v) - a(v) \quad \forall N = 1, 2, \ldots;$$

(ii) For every program $\{z_i\}_{i=0}^{\infty}$ *the sequence (1) is either bounded or it diverges to infinity;*
(iii) For every initial value z_0 *there is a program* $\{z_i\}_{i=0}^{\infty}$ *which satisfies*

$$\left|\sum_{i=0}^{N}[v(z_i, z_{i+1}) - \mu(v)]\right| \leq 4|a(v) - b(v)| \quad \forall N = 1, 2, \ldots.$$

Theorem 2. *Let* $v \in C(K \times K)$ *and define*

$$\pi^v(x) = \inf\left\{\liminf_{N\to\infty}\sum_{i=0}^{N-1}[v(z_i, z_{i+1}) - \mu(v)]\colon \mathbf{z} \in K, z_0 = x\right\}.$$

Then we can represent $v(x, y)$ *in the form*

$$v(x, y) = \theta^v(x, y) + \mu(v) + \pi^v(x) - \pi^v(y) \text{ for } x, y \in K \tag{2}$$

where $\theta^v(x, y)$ *is defined by (2), and where* π^v, θ^v *are continuous functions,* θ^v *is nonnegative and* $E(x) = \{y \in K\colon \theta^v(x, y) = 0\}$ *is nonempty for every* $x \in K$.

In [14] these theorems were established when K was a compact subset of R^n, but their proofs remain in force also when K is any compact metric space.

3 Existence and Structure of Optimal Solutions for Discrete-Time Control Systems on Compact Spaces

Let K be a compact metric space. For a program $\mathbf{x}$ we denote by $\omega(\mathbf{x})$ the set of all points $z \in K$ such that some subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ converge to z, and denote by $\Omega(\mathbf{x})$ the set of all points $(z_1, z_2) \in K \times K$ such that some subsequence $\{(x_{i_k}, x_{i_k+1})\}_{k=1}^{\infty}$ converge to (z_1, z_2). Denote by d the metric on K and define the metric d_1 on $K \times K$ by

$$d_1((x_1, x_2), (y_1, y_2)) = d(x_1, y_1) + d(x_2, y_2) \quad (x_1, x_2, y_1, y_2 \in K).$$

We denote $d(x, B) = \inf\{d(x, y)\colon y \in B\}$ for $x \in K, B \subset K$ and

$$d_1((x_1, x_2), A) = \inf\{d_1((x_1, x_2), (y_1, y_2))\colon (y_1, y_2) \in A\}$$

for $(x_1, x_2) \in K \times K$ and $A \subset K \times K$. Denote by $\mathrm{dist}(A, B)$ the Hausdorff metric for two sets $A \subset K$ and $B \subset K$ and denote by $\mathrm{Card}(A)$ the cardinality of a set A.

We call a sequence $\{x_i\}_{i=-\infty}^{\infty} \subset K$ *almost periodic* if for every $\epsilon > 0$ there exists an integer $m \geq 1$ such that the relation $d(x_i, x_{i+mj}) \leq \epsilon$ holds for any integer i and any integer j. We call a program $\{x_i\}_{i=0}^{\infty}$ *asymptotically almost periodic* if for every $\epsilon > 0$ there exist integers $k \geq 1, m \geq 1$ such that $d(x_i, x_{i+mj}) \leq \epsilon$ for any integer $i \geq k$ and any integer $j \geq 1$.

In [28, 29] we proved the existence of a set $F \subset C(K \times K)$ which is a countable intersection of open everywhere dense sets in $C(K \times K)$ and for which the following results are valid.

Theorem 3. *(i) For every $u \in F$ there are closed sets $H(u) \subset K \times K$, $H_0(u) \subset K$ such that for every (u)-good program $\mathbf{x}$ we have $\Omega(\mathbf{x}) = H(u)$, $\omega(\mathbf{x}) = H_0(u)$.*
(ii) Let $u \in F$. Then every (u)-good program $\mathbf{x}$ is asymptotically almost periodic.
(iii) Let $u \in F, \delta > 0$. Then there is a neighborhood $W(u)$ of u in $C(K \times K)$ such that for every $w \in W(u)$ and for every (w)-good program $\mathbf{x}$ we have $dist(H(u), \Omega(\mathbf{x})) \leq \delta$.

Theorem 4. *Let $u \in F$, and let $\{x_i\}_{i=0}^{\infty}$ be a program such that*

$$\theta^u(x_i, x_{i+1}) = 0 \quad \forall i = 0, 1, \ldots$$

Then $\{x_i\}_{i=0}^{\infty}$ is a (u)-weakly optimal program. Moreover, there exists a strictly increasing sequence of natural numbers $\{i_k\}_{k=1}^{\infty}$ such that for every program $\{y_i\}_{i=0}^{\infty}$ satisfying $y_0 = x_0$ the inequality

$$\liminf_{k \to \infty} \sum_{j=0}^{i_k - 1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] \geq 0$$

holds, and if for some program $\{y_i\}_{i=0}^{\infty}$ *satisfying* $y_0 = x_0$,

$$\liminf_{k\to\infty} \sum_{j=0}^{i_k-1} [u(y_j, y_{j+1}) - u(x_j, x_{j+1})] = 0,$$

then $\theta^u(y_j, y_{j+1}) = 0$, $\forall j = 0, 1, \ldots$.

For every $u \in C(K \times K)$, every number $\Delta > 0$, and every integer $N \geq 1$ we denote by $A(u, N, \Delta)$ the set of all sequences $\{y_i\}_{i=0}^{N} \subset K$ such that for every sequence $\{z_i\}_{i=0}^{N} \subset K$ satisfying $z_0 = y_0, z_N = y_N$ the following inequality holds:

$$\sum_{i=0}^{N-1} [u(y_i, y_{i+1}) - u(z_i, z_{i+1})] \leq \Delta.$$

Recall the representation formula (2) and define $L\colon C(K \times K) \to R^1 \times C(K) \times C(K \times K)$ by $L(v) = (\mu(v), \pi^v, \theta^v)$, $v \in C(K \times K)$.

The second assertions of Theorems 5 and 6 establish a turnpike property for every $u \in F$ [16, 17, 21, 22, 25].

Theorem 5. *(i) L is continuous at every point of F.*
(ii) Let $u \in F$ *and* $\delta > 0$. *Then there are a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *and positive numbers* Q_1, Q_2 *such that for every* $w \in W(u)$, *for every integer* $N \geq 1$, *for every integer* $M > 0$ *and every program* $\{y_i\}_{i=0}^{N} \in A(w, N, M)$ *the following relation holds:*

$$Card\{i \in \{0, \ldots, N-1\}\colon d_1((y_i, y_{i+1}), H(u)) > \delta\} \leq Q_1 + MQ_2.$$

Theorem 6. *(i) Let* $u \in F$, $\epsilon > 0$. *Then there exist a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *and* $\delta > 0$ *such that for every* $w \in W(u)$ *and for every program* $\{x_i\}_{i=0}^{\infty}$ *satisfying* $\theta^w(x_i, x_{i+1}) = 0$, $i = 0, 1, \ldots$ *and* $d(x_0, H_0(u)) \leq \delta$, *the relation* $d_1((x_i, x_{i+1}), H(u)) \leq \epsilon$ *holds for* $i = 0, 1, \ldots$
(ii) Let $u \in F$, $\epsilon > 0$. *Then there exist a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *and an integer* $N \geq 1$ *such that for every* $w \in W(u)$ *and for every program* $\{x_i\}_{i=0}^{\infty}$ *satisfying* $\theta^w(x_i, x_{i+1}) = 0$, $i = 0, 1, \ldots$, *the inequality* $d_1((x_i, x_{i+1}), H(u)) \leq \epsilon$ *holds for every* $i \geq N$.

Corollary 1. *Let* $u \in F$, $\{x_i\}_{i=-\infty}^{\infty}$ *be a program such that* $\theta^u(x_i, x_{i+1}) = 0$, $i = 0, \pm 1, \ldots$. *Then* $(x_i, x_{i+1}) \in H(u)$, $i = 0, \pm 1, \ldots$

Corollary 2. *Let* $u \in F$, $\epsilon > 0$. *Then there exists a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *such that for every* $w \in W(u)$ *and for every program* $\{x_i\}_{i=-\infty}^{\infty}$ *satisfying* $\theta^w(x_i, x_{i+1}) = 0$, $i = 0, \pm 1, \ldots$ *the relation* $d_1((x_i, x_{i+1}), H(u)) \leq \epsilon$ *holds for every integer* i.

Theorem 7. *Let* $u \in F$. *Then every sequence* $\{y_i\}_{i=-\infty}^{\infty}$ *which satisfies* $\theta^u(y_i, y_{i+1}) = 0$, $i = 0, \pm 1, \ldots$ *is almost periodic. Moreover, for every* $\epsilon > 0$ *there exist a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *and an integer* $m \geq 1$

such that for every $w \in W(u)$ *and for every program* $\{y_i\}_{i=-\infty}^{\infty}$ *satisfying* $\theta^w(y_i, y_{i+1}) = 0$, $i = 0, \pm 1, \ldots$, *the relation* $d(y_i, y_{i+jm}) \leq \epsilon$ *holds for any integers* i *and* j.

Theorem 8. *(i) Let* $u \in F$ *and let* $\{x_i\}_{i=0}^{\infty}$ *be a* (u)*-good program. Then there exists a program* $\{y_i\}_{i=-\infty}^{\infty}$ *such that* $\theta^u(y_i, y_{i+1}) = 0$, $i = 0, \pm 1, \ldots$, *and* $\lim_{i \to \infty} d(x_i, y_i) = 0$.
(ii) Let $u \in F$ *and let* $\{x_i\}_{i=-\infty}^{\infty}$ *be a* (u)*-minimal energy configuration. Then there exist programs* $\{y_i\}_{i=-\infty}^{\infty}$, $\{z_i\}_{i=-\infty}^{\infty}$ *such that* $\theta^u(y_i, y_{i+1}) = 0$, $\theta^u(z_i, z_{i+1}) = 0$, $i = 0, \pm 1, \ldots$, $\lim_{i \to \infty} d(x_i, y_i) = 0$, $\lim_{i \to -\infty} d(x_i, z_i) = 0$.

We define

$$C_0(K \times K) = \{v \in C(K \times K):\ \mu(v) = \min\{v(x, x):\ x \in K\}\}.$$

It is easy to see that $C_0(K \times K)$ is a closed subspace of $C(K \times K)$. The space $C_0(K \times K)$ also has the topology of uniform convergence. In [29] we reinforced the previous theorems for $u \in C_0(K \times K)$ and proved the existence of a set $F_0 \subset F \cap C_0(K \times K)$ which is a countable intersection of open everywhere dense subsets of $C_0(K \times K)$ such that Theorem 9 holds for F_0. This result shows that for every $u \in F_0$ and every $x \in K$ there is a (u)-overtaking optimal program $\{x_i\}_{i=0}^{\infty}$ satisfying $x_0 = x$. It also establishes the strong turnpike theorem for $u \in F_0$.

Theorem 9. *(i)* $Card(H_0(u)) = 1$ *for* $u \in F_0$.
(ii) Let $u \in F_0$, $\delta > 0$. *Then there exists a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$ *such that for every* $w \in W(u)$, *and for every* (w)*-good program* $\mathbf{x}$ *the relation* $dist(\Omega(\mathbf{x}), (H_0(u) \times H_0(u))) \leq \delta$ *holds.*
(iii) Let $u \in F_0$, $\{x_i\}_{i=0}^{\infty}$ *be a program for which* $\theta^u(x_i, x_{i+1}) = 0$, $i = 0, 1, \ldots$. *Then* $\{x_i\}_{i=0}^{\infty}$ *is a* (u)*-overtaking optimal program and, moreover, if* $\{y_i\}_{i=0}^{\infty}$ *is a program such that* $y_0 = x_0$ *and*

$$\liminf_{N \to \infty} \sum_{i=0}^{N-1} [u(y_i, y_{i+1}) - u(x_i, x_{i+1})] = 0,$$

then $\theta^u(y_i, y_{i+1}) = 0$, $i = 0, 1, \ldots$
(iv) Let $u \in F_0$, $\epsilon > 0$. *Then there exist a neighborhood* $W(u)$ *of* u *in* $C(K \times K)$, *an integer* $Q \geq 1$ *and* $\epsilon_0 \in (0, \epsilon)$ *such that for every* $w \in W(u)$, *for every integer* $N \geq 2Q$ *and every program* $\{y_i\}_{i=0}^{N} \in A(w, N, \epsilon_0)$ *the following relation holds:* $d(y_i, H_0(u)) \leq \epsilon$, $i = Q, \ldots N - Q$ *and if* $d(y_0, H_0(u)) \leq \epsilon_0$, *then* $d(y_i, H_0(u)) \leq \epsilon$, $i = 0, \ldots N - Q$.

4 Turnpike Result for Convex Infinite Dimensional Discrete-Time Control Systems

In this section we discuss the structure of "approximate" solutions for an infinite dimensional discrete-time optimal control problem determined by a

convex function $v : K \times K \to R^1$, where K is a convex closed bounded subset of a Banach space. In [34] we showed that for a generic function v there exists $y_v \in K$ such that each "approximate" optimal solution $\{x_i\}_{i=0}^n \subset K$ is a contained in a small neighborhood of y_v for all $i \in \{N, \dots, n-N\}$, where N is a constant which depends on the neighborhood and does not depend on n. This result is a generalization of the main result of [33] which was established for convex uniformly continuous functions.

Let $(X, ||\cdot||)$ be a Banach space and let $K \subset X$ be a nonempty closed convex bounded set. Denote by $\mathcal{A}$ the set of all bounded convex functions $v : K \times K \to R^1$ which are continuous at a point (x, x) for any $x \in K$. Denote by $\mathcal{A}_l$ the set of all lower semicontinuous functions $v \in \mathcal{A}$, by $\mathcal{A}_c$ the set of all continuous functions $v \in \mathcal{A}$ and by $\mathcal{A}_u$ the set of all functions $v \in \mathcal{A}$ which satisfy the following uniform continuity assumption:

> $\forall \epsilon > 0$ there exists $\delta > 0$ such that for each $x_1, x_2, y_1, y_2 \in K$ satisfying $||x_i - y_i|| \le \delta$, $i = 1, 2$ the relation $|v(x_1, x_2) - v(y_1, y_2)| \le \epsilon$ holds.

We equip the space $\mathcal{A}$ with the metric

$$\rho(u, v) = \sup\{|v(x, y) - u(x, y)| : x, y \in K\}, \quad u, v \in \mathcal{A}.$$

Evidently the metric space $(\mathcal{A}, \rho)$ is complete and $\mathcal{A}_l$, $\mathcal{A}_c$ and $\mathcal{A}_u$ are closed subsets of $(\mathcal{A}, \rho)$. We equip the sets $\mathcal{A}_l$, $\mathcal{A}_c$ and $\mathcal{A}_u$ with the metric ρ.

In [33, 34] we investigated the structure of "approximate" solutions of the optimization problem

$$\text{minimize } \sum_{i=0}^{n-1} v(x_i, x_{i+1}) \text{ subject to } \{x_i\}_{i=0}^n \subset K, \ x_0 = y, \ x_n = z \tag{3}$$

for $v \in \mathcal{A}$, $y, z \in K$ and $n \ge 1$. In [33] we showed that for a generic function $v \in \mathcal{A}_u$ the following property holds:

> there is $y_v \in K$ such that for all large enough n and each $y, z \in K$ an "approximate" solution $\{x_i\}_{i=0}^n$ of (3) is contained in a small neighborhood of y_v for all $i \in \{N, \dots, n-N\}$ where N is a constant which depends on the neighborhood and does not depend on n.

This phenomenon which is called the turnpike property (TP) is well known in mathematical economics. The term was first coined by Samuelson [25] in 1958 where he showed that an efficient expanding economy would spend most of the time in the vicinity of a balanced equilibrium path (also called a von Neumann path). This property was further investigated in in mathematical economics (see [10, 11, 16-19, 21-24]) for optimal trajectories of models of economic dynamics. A related weak version of the turnpike property was considered in Section 3 with a nonconvex function $v : K \times K \to R^1$ and a compact metric space K.

When we say that a certain property holds for a generic element of a complete metric space Y we mean that the set of points which have this property contains a G_δ everywhere dense subset of Y. Such an approach, when a certain property is investigated for the whole space Y and not just for a single point in Y, has already been successfully applied in many areas of Analysis. In [34] we generalized the main result of [33] and showed that the turnpike property holds for a generic function $v \in \mathcal{A}$.

In almost all studies of discrete time control systems the turnpike property was considered for a single cost function v and a space of states K which was a compact convex set in a finite dimensional space. In these studies the compactness of K playes an important role. Specifically for the optimization problems considered in this section if a function v has the turnpike property then its "turnpike" y_v is a unique solution of the following optimization problem

$$\text{minimize } v(x,x) \text{ subject to } x \in K.$$

The existence of a solution of this problem is guaranteed only if K satisfies some compactness assumptions. To obtain the uniqueness of the solution we need additional assumptions on v such as its strict convexity.

Here, instead of considering the turnpike property for a single cost function v, we investigate it for spaces of all such functions equipped with some natural metric, and show that this property holds for most of these functions. In [33] we established the turnpike property without compactness assumption on the space of states for a generic convex uniform continuous cost function. In [34] we established the turnpike property for a generic convex cost function without this assumption.

For each $v \in \mathcal{A}$, integers $m_1, m_2 > m_1$ and $y_1, y_2 \in K$ we define

$$\sigma(v, m_1, m_2) = \inf\{\sum_{i=m_1}^{m_2-1} v(z_i, z_{i+1}) : \{z_i\}_{i=m_1}^{m_2} \subset K\},$$

$$\sigma(v, m_1, m_2, y_1, y_2) = \inf\{\sum_{i=m_1}^{m_2-1} v(z_i, z_{i+1}) : \{z_i\}_{i=m_1}^{m_2} \subset K,\ z_{m_1} = y_1, z_{m_2} = y_2\},$$

and the minimal growth rate

$$\mu(v) = \inf\{\liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(z_i, z_{i+1}) : \{z_i\}_{i=0}^{\infty} \subset K\}.$$

In [34, Prop. 2.1] we showed that $\mu(v) = \inf\{v(z,z) : z \in K\}$ for any $v \in \mathcal{A}$. In the same work we constructed a set $\mathcal{F}$ ($\mathcal{F}_l, \mathcal{F}_c, \mathcal{F}_u$, respectively) which is a countable intersection of open everywhere dense subsets of $\mathcal{A}$ ($\mathcal{A}_l, \mathcal{A}_c, \mathcal{A}_u$, respectively) and such that $\mathcal{F}_l \subset \mathcal{A}_l \cap \mathcal{F}$, $\mathcal{F}_c \subset \mathcal{A}_c \cap \mathcal{F}$, $\mathcal{F}_u \subset \mathcal{A}_u \cap \mathcal{F}$. We also established the following two theorems.

Theorem 10. *Let $v \in \mathcal{F}$. Then there exists a unique $y_v \in K$ such that $v(y_v, y_v) = \mu(v)$ and the following assertion holds:*

For each $\epsilon > 0$ there exist a neighborhood $\mathcal{U}$ of v in $\mathcal{A}$ and $\delta > 0$ such that for each $u \in \mathcal{U}$ and each $y \in K$ satisfying $u(y, y) \leq \mu(u) + \delta$ the relation $||y - y_v|| \leq \epsilon$ holds.

Theorem 11. *Let $v \in \mathcal{F}$ and $\epsilon > 0$. Then there exist $\delta \in (0, \epsilon)$, a neighborhood $\mathcal{U}$ of v in $\mathcal{A}$ and an integer $N \geq 1$ such that for each $u \in \mathcal{U}$, each integer $n \geq 2N$ and each sequence $\{x_i\}_{i=0}^n \subset K$ satisfying*

$$\sum_{i=0}^{n-1} u(x_i, x_{i+1}) \leq \sigma(u, 0, n, x_0, x_n) + \delta,$$

there exist $\tau_1 \in \{0, \dots, N\}$ and $\tau_2 \in \{n - N, \dots, n\}$ such that

$$||x_t - y_v|| \leq \epsilon, \quad t = \tau_1, \dots, \tau_2.$$

Moreover, if $||x_0 - y_v|| \leq \delta$, then $\tau_1 = 0$, and if $||y_v - x_n|| \leq \delta$, then $\tau_2 = n$.

5 Turnpike Result for Nonconvex Control Systems on Compact Metric Spaces

Let (K, d) be a compact metric space, $C(K \times K)$ the space of all continuous functions v: $K \times K \to R^1$ with the topology of the uniform convergence. Let $C(K)$ be the space of all continuous functions v: $K \to R^1$ with the topology of uniform convergence.

In this section we continue to discuss the structure of "approximate" solutions of the optimization problem (3) for $v \in C(K \times K)$, $y, z \in K$ and $n \geq 1$. Recall that a sequence $\{z_i\}_{i=0}^\infty \subset K$ is called (v)-good if the sequence (1) is bounded. Assume that $v \in C(K \times K)$ and there exists $x_v \in K$ such that each (v)-good sequence $\{x_i\}_{i=1}^\infty \subset K$ converges to x_v. In [35] we showed that the following turnpike property holds:

> for all large enough n and each $y, z \in K$ an "approximate" solution $\{x_i\}_{i=0}^n$ of (3) is contained in a small neighborhood of x_v for all $i \in \{N, \dots, n - N\}$ where N is a constant which depends on the neighborhood and does not depend on n.

Namely, in [35] we proved the following result.

Theorem 12. *Let $v \in C(K \times K)$. Assume that there exists $x_v \in K$ such that each (v)-good sequence $\{x_i\}_{i=1}^\infty \subset K$ converges to x_v. Let $\epsilon \in (0, 1)$. Then there exists $\delta \in (0, \epsilon)$, a neighborhood $\mathcal{U}$ of v in $C(K \times K)$ and an integer $N \geq 1$ such that for each $u \in \mathcal{U}$, each integer $n \geq 2N$ and each sequence $\{x_i\}_{i=0}^n \subset K$ satisfying*

$$\sum_{i=0}^{n-1} u(x_i, x_{i+1}) \le \sigma(u, 0, n, x_0, x_n) + \delta$$

there exist $\tau_1 \in \{0, \dots, N\}$ *and* $\tau_2 \in \{n - N, \dots, n\}$ *such that*

$$||x_t - x_v|| \le \epsilon, \ t = \tau_1, \dots, \tau_2.$$

Moreover, if $||x_0 - x_v|| \le \delta$, *then* $\tau_1 = 0$ *and if* $||x_n - x_v|| \le \delta$, *then* $\tau_2 = n$.

6 Minimal Solutions for Discrete-Time Control Systems in Metric Spaces

In this section we study the structure of minimal solutions for an autonomous discrete-time control system in a metric space X determined by a continuous function $v : X \times X \to R^1$. A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset X$ is called (v)-minimal if for each pair of integers $m_2 > m_1$ and each sequence $\{y_i\}_{i=m_1}^{m_2}$ satisfying $y_j = x_j$, $j = m_1, m_2$ the inequality

$$\sum_{i=m_1}^{m_2-1} v(x_i, x_{i+1}) \le \sum_{i=m_1}^{m_2-1} v(y_i, y_{i+1})$$

is valid. In [36] we considered a space of functions $v : X \times X \to R^1$ equipped with a natural complete metric and showed that for a generic function v there exists a (v)-minimal sequence.

Let (X, d) be a complete metric space. We equip the set $X \times X$ with the metric d_1 (defined in a similar way as in Section 3). Clearly the metric space $(X \times X, d_1)$ is complete. Denote by $\mathcal{A}$ the set of all continuous functions $v : X \times X \to R^1$ which satisfy the following two assumptions:

(i) (uniform boundedness) $\sup\{|v(x, y)| : \ x, y \in X\} < \infty$;

(ii) (uniform continuity) $\forall \epsilon > 0$, $\exists \delta > 0$ such that $|v(x_1, x_2) - v(y_1, y_2)| \le \epsilon$ for each $x_i, y_i \in X$, $i = 1, 2$ which satisfy $d(x_i, y_i) \le \delta$, $i = 1, 2$.

Define $\rho : \mathcal{A} \times \mathcal{A} \to R^1$ by $\rho(v, w) = \sup\{|v(x, y) - w(x, y)| : \ x, y \in X\}$. Clearly the metric space $(\mathcal{A}, \rho)$ is complete.

In this section we consider the optimization problem

$$\text{minimize} \sum_{i=k_1}^{k_2-1} v(x_i, x_{i+1}) \ \text{subject to} \ \{x_i\}_{i=k_1}^{k_2} \subset X, \ x_{k_1} = y, \ x_{k_2} = z \quad (4)$$

where $v \in \mathcal{A}$, $y, z \in X$ and $k_2 > k_1$ are integers.

Note that the problem (4) was considered in Sections 2 and 3 with a compact metric space X and in Section 4 when X was a bounded closed convex subset of a Banach space and the function v was convex.

If the space of states X is compact then the problem (4) has a solution for each $v \in \mathcal{A}$, $y, z \in X$ and each pair of integers $k_2 > k_1$. For the noncompact space X the existence of solutions of the problem (4) is not guaranteed.

For each $v \in \mathcal{A}$, each natural number m and each $y_1, y_2 \in X$ we set

$$||v|| = \sup\{|v(x,y)| : x, y \in X\},$$

$$\mu(v) = \inf\{\liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(x_i, x_{i+1}) : \{x_i\}_{i=0}^{\infty} \subset X\},$$

$$\sigma(v,m) = \inf\{\sum_{i=0}^{m-1} v(x_i, x_{i+1}) : \{x_i\}_{i=0}^{m} \subset X\},$$

$$\sigma_{per}(v,m) = \inf\{\sum_{i=0}^{m-1} v(x_i, x_{i+1}) : \{x_i\}_{i=0}^{m} \subset X,\ x_0 = x_m\},$$

$$\sigma(v,m,y_1,y_2) = \inf\{\sum_{i=0}^{m-1} v(x_i, x_{i+1}) : \{x_i\}_{i=0}^{m} \subset X,\ x_0 = y_1,\ x_m = y_2\}.$$

Let $v \in \mathcal{A}$. A sequence $\{x_i\}_{i=-\infty}^{\infty} \subset X$ is called (v)-minimal if for each pair of integers $m_2 > m_1$

$$\sum_{i=m_1}^{m_2-1} v(x_i, x_{i+1}) = \sigma(v, m_2 - m_1, x_{m_1}, x_{m_2}).$$

If the space of states X is compact then a (v)-minimal sequence can be constructed as a limit of a sequence of optimal solutions on finite intervals. For the noncompact space X the problem is more difficult and less understood. The difficulty is that for any problem of type (4) the existence of its solution is not guaranteed and that a (v)-minimal sequence is an exact solution of a countable number of optimization problems of type (4). In [36] we showed that for a generic function v taken from the space $\mathcal{A}$ there exists a (v)-minimal sequence.

A sequence $\{x_i\}_{i=0}^{\infty} \subset X$ is called (v)-good if there exists a number $M > 0$ such that for each natural number m

$$\sum_{i=0}^{m-1} v(x_i, x_{i+1}) \leq \sigma(v, m, x_0, x_m) + M.$$

It is not difficult to see that the following proposition holds.

Proposition 1. *Let $v \in \mathcal{A}$ and $\{z_i\}_{o=0}^{\infty} \subset X$ be a (v)-good sequence. Then for each $x \in X$ there is a (v)-good sequence $\{x_i\}_{i=0}^{\infty} \subset X$ such that $x_0 = x$.*

For each $\{x_i\}_{i=0}^{\infty} \subset X$ denote by $\omega(\{x_i\}_{i=0}^{\infty})$ the set of all $y \in X$ for which there exists a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} x_{i_k} = y$. In [36] we proved the following result.

Theorem 13. *There exists a set $\mathcal{F} \subset \mathcal{A}$ which is a countable intersection of open everywhere dense subsets of $\mathcal{A}$ such that for each $v \in \mathcal{F}$ there exists a nonempty compact set $\Omega(v) \subset X$ which satisfies the following two conditions:*

(i) there is a (v)-minimal sequence $\{x_i^{(v)}\}_{i=-\infty}^{\infty} \subset \Omega(v)$;

(ii) for each (v)-good sequence $\{y_i\}_{i=0}^{\infty} \subset X$ there exists a (v)-minimal sequence $\{x_i\}_{i=-\infty}^{\infty} \subset \Omega(v) \cap \omega(\{y_i\}_{i=0}^{\infty})$.

7 Turnpike Result for Control Systems on Metric Spaces

Let (X, ρ) be a metric space. Denote by $\mathcal{A}$ the set of all bounded functions $v : X \times X \to R^1$. Set $\Delta = X \times X$. We equip the set $\mathcal{A}$ with the metric

$$d(u, v) = \sup\{|v(x, y) - u(x, y)| : \ x, y \in K\}, \ u, v \in \mathcal{A}.$$

Evidenly $(\mathcal{A}, d)$ is a complete metric space. Denote by $\mathcal{A}_l$ the set of all lower semicontinuous functions $v \in \mathcal{A}$, by $\mathcal{A}_c$ the set of all continuous functions $v \in \mathcal{A}$ and by $\mathcal{A}_u$ the set of all uniformly continuous functions $v \in \mathcal{A}$. Clearly $\mathcal{A}_l, \mathcal{A}_c$ and $\mathcal{A}_u$ are closed subsets of the complete metric space $(\mathcal{A}, d)$.

Let $v \in \mathcal{A}$. Define a minimal growth rate $\mu(v)$ as in Section 6. Clearly

$$\mu(v) \leq \inf\{v(x, x) : \ x \in X\}.$$

Denote by $\mathcal{A}_*$ the set of all $v \in \mathcal{A}$ such that $\mu(v) = \inf\{v(x, x) : \ x \in X\}$. Clearly $\mathcal{A}_*$ is a closed subset of $(\mathcal{A}, d)$. Set

$$\mathcal{A}_{*l} = \mathcal{A}_* \cap \mathcal{A}_l, \ \mathcal{A}_{*c} = \mathcal{A}_* \cap \mathcal{A}_c, \ \mathcal{A}_{*u} = \mathcal{A}_* \cap \mathcal{A}_u.$$

Clearly $\mathcal{A}_* \neq \emptyset$. For example, if $v(x, y) = c$ for all $(x, y) \in X \times X$ where c is a constant, then $v \in \mathcal{A}_{*c}$.

The following proposition will be proved in Section 8.

Proposition 2. *Let $v \in \mathcal{A}_*$. Then for each $x \in X$ there is a sequence $\{x_i\}_{i=0}^{\infty} \subset X$ such that $x_0 = x$ and*

$$\limsup_{N \to \infty} N^{-1} \sum_{t=0}^{N-1} v(x_i, x_{i+1}) = \mu(v).$$

For each $v \in \mathcal{A}$ set $||v|| = \sup\{|v(x, y)| : \ x, y \in X\}$. For $x \in X$ and $B \subset X$ set $\rho(x, B) = \inf\{\rho(x, y) : \ y \in B\}$. Denote by $\mathcal{F}$ the set of all $v \in \mathcal{A}_*$ which have the following property:

(P) for each $\epsilon > 0$ there exist $\delta > 0$ and a neighborhood $\mathcal{V}$ of v in $\mathcal{A}$ such that for each $u \in \mathcal{V}$ and each sequence $\{x_i\}_{i=0}^{\infty} \subset X$ which satisfies

$$\limsup_{N \to \infty} N^{-1} \sum_{i=0}^{N-1} u(x_i, x_{i+1}) \leq \mu(u) + \delta, \tag{5}$$

one has $\limsup_{N\to\infty} N^{-1}\mathrm{Card}\{i \in \{0,\dots,N-1\} : \rho(x_i, x_{i+1}) \geq \epsilon\} \leq \epsilon$.

If $v \in \mathcal{A}_*$ has the property (P) then all good programs spend most of time in a small neighborhood of the diagonal Δ.

We show that most elements of $\mathcal{A}_*$ (in the sense of Baire's categories) have the property (P). Moreover, we show that the complement of the set of all functions which have the property (P) is not only of the first category, but also σ-porous.

Before we continue we recall the concept of porosity [8, 9]. Let (Y, ρ) be a complete metric space. We denote by $B(y, r)$ the closed ball of center $y \in Y$ and radius $r > 0$. A subset $E \subset Y$ is called porous in (Y, d) if there exist $\alpha \in (0, 1)$ and $r_0 > 0$ such that for each $r \in (0, r_0]$ and each $y \in Y$ there exists $z \in Y$ for which $B(z, \alpha r) \subset B(y, r) \setminus E$. A subset of the space Y is called σ-porous in (Y, ρ) if it is a countable union of porous subsets in (Y, ρ).

Since porous sets are nowhere dense, all σ-porous sets are of the first category. If Y is a finite-dimensional Euclidean space, then σ-porous sets are of Lebesgue measure 0. In fact, the class of σ-porous sets in such a space is much smaller than the class of sets which have measure 0 and are of the first category.

In this paper we will establish the following result.

Theorem 14. *The set $\mathcal{A}_* \setminus \mathcal{F}$ ($\mathcal{A}_{*l} \setminus \mathcal{F}$, $\mathcal{A}_{*c} \setminus \mathcal{F}$, $\mathcal{A}_{*u} \setminus \mathcal{F}$, respectively) is a σ-porous subset of $\mathcal{A}_*$ ($\mathcal{A}_{*l}$, $\mathcal{A}_{*c}, \mathcal{A}_{*u}$, respectively).*

8 Proof of Proposition 2

Let $x \in X$. For each natural number n there is $z_n \in X$ such that

$$v(z_n, z_n) \leq \mu(v) + 2^{-n}. \tag{6}$$

Define a sequence $\{x_n\}_{n=0}^{\infty} \subset X$ as follows:

$$x_0, x_1 = x,\ x_n = z_k,\ n = 2^k, \dots, 2^{k+1} - 1,\ k = 1, 2, \dots \tag{7}$$

We show that

$$\limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(x_i, x_{i+1}) = \mu(v).$$

Clearly

$$\limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(x_i, x_{i+1}) \geq \mu(v).$$

Let $N \geq 9$ be a natural number. There is a natural number $k = k(N)$ such that $2^k \leq N < 2^{k+1}$. Set $p = N - 2^k$. Then by (7) and (6)

$$\sum_{i=0}^{N-1} v(x_i, x_{i+1}) = v(x, x) + v(x, z_1) + v(z_1, z_1)$$

$$+\sum_{j=2}^{k-1}\left[v(z_{j-1},z_j)+(2^j-1)v(z_j,z_j)\right]+(v(z_{k-1},z_k)+pv(z_k,z_k))$$

$$\le 3||v||+\sum_{j=2}^{k-1}\left[2^j v(z_j,z_j)+2||v||\right]+\left[(p+1)v(z_k,z_k)+2||v||\right]$$

$$\le \sum_{j=2}^{k-1}2^j v(z_j,z_j)+(p+1)v(z_k,z_k)+2||v||(k+1)$$

$$\le \sum_{j=2}^{k-1}2^j(\mu(v)+2^{-j})+(p+1)[\mu(v)+2^{-k}]+2||v||(k+1)$$

$$=\mu(v)\Big[\sum_{j=2}^{k-1}2^j+(p+1)\Big]+k+2||v||(k+1)$$

$$\le \mu(v)\left[2^k-1-3+(p+1)\right]+(k+1)(2||v||+1)$$

$$\le \mu(v)(N-3)+(\log_2 N+1)(2||v||+1).$$

This relation implies that

$$\limsup_{N\to\infty} N^{-1}\sum_{i=0}^{N-1}v(x_i,x_{i+1})$$

$$\le \limsup_{N\to\infty}[N^{-1}\mu(v)(N-3)+(2||v||+1)(\log_2 N+1)N^{-1}]=\mu(v).$$

Proposition 2 is proved. □

9 An Auxiliary Result for Theorem 14

Let $v\in\mathcal{A}_*$, $\gamma\in(0,1]$. Define

$$v_\gamma(x,y)=v(x,y)+\gamma\min\{1,\rho(x,y)\},\ x,y\in X. \tag{8}$$

Clearly

$$v_\gamma\in\mathcal{A}_*,\ \mu(v_\gamma)=\mu(v), \tag{9}$$

if $v\in\mathcal{A}_{*l}$ ($\mathcal{A}_{*c}$, $\mathcal{A}_{*u}$ respectively), then $v_\gamma\in\mathcal{A}_{*l}$ ($\mathcal{A}_{*c}$, $\mathcal{A}_{*u}$ respectively).

Lemma 1. *Let $\delta>0$, $u\in\mathcal{A}_*$ satisfy $d(u,v_\gamma)\le\delta$, and let $\{x_i\}_{i=0}^{\infty}\subset X$ satisfy*

$$\limsup_{N\to\infty}N^{-1}\sum_{i=0}^{N-1}u(x_i,x_{i+1})\le\mu(u)+\delta. \tag{10}$$

Then, for each $\epsilon\in(0,1]$,

$$\limsup_{N\to\infty}N^{-1}\,Card\{i\in\{0,\dots,N-1\}:\rho(x_i,x_{i+1})\ge\epsilon\}\le(3\delta)(\gamma\epsilon)^{-1}.$$

Proof. Let $\epsilon \in (0,1]$. Due to $d(u, v_\gamma) \le \delta$, one has

$$|\limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} u(x_i, x_{i+1}) - \limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v_\gamma(x_i, x_{i+1})| \le \delta, \tag{11}$$

$$|\mu(u) - \mu(v_\gamma)| \le \delta. \tag{12}$$

In view of (9)-(12),

$$\limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v_\gamma(x_i, x_{i+1}) \le \mu(v_\gamma) + 3\delta = \mu(v) + 3\delta. \tag{13}$$

It follows from (13), (8) and the definition of $\mu(v)$ that

$$\mu(v) + 3\delta \ge \limsup_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} \left[v(x_i, x_{i+1}) + \gamma \min\{1, \rho(x_i, x_{i+1})\}\right]$$

$$\ge \limsup_{N\to\infty} N^{-1}[\sum_{i=0}^{N-1} v(x_i, x_{i+1}) + \gamma\epsilon \mathrm{Card}\{i \in \{0, \dots, N-1\} : \rho(x_i, x_{i+1}) \ge \epsilon\}]$$

$$\ge \limsup_{N\to\infty} N^{-1}\gamma\epsilon \mathrm{Card}\{i \in \{0, \dots, N-1\} : \rho(x_i, x_{i+1}) \ge \epsilon\}$$

$$+ \liminf_{N\to\infty} N^{-1} \sum_{i=0}^{N-1} v(x_i, x_{i+1})$$

$$\ge \limsup_{N\to\infty} \gamma\epsilon N^{-1} \mathrm{Card}\{i \in \{0, \dots, N-1\} : \rho(x_i, x_{i+1}) \ge \epsilon\} + \mu(v).$$

This inequality implies that

$$\limsup_{N\to\infty} N^{-1} \mathrm{Card}\{i \in \{0, \dots, N-1\} : \rho(x_i, x_{i+1}) \ge \epsilon\} \le (3\delta)(\gamma\epsilon)^{-1}.$$

Lemma 1 is proved. □

10 Proof of Theorem 14

For each natural number n denote by $\mathcal{F}_n$ the set of all $v \in \mathcal{A}_*$ which have the following property:

(P1) there exist $\delta > 0$ and a neighborhood $\mathcal{V}$ of v in $\mathcal{A}_*$ such that for each $u \in \mathcal{V}$ and each sequence $\{x_i\}_{i=0}^{\infty} \subset X$ which satisfies (5), one has

$$\limsup_{N\to\infty} N^{-1} \mathrm{Card}\{i \in \{0, \dots, N-1\} : \rho(x_i, x_{i+1}) \ge n^{-1}\} \le 1/n.$$

It is not difficult to see that $\mathcal{F} = \cap_{n=1}^{\infty} \mathcal{F}_n$. In order to prove Theorem 14 it is sufficient to show that for each integer $n \geq 1$, $\mathcal{A}_* \setminus \mathcal{F}_n$ ($\mathcal{A}_{*l} \setminus \mathcal{F}_n$, $\mathcal{A}_{*c} \setminus \mathcal{F}_n$, $\mathcal{A}_{*u} \setminus \mathcal{F}_n$, respectively) is a porous subset of $\mathcal{A}_*$ ($\mathcal{A}_{*l}$, $\mathcal{A}_{*c}$, $\mathcal{A}_{*u}$, respectively).

Let n be a natural number. Set

$$\alpha = (32n^2)^{-1}. \tag{14}$$

Assume that $v \in \mathcal{A}_*$, $r \in (0, 1]$. Put

$$\gamma = 6\alpha r n^2 < r/4 \tag{15}$$

and define

$$v_\gamma(x, y) = v(x, y) + \gamma \min\{1, \rho(x, y)\}, \ x, y \in X. \tag{16}$$

Clearly $v_\gamma \in \mathcal{A}_*$, $\mu(v_\gamma) = \mu(v)$, and if $v \in \mathcal{A}_{*l}$ ($\mathcal{A}_{*c}$, $\mathcal{A}_{*u}$ respectively), then $v_\gamma \in \mathcal{A}_{*l}$ ($\mathcal{A}_{*c}$, $\mathcal{A}_{*u}$ respectively). By (15)-(16),

$$d(v, v_\gamma) \leq \gamma \leq r/4. \tag{17}$$

Assume that

$$u \in \mathcal{A}_*, \ d(u, v_\gamma) \leq 2\alpha r, \tag{18}$$

$$\{x_i\}_{i=0}^{\infty} \subset X, \ \limsup_{N \to \infty} N^{-1} \sum_{i=0}^{N-1} u(x_i, x_{i+1}) \leq \mu(u) + 2\alpha r. \tag{19}$$

In view of (17)-(18) and (14),

$$d(u, v) \leq r. \tag{20}$$

It follows from (18)-(19), Lemma 1 (with $\delta = 2\alpha r$, $\epsilon = 1/n$) and (15) that

$$\limsup_{N \to \infty} N^{-1} \mathrm{Card}\{i \in \{0, \ldots, N-1\} : \ \rho(x_i, x_{i+1}) \geq 1/n\}$$

$$\leq (6\alpha r)(\gamma/n)^{-1} = 6\alpha r \gamma^{-1} n \leq 1/n.$$

We have shown that each $u \in \mathcal{A}_*$ satisfying $d(u, v_\gamma) \leq \alpha r$ belongs to $\mathcal{F}_n$ and satisfies (20). This completes the proof of Theorem 14. □

References

1. B.D.O. Anderson and J.B. Moore. Linear Optimal Control. Prentice-Hall, Englewood Cliffs, NJ, 1971.
2. H. Atsumi. Neoclassical growth and the efficient program of capital accumulation. Review of Econ. Studies, 32: 127-136, 1965.
3. S. Aubry and P.Y. Le Daeron. The discrete Frenkel-Kontorova model and its extensions. Physica D, 8: 381-422, 1983.
4. J. Blot and P. Cartigny. Optimality in infinite-horizon variational problems under sign conditions. J. Optim. Theory Appl., 106: 411-419, 2000.

5. J. Blot and P. Michel. The value-function of an infinite-horizon linear quadratic problem. Appl. Math. Lett., 16: 71-78, 2003.
6. W.A. Brock. On existence of weakly maximal programs in a multi-sector economy. Review of Econ. Studies, 37: 275-280, 1970.
7. B.D. Coleman, M. Marcus and V.J. Mizel. On the thermodynamics of periodic phases. Arch. Rational Mech. Anal., 117: 321-347, 1992.
8. F.S. De Blasi and J. Myjak. Sur la porosité des contractions sans point fixe. C. R. Acad. Sci. Paris, 308: 51-54, 1989.
9. F. S. De Blasi, J. Myjak and P. L. Papini. Porous sets in best approximation theory. J. London Math. Soc., 44: 135-142, 1991.
10. Z. Dzalilov, A.F. Ivanov and A.M. Rubinov. Difference inclusions with delay of economic growth. Dynam. Systems Appl., 10: 283-293, 2001.
11. Z. Dzalilov, A.M. Rubinov and P.E. Kloeden. Lyapunov sequences and a turnpike theorem without convexity. Set-Valued Analysis, 6: 277-302, 1998.
12. D. Gale. On optimal development in a multisector economy. Rev. of Econ. Studies, 34: 1-19, 1967.
13. M. Ali Khan and T. Mitra. On choice of technique in the Robinson-Solow-Srinivasan model. Int. J. Econ. Theory, accepted.
14. A. Leizarowitz. Infinite horizon autonomous systems with unbounded cost. Appl. Math. Optim., 13: 19-43, 1985.
15. A. Leizarowitz and V.J. Mizel. One dimensional infinite horizon variational problems arising in continuum mechanics. Arch. Rational Mech. Anal. , 106: 161-194, 1989.
16. V.L. Makarov, M.J. Levin and A.M. Rubinov. Mathematical Economic Theory: Pure and Mixed Types of Economic Mechanisms. North-Holland, Amsterdam, 1995.
17. V.L. Makarov and A.M. Rubinov. Mathematical Theory of Economic Dynamics and Equilibria. Nauka, Moscow, 1973, English trans. Springer-Verlag, New York, 1977.
18. M.A. Mamedov and S. Pehlivan. Statistical convergence of optimal paths. Math. Japon., 52: 51-55, 2000.
19. M.A. Mamedov and S. Pehlivan. Statistical cluster points and turnpike theorem in nonconvex problems. J. Math. Anal. Appl., 256: 686-693, 2001.
20. M. Marcus and A.J. Zaslavski. The structure of extremals of a class of second order variational problems. Ann. Inst. H. Poincaré, Anal. non linéaire, 16: 593-629, 1999.
21. L.W. McKenzie. Turnpike theory. Econometrica 44: 841-866, 1976.
22. R. Radner. Path of economic growth that are optimal with regard only to final states; a turnpike theorem. Rev. Econom. Stud., 28: 1961, 98-104, 1961.
23. A.M. Rubinov. Superlinear Multivalued Mappings and their Applications in Economic Mathematical Problems. Nauka, Leningrad, 1980.
24. A.M. Rubinov. Economic dynamics. J. Soviet Math., 26: 1975-2012, 1984.
25. P.A. Samuelson. A catenary turnpike theorem involving consumption and the golden rule. Amer. Econ. Review, 55: 486-496, 1965.
26. C.C. von Weizsacker. Existence of optimal programs of accumulation for an infinite horizon. Rev. Econ. Studies, 32: 85-104, 1965.
27. A.J. Zaslavski. Ground states in Frenkel-Kontorova model. Math. USSR Izvestiya, 29: 323-354, 1987.
28. A.J. Zaslavski. Optimal programs on infinite horizon 1. SIAM J. Control Optim., 33: 1643-1660, 1995.

29. A.J. Zaslavski. Optimal programs on infinite horizon 2. SIAM J. Control Optim., 33: 1661-1686, 1995.
30. A.J. Zaslavski. Dynamic properties of optimal solutions of variational problems. Nonlinear Analysis, 27: 1996, 895-932, 1996.
31. A.J. Zaslavski. Structure of extremals for one-dimensional variational problems arising in continuum mechanics. J. Math. Anal. Appl., 198: 893-921, 1996.
32. A.J. Zaslavski. Existence and structure of optimal solutions of variational problems. Proc. Special Session on Optimization and Nonlinear Analysis, Joint AMS-IMU Conference, Jerusalem, May 1995, Contemporary Mathematics, 204: 247-278, 1997.
33. A.J. Zaslavski. Turnpike theorem for convex infinite dimensional discrete-time control systems. J. Convex Analysis, 5: 237-248, 1998.
34. A.J. Zaslavski. Turnpike property for infinite dimensional convex discrete-time control systems in a Banach space. Int. J. Pure and Applied Math., 7: 295-309, 2003.
35. A.J. Zaslavski. Turnpike theorem for a class of discrete time optimal control problems. Proc. 2001 of the 7th Int. Conf. on Nonlinear Functional Analysis and Applications. Fixed Point Theory and Applications, Nova Science Publishers, Inc., New York, 5: 175-182, 2003.
36. A.J. Zaslavski. Minimal solutions for discrete-time control systems in metric spaces. Numerical Func. Anal. Optim., 24: 637-651, 2003.
37. A.J. Zaslavski and A. Leizarowitz. Optimal solutions of linear control systems with nonperiodic integrands. Math. Oper. Res., 22: 726-746, 1997.
38. A.J. Zaslavski and A. Leizarowitz. Optimal solutions of linear periodic control systems with convex integrands. Appl. Math. Optim., 37: 127-150, 1998.

Numerical Methods for Optimal Control with Binary Control Functions Applied to a Lotka-Volterra Type Fishing Problem*

Sebastian Sager[1,2], Hans Georg Bock[1], Moritz Diehl[1], Gerhard Reinelt[1], and Johannes P. Schlöder[1]

[1] IWR Heidelberg, Germany
[2] sebastian.sager@iwr.uni-heidelberg.de

Summary. We investigate possibilities to deal with optimal control problems that have special integer restrictions on the time dependent control functions, namely to take only the values of 0 or 1 on given time intervals. A heuristic penalty term homotopy and a Branch and Bound approach are presented, both in the context of the direct multiple shooting method for optimal control. A tutorial example from population dynamics is introduced as a benchmark problem for optimal control with $0-1$ controls and used to compare the numerical results of the different approaches.

1 Introduction

Optimal control problems have long been under investigation and it is well known that for certain systems, in particular linear ones, bang-bang controls are optimal. On the other hand it is not clear what to do if the feasible set of a control is a priori restricted to two (or more) discrete values only and the optimal switching structure cannot be guessed due to the complexity of the model under consideration.

Optimal control problems with the mentioned restriction to 0-1 values in the controls arise whenever a yes-no decision has to be made, as is e.g. the case for certain types of valves or pumps in engineering, certain investments in economics, discrete stages in transport or application of laws in given time periods. Such problems are typically nonlinear and already difficult to solve without combinatorial apects.

Although some mixed integer dynamic optimisation problems, namely the optimisation of New York subway trains that are equipped with discrete acceleration stages, were solved in the early eighties [3], the so-called indirect methods used there do not seem appropriate for generic large-scale optimal control

* Work supported by the Deutsche Forschungsgemeinschaft (DFG) within the graduate program Complex Processes: Modeling, Simulation and Optimization.

problems with underlying nonlinear differential algebraic equation (DAE) systems. Therefore efforts have to be undertaken to bring together methodology of and new results for indirect methods in this context (see e.g. [24]) and the so-called direct methods, particularly the direct multiple shooting method [4] and direct collocation [23, 29].

Several authors have been working on optimal control problems with discrete valued control functions: [7] investigate a water distribution network in Berlin with such on/off pumps, using a problem specific nonlinear continuous reformulation of the control functions; [28] treat powertrain control of heavy duty trucks with a tailored heuristics in the context of direct multiple shooting that fits into the model predictive control context; [15, 22] use a switching time approach related to the one described in Section 3.3 to deal with problems where only a finite set of controls, e.g. velocities of submarine vessels, is feasible; [8] focus on problems in robotics, applying a combination of Branch and Bound and direct collocation [26].

Other publications in the field of mixed integer dynamic optimisation deal with time independent integer variables (e.g. [20]) or state dependent (autonomous) switches (e.g. [6]) that are both not the topic of this paper.

The paper is organised as follows. In Section 2 a short introduction to numerical methods for optimal control is given, in particular to the direct multiple shooting method [4]. In Section 3 extensions to treat additional integer restrictions are presented. An optimal control problem with a 0-1 restriction in the controls is presented in Section 4 and is used as a benchmark problem further on. Numerical results are given and compared in Section 5. Section 6 concludes.

2 Numerical Methods for Optimal Control

The optimal control problems we refer to in this section and that are later on to be extended, are of the form

$$\begin{aligned}
\min_{p,u,x,z} \; & \int_{t_0}^{T} L(x(t), z(t), u(t), p)dt \; + E(x(T), z(T), p) \\
\text{s.t.} \quad & \dot{x}(t) = f(t, x(t), z(t), u(t), p), \quad t \in [t_0, T] \\
& 0 = g(t, x(t), z(t), u(t), p), \quad t \in [t_0, T] \\
& 0 \leq c(t, x(t), z(t), u(t), p), \quad t \in [t_0, T] \\
& 0 \leq r_i(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p) \\
& 0 = r_e(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p)
\end{aligned} \tag{1}$$

The system state is described by the differential and algebraic state vectors $x(t) \in \mathbb{R}^{n_x}$ and $z(t) \in \mathbb{R}^{n_z}$. The system behaviour is controlled by the control vectors $u(t) \in \mathbb{R}^{n_u}$ and the global design parameter vector $p \in \mathbb{R}^{n_p}$. The objective functional is of generalised Bolza type, containing Lagrange and Mayer terms. The differential and algebraic right hand side functions f respectively

g describe the dynamical system behavior, while the vector valued functions r_i, r_e are additional interior point constraints for given time points t_i (see Section 2.1) and c contains path constraints. The Jacobian $\partial g/\partial z \in \mathbb{R}^{n_z \times n_z}$ is assumed to be invertible, resulting in an index 1 DAE.

There are several approaches to treat optimal control problems of this form. For an overview and comparison between indirect and direct methods, sequential and simultaneous approaches (in particular single shooting, multiple shooting and collocation) we refer to [1]. We investigate extensions in the context of the direct multiple shooting method, therefore we will give a brief introduction in Section 2.1.

2.1 Direct Multiple Shooting

The direct multiple shooting method [4, 17] is used to transform the infinite dimensional optimisation problem (1) into a finite dimensional one that can be treated efficiently with tailored nonlinear optimisation methods, e.g. sequential quadratic programming (SQP). This transformation is performed by a piecewise parameterisation of the control functions, a relaxation of the path constraints to grid points and a discretisation of the state variables. To this end the time horizon $[t_0, T]$ is divided into a number of m subintervals $[t_i, t_{i+1}]$ with $t_0 < t_1 < \cdots < t_m = T$, the so-called multiple shooting intervals.

Parameterisation of the controls. For each interval the function space that the optimal control function $u(t)$ can be chosen from is reduced to a finite dimensional one. Then a piecewise approximation $\hat{u}$ of the control functions u on this grid is defined by

$$\hat{u}(t) := \varphi_i(t, q_i), \quad t \in [t_i, t_{i+1}], \quad i = 0, \ldots, m-1 \tag{2}$$

using "local" control parameters q_i. The functions φ_i are typically vectors of constant, linear or cubic functions.

State discretisation. The basic concept of the multiple shooting method is to solve the DAE-constraints independently on each of the multiple shooting intervals. On interval i the initial value for the DAE solution is given by the so-called *node values* s_i^x, s_i^z for differential and algebraic states. The algebraic equations are relaxed (see [2], [16]). They enter as conditions in t_i into the NLP. Continuity of the state trajectory at the multiple shooting grid points

$$s_{i+1}^x = x_i(t_{i+1}) \tag{3}$$

is also incorporated by constraints into the nonlinear program (NLP). Here $x_i(t)$ denotes the differential part of the DAE solution on interval $[t_i, t_{i+1}]$ with initial values s_i^x, s_i^z. These equations are required to be satisfied only at the solution of the problem, not necessarily during the SQP iterations.

Resulting NLP. The local variables q_i, the global parameters p, that may include the time horizon length $h = T - t_0$, and the node values s_i^x, s_i^z are

the degrees of freedom of the parameterised optimal control problem. If we write them in one vector $\xi = (q_i, p, s_i^x, s_i^z)$, rewrite the objective function as $F(\xi)$, subsume all equality constraints with the continuity conditions (3) into a function $G(\xi)$ and all inequality constraints into a function $H(\xi)$, then the resulting NLP can be written as

$$\min_{\xi} F(\xi) \quad \text{s.t.} \quad 0 = G(\xi), \;\; 0 \leq H(\xi) \tag{4}$$

This NLP can be solved with tailored iterative methods, exploiting the structure of the problem. For more details, see [4, 16, 17]. An efficient implementation of the described method is the software package `MUSCOD-II` [9].

3 Treatment of Binary Control Functions

We are interested in an extension of problem (1), where some or all of the control functions have the additional restriction to have values in $\{0, 1\}$. If we denote these control functions by $w(t)$ we can formulate an optimal control problem with binary valued control functions. We want to minimise the functional

$$\Phi[x, z, w, u, p] := \int_{t_0}^{T} L(x(t), z(t), w(t), u(t), p)dt \; + E(x(T), z(T), p) \tag{5}$$

subject to a system of DAEs, path and interior point constraints and additional restrictions

$$w(t) \in \{0, 1\}^{n_w}, \quad t \in [t_0, T] \tag{6}$$

that turn the problem into a combinatorial one.

For some applications restriction (6) is still too general. A certain limitation on the number of switchings must be taken into consideration, as an infinite switching from one value to the other is not applicable in practice. This might be achieved by an upper limit on the number of switches or a penalisation. In the direct multiple shooting approach a fixed finite control parameterisation inhibts infinite switching automatically.

Another possible limitation occurs when switching can only take place at time points from a prefixed given set. This limitation is motivated by machines that can only switch in discrete time steps and by laws or investments that can only be applied resp. made at certain times, e.g. on the first of a month or year. Thus we replace restriction (6) by the more general restriction

$$w(t) \in \Omega(\Psi), \quad t \in [t_0, T] \tag{7}$$

where $\Omega(\Psi)$ is defined as

$$\Omega(\Psi) := \left\{ w(t) \in \{0, 1\}^{n_w}, \text{ with discontinuities only at times } \hat{t}_i \in \Psi \right\}$$

with either

$$\Psi = \{\tau_1, \tau_2, \ldots, \tau_{n_\tau}\} \tag{8}$$

being a finite set of possible switching times or with

$$\Psi = [t_0, T] \tag{9}$$

corresponding to (6). If we write $\bar{\Omega}(\Psi)$, we mean the relaxed function space where $\{0,1\}^{n_w}$ is replaced by $[0,1]^{n_w}$. Summing up, the optimal control problems under consideration can be formulated in the following way:

$$\begin{aligned}
\min_{x,z,w,u,p} \; & \int_{t_0}^{T} L(x(t), z(t), w(t), u(t), p)dt \; + E(x(T), z(T), p) \\
\text{s.t.} \quad & \dot{x}(t) = f(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\
& 0 = g(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\
& 0 \leq c(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\
& 0 \leq r_i(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p) \\
& 0 = r_e(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p) \\
& w(t) \in \Omega(\Psi), \quad t \in [t_0, T]
\end{aligned} \tag{10}$$

In the following we will choose the control parameterisation intervals $[t_i, t_{i+1}]$ such that they coincide with the intervals $[\tau_i, \tau_{i+1}]$. More precisely, we choose $m = n_\tau$ and $t_i = \tau_i, i = 1, \ldots, m$. Furthermore we will use a control parameterisation (2) that is constant on these intervals.

We want to investigate possibilities to solve problem (10). In Section 3.1 we have a look at relaxations of the integer constraints and in Section 3.2 we describe a Branch and Bound algorithm for mixed integer dynamic optimisation problems. In Section 3.3 a reformulation based on optimisation of the continuous switching times is given and discussed.

3.1 Heuristics Based on Relaxation

A first approach to solve problem (10) consists of relaxing the integer requirement $w(t) \in \Omega(\Psi)$ to $\bar{w}(t) \in \bar{\Omega}(\Psi)$ and to solve a relaxed problem of form (1) first. The obtained solution $\bar{w}(t)$ can then be investigated – in the best case it is an integer feasible bang-bang solution and we have found an optimal solution for the integer problem. In case the relaxed solution is not integer, one of the following rounding strategies can be applied:

- Rounding strategy 1
  The values $w_{j,i}(t)$ of the control functions $j = 1, \ldots, n_w$ on the intervals $[t_i, t_{i+1}]$ are fixed to

$$w_{j,i}(t) = \begin{cases} 1 & \text{if } \bar{w}_{j,i}(t) \geq 0.5 \\ 0 & \text{else} \end{cases}, \quad i = 0, \ldots, m-1$$

- Rounding strategy 2
  The values of $\bar{w}_{j,i}(t)$ are summed up over the intervals, more precisely

$$w_{j,i}(t) = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \bar{w}_{j,k}(t) - \sum_{k=0}^{i-1} w_{j,k}(t) \geq 1 \\ 0 & \text{else} \end{cases}, \quad i = 0, \ldots, m-1$$

- Rounding strategy 3
  As 2, but with a different threshold:

$$w_{j,i}(t) = \begin{cases} 1 & \text{if } \sum_{k=0}^{i} \bar{w}_{j,k}(t) - \sum_{k=0}^{i-1} w_{j,k}(t) \geq 0.5 \\ 0 & \text{else} \end{cases}, \quad i = 0, \ldots, m-1$$

In case the relaxed solution is not integer and the gap between the objective values of relaxed and rounded problems is important, we propose the following approach to drive the values of the control function to its borders.

- Penalty term homotopy
  We consider an optimal control problem $P^k, k \in \mathbb{N}_0$ defined by

$$\begin{aligned} \min_{x,z,w,u,p} \quad & \Phi[x,z,w,u,p] + \sum_{i=1}^{n_w} \epsilon_i^k \int_{t_0}^{T} (1 - w_i(t))\, w_i(t)\, dt \\ \text{s.t.} \quad & \dot{x}(t) = f(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\ & 0 = g(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\ & 0 \leq c(t, x(t), z(t), w(t), u(t), p), \quad t \in [t_0, T] \\ & 0 \leq r_i(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p) \\ & 0 = r_e(x(t_0), z(t_0), x(t_1), z(t_1), \ldots, x(T), z(T), p) \\ & w(t) \in \bar{\Omega}(\Psi), \quad t \in [t_0, T] \end{aligned} \tag{11}$$

  with penalty parameters $\epsilon_i^k \geq 0$ for $i = 1, \ldots, n_w$. P^k is similar to the relaxed version of problem (10), but additionally penalises all measurable violations of the integer requirements with a concave quadratic penalty term. The proposed penalty term homotopy consists of solving a series of continuous optimal control problems $\{P^k\}, k \in \mathbb{N}_0$ with relaxed $w(t)$. Problem P^{k+1} is initialised with the solution of P^k and $\epsilon_i^0 = 0$ so that P^0 is the relaxed version of problem (10). The penalty parameters ϵ_i^k are raised monotonically until all $w_i(t)$ are 0 or 1.

Remark 1. The algorithm may of course get stuck if the solution is driven towards an infeasible solution. This can e.g. be observed by a technique controlling the changes in the optimisation variables from one problem to the next. In such a situation several remedies are possible, e.g. a complete restart with different initial data, backtracking with a different choice of ϵ_i^k, another penalisation to get away from the current point or a transformation of the problem with an approach as described in Section 3.3.

Remark 2. A good choice for the ϵ_i^k is crucial for the behaviour of the method. A too fast increase in the penalty parameters results in less accuracy and is

getting closer to simple rounding, while a slow increase leads to an augmentation in the number of QPs that have to be solved. In Section 5.1 problem specific parameters are given. A general formula is topic of current research.

Remark 3. Another possibility to penalise the nonintegrality is proposed by [25]. They introduce additional inequalities, prohibiting nonintegral domains of the optimisation space.

3.2 Branch and Bound

Mixed integer dynamic optimisation problems can be solved with methods used in mixed integer nonlinear optimisation (MINLP), see [10]. This can be accomplished by parameterising problem (10) in a way as described in Section 2. Instead of a NLP (4) the result would be a MINLP of the form

$$\begin{aligned} \min_{\xi,\omega} \; & F(\xi,\omega) \\ \text{s.t.} \; & 0 \;\; = \;\; G(\xi,\omega) \\ & 0 \;\; \leq \;\; H(\xi,\omega) \\ & \omega_i \;\; \in \;\; \{0,1\}, \quad i = 1,\ldots,n_w \end{aligned} \tag{12}$$

that can be solved with methods as Branch and Bound or Outer Approximation. In the following we assume that the objective function and the feasible set are convex. In our study we apply a Branch and Bound algorithm that is performing a tree search in the space of the binary variables. We first solve a relaxed problem with $\omega \in [0,1]^{n_w}$ and decide on which of the variables with non-integral value we shall branch, say ω_i. Two new subproblems are then created with ω_i fixed to 0 and 1, respectively. These new subproblems are added to a list and the father problem is removed from it. This procedure is repeated for all problems of the list until none is left. There are three exceptions to this rule, when a node is not branched on, but abandoned directly:

1. The relaxed solution is an integer solution. Then we have found a feasible solution of the MINLP and can compare the objective value with the current upper bound (and update it, if possible).
2. The problem is infeasible. Then all problems on the subtree will be infeasible, too.
3. The objective value is higher than the current upper bound. As it is a lower bound on the objective values of all problems on the subtree, they can be abandoned from the tree search.

A more detailed description of nonlinear Branch and Bound methods and a survey about branching rules can e.g. be found in [11]. We used depth-first search and most violation branching in our implementation.

Remark 4. On each node of the search tree a NLP resulting from an optimal control problem has to be solved, which is very costly. A more efficient way of integrating the Branch and Bound scheme and SQP is proposed by [5] and [18].

Remark 5. If the functions are non-convex, the nodes cannot be fathomed any more as feasible or better solutions may be cut off. A heuristics to overcome this is proposed in [18]. An approach using underestimations of the dynamical system is described in [21].

3.3 Switching Time Approach

Another possibility to solve problem (10) is motivated by the idea to optimise the switching structure and to take the values of the controls fixed on given intervals, as is done for bang-bang arcs in indirect methods. Of course this is only valid for feasible sets $\Omega(\Psi)$ where Ψ is given by (9). Instead of (7) we have, assuming for the sake of notational simplicity a one-dimensional control, a fixed $\hat{w}(t;\hat{t},n_{\mathrm{sw}})$ given by

$$\hat{w}(t;\hat{t},n_{\mathrm{sw}}) = \begin{cases} 0 & \text{if } t \in [\hat{t}_i,\hat{t}_{i+1}], \quad \text{i even} \\ 1 & \text{if } t \in [\hat{t}_i,\hat{t}_{i+1}], \quad \text{i odd} \end{cases}, \quad i = 0,\ldots,n_{\mathrm{sw}} \tag{13}$$

with $t_0 = \hat{t}_0 \le \hat{t}_1 \le \cdots \le \hat{t}_{n_{\mathrm{sw}}+1} = T$. The number n_{sw} and the locations $\hat{t}_j$ of the switching times are then to be optimised and we obtain

$$\begin{aligned} \min_{x,z,u,p,\hat{t},n_{\mathrm{sw}}} \quad & \int_{t_0}^{T} L(x(t),z(t),\hat{w}(t;\hat{t},n_{\mathrm{sw}}),u(t),p)dt \; + E(x(T),z(T),p) \\ \text{s.t.} \quad & \dot{x}(t) = f(t,x(t),z(t),\hat{w}(t;\hat{t},n_{\mathrm{sw}}),u(t),p), \quad t \in [t_0,T] \\ & 0 = g(t,x(t),z(t),\hat{w}(t;\hat{t},n_{\mathrm{sw}}),u(t),p), \quad t \in [t_0,T] \\ & 0 \le c(t,x(t),z(t),\hat{w}(t;\hat{t},n_{\mathrm{sw}}),u(t),p), \quad t \in [t_0,T] \\ & 0 \le r_i(x(t_0),z(t_0),x(t_1),z(t_1),\ldots,x(T),z(T),p) \\ & 0 = r_e(x(t_0),z(t_0),x(t_1),z(t_1),\ldots,x(T),z(T),p) \end{aligned} \tag{14}$$

with fixed $\hat{w}(t;\hat{t},n_{\mathrm{sw}})$ and $\hat{t}_i$ and n_{sw} as above.

If we allow that switching times fall together, $\hat{t}_j = \hat{t}_{j+1}$, this formulation can be extended in a straightforward way to n_w binary control functions and every solution $(p,w(t),u(t),x(t),z(t))$ of system (10) with a finite number of switches in $w(t)$ has an equivalent solution $(p,n_{\mathrm{sw}},\hat{t},u(t),x(t),z(t))$ of system (14) and vice versa.

For fixed n_{sw} we then have an optimal control problem that fits into the definition of problem (1) and can be solved with standard methods, where the interval lengths $\hat{t}_{j+1} - \hat{t}_j$ take the role of parameters that have to be determined. Special care has to be taken to treat the case where interval lengths diminish during the optimisation procedure, causing the problem to become singular. [12, 13, 19] propose an algorithm to eliminate such non-optimal bang-bang intervals.

Some authors propose to iterate on n_{sw} until there is no further decrease in the objective function of the corresponding optimal solution [22, 12, 13]. But it should be stressed that this can only be applied to more complex systems, if good initial values for the location of the switching points are available, as they are essential for the convergence behaviour of the underlying method. In

Section 5.3 we will see how this approach may drift off to an arbitrary local optimum even in the case of a one-dimensional control, only few switching events and reasonable initialisations.

points are available, as they are essential for the convergence behaviour of the underlying method. In Section 5.3 we will see how this approach may drift off to an arbitrary local optimum even in case of a one-dimensional control, only few switching events and reasonable initialisations.

4 A Fish Population Optimal Control Problem

In this section we introduce a fish population control model as a benchmark problem for optimal control with binary control functions. This model has some oscillations that we want to bring close to a steady state. Such an optimisation objective might also be the topic of other applications, e.g. in control of pattern self-aggregation [14].

In Section 4.1 the standard textbook ordinary differential equation (ODE) model of Lotka–Volterra type is brought back to memory. This ODE model is then extended in 4.2 to a control problem by introducing a fishing allowance. In Section 4.3 we have a look at some details of this model, e.g. at the optimal relaxed solutions obtained either by a direct approach or by Pontryagins maximum principle.

4.1 ODE Model

Lotka-Volterra systems have already been under investigation for a long time and are very well studied, see e.g. [27] for an overview. In a two-species predator-prey model there are two differential states, namely the biomass of the prey $x_0(t)$ that is assumed to grow exponentially and the biomass of the predator species $x_1(t)$ that is assumed to decrease exponentially. A second coupling term standing for the probability of a contact between the two species gives a decrease in the biomass of prey and an increase in that of the predator due to eating. The system is assumed to be in a given state $x(t_0) = x_0 \geq 0$ at time t_0. All parameters typically in use in such models are assumed to be 1 for the sake of notational simplicity.

$$\begin{aligned} \dot{x}_0(t) &= x_0(t) - x_0(t)x_1(t) \\ \dot{x}_1(t) &= -x_1(t) + x_0(t)x_1(t) \\ x_i(t_0) &= x_{i0} \end{aligned} \tag{15}$$

The plots in Figure 1 show the periodic oscillating nature of this model for a given initial state $x = (0.5, 0.7)^T$.

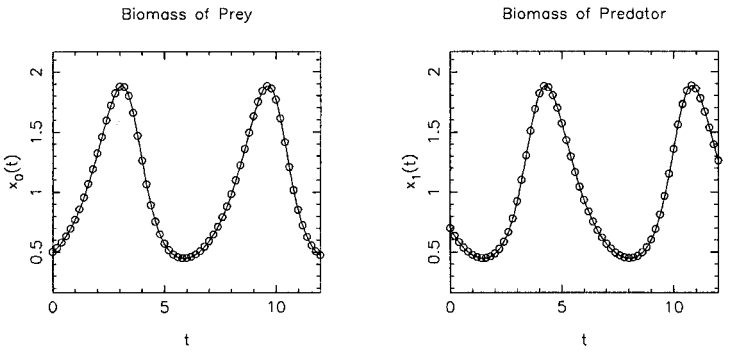


Fig. 1. Simulation of the ODE (15) for a time horizon $[t_0, T] = [0, 12]$.

4.2 Optimal Control Problem

As D'Ancona and Volterra [30] observed due to an unexpected decrease in the fishing quota after World War I – everybody expected an increase as fishing was almost completely abandoned in the war years – that there is an interconnection between the evolution of the biomasses of a system of type (15) and fishing. A very simple way to model an additional fishing aspect is the following model:

$$\begin{aligned} \dot{x}_0(t) &= x_0(t) - x_0(t)x_1(t) \quad - c_0 x_0(t)\, w(t) \\ \dot{x}_1(t) &= -x_1(t) + x_0(t)x_1(t) - c_1 x_1(t)\, w(t) \\ x_i(t_0) &= x_{i0}\,, \quad w(t) \in [0,1] \end{aligned} \tag{16}$$

Here $w(t)$ is a function describing the percentage of the fleet that is actually fishing at time t. The parameters c_0 and c_1 indicate how many fish would be caught by the entire fleet, we choose arbitrarily $c_0 = 0.4$ and $c_1 = 0.2$. The plots in Figure 2 show that amplitude and phase offset have changed, but that the periodic oscillating nature is kept for $w(t) = 1$.

One might be interested in bringing such a system close to a steady state to avoid the high fluctuations shown in Figure 2 that cause economical problems. One way to achieve this is to vary the fishing quota for a certain time span $T - t_0$. Adding an objective functional that punishes deviation from the steady state $\tilde{x} = (1,1)^T$ for $w(t) = 0$ resp. $\tilde{x} = (1 + c_1, 1 - c_0)^T$ for $w(t) = 1$ [3]

$$\min_{x,w} \int_{t_0}^{T} (x_0(t) - 1)^2 + (x_1(t) - 1)^2 \, dt$$

leads us to the following optimal control problem

[3] for the sake of notational simplicity we will stick to the first case, wanting a steady state for a system left alone after time T

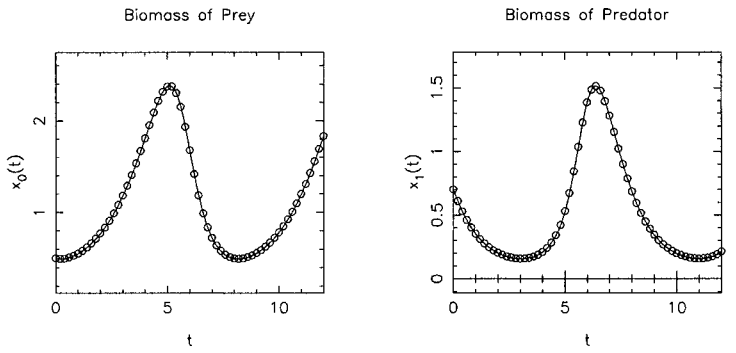


Fig. 2. Simulation of the ODE (16) with fishing for a time horizon $[t_0, T] = [0, 12]$.

$$\begin{aligned} &\min_{x,w} \int_{t_0}^{T} (x_0(t) - 1)^2 + (x_1(t) - 1)^2 \, dt \\ &\text{s.t. } \dot{x}_0(t) = x_0(t) - x_0(t)x_1(t) - c_0 x_0(t)\, w(t) \\ &\qquad \dot{x}_1(t) = -x_1(t) + x_0(t)x_1(t) - c_1 x_1(t)\, w(t) \\ &\qquad x_i(t_0) = x_{i0}\,, \quad w(t) \in [0, 1] \end{aligned} \tag{17}$$

This optimal control problem can e.g. be solved by indirect methods or with the direct multiple shooting method. Having a look at the optimal control function $w(t)$ (see Figure 3) one notices that the percentage of the fleet that is fishing at a given time t is varying strongly on the singular arc, which would be practically very hard to achieve in the fishing example. From an economic point of view it would be easier to operate either the entire fleet or to do no fishing at all and use the manforce for different things in the meantime (on fixed time intervals corresponding to weeks). This could be achieved by laws that prohibit fishing for a certain time span. This would lead us to a model of the form

$$\begin{aligned} &\min_{x,w} \int_{t_0}^{T} (x_0(t) - 1)^2 + (x_1(t) - 1)^2 \, dt \\ &\text{s.t. } \dot{x}_0(t) = x_0(t) - x_0(t)x_1(t) - c_0 x_0(t)\, w(t) \\ &\qquad \dot{x}_1(t) = -x_1(t) + x_0(t)x_1(t) - c_1 x_1(t)\, w(t) \\ &\qquad x_i(t_0) = x_{i0}\,, \quad w(t) \in \Omega(\Psi) \end{aligned} \tag{18}$$

where the control function is restricted to take values of either 0 or 1 and change its value only at given time points, see definition (7) of $\Omega(\Psi)$. In our case we assume a hime horizon $[t_0, T] = [0, 12]$ and $n_\tau = 60$ equidistant time points, e.g. the start of a working week or a month, to be feasible switching points. Therefore all calculations in this paper are done using a control parameterisation of $m = 60$ intervals.

4.3 Relaxed Solutions

The relaxed fishing problem (17) can be solved efficiently with the standard direct multiple shooting method. It is also possible to apply indirect methods, which is typically much harder for higher dimensional optimal control problems with a complex switching structure. Here we give some details to deliver a more comprising understanding of the system under investigation. We write x_i for $x_i(t)$, w for $w(t)$ and λ_i for $\lambda_i(t)$. The Hamiltonian in normalized form H and the adjoint equations of (17) are given by

$$\begin{aligned} H &:= -L(x) + \lambda^T f(x,w) \\ &= -(x_0-1)^2 - (x_1-1)^2 \\ &\quad +\lambda_0(x_0 - x_0x_1 - c_0x_0w) + \lambda_1(-x_1 + x_0x_1 - c_1x_1w) \\ \dot\lambda_0 &:= -H_{x_0} = 2(x_0-1) - \lambda_0(1 - x_1 - c_0w) - \lambda_1 x_1 \\ \dot\lambda_1 &:= -H_{x_1} = 2(x_1-1) + \lambda_0 x_0 - \lambda_1(-1 + x_0 - c_1w) \end{aligned}$$

The switching function $S(x,\lambda)$ is given by $S(x,\lambda) := H_w = -c_1\lambda_1x_1 - c_2\lambda_2x_2$. It can be shown, e.g. with the methods presented in [19], that the optimal relaxed control (neglecting $\Psi = \{\tau_1, \tau_2, \ldots, \tau_{60}\}$) has the form

$$w(t, x(t), \hat t) := \begin{cases} 0 & \text{for } t \in [t_0, \hat t_1] \\ 1 & \text{for } t \in [\hat t_1, \hat t_2] \\ w_{\text{sing}}(x) & \text{for } t \in [\hat t_2, T] \end{cases} \tag{19}$$

The singular control is of *order one* since the second total time derivative $S^2(x,\lambda,u)$ of the switching function $S(x,\lambda)$ contains the control explicitly. Then along a singular arc the equations $S(x,\lambda) = 0, S^1(x,\lambda) = \frac{d}{dt}S(x,\lambda) = 0$ and $S^2(x,\lambda,u) = 0$ hold from which one can compute the singular control in feedback form (the adjoint variables λ can be eliminated),

$$\begin{aligned} w_{\text{sing}} = (\ & c_0^3x_0^3 - c_1^3x_1^3 + c_0^3x_0^2x_1 - c_1^3x_0x_1^2 + 2c_0x_0x_1^2c_1^2 - 2c_1x_0^2x_1c_0^2 \\ & - 4c_0^2x_0c_1x_1^2 + 2c_0^2x_0c_1x_1 + 4c_1^2x_1c_0x_0^2 - 2c_1^2x_1c_0x_0 \\ & - x_0^3x_1c_0^3 + x_0^2x_1^2c_1^3 + x_0x_1^3c_1^3 - 2x_0^2x_1^2c_1^2c_0 + x_0^3x_1c_1c_0^2 \\ & - x_0x_1^3c_0c_1^2 - x_0^3x_1c_1^2c_0 - x_0^2x_1^2c_0^3 + 2x_0^2x_1^2c_0^2c_1 + x_0x_1^3c_0^2c_1\ ) \\ / (\ & c_0^4x_0^3 + 2c_0^2x_0^2c_1^2x_1 - 2c_0^2x_0c_1^2x_1 + 2c_1^2x_1^2c_0^2x_0 \\ & + c_1^4x_1^3 - c_0^3x_0c_1x_1^2 + c_0^3x_0c_1x_1 - c_1^3x_1c_0x_0^2 + c_1^3x_1c_0x_0\ ) \end{aligned} \tag{20}$$

for the singular arc $[\hat t_2, T]$. The parameters $\hat t_1$ and $\hat t_2$ can be determined to $\hat t_1 = 2.43670$ and $\hat t_2 = \hat t_1 + 1.50526$ by solving a boundary value problem. The computed initial values of the adjoint states are $\lambda_0(0) = 5.83903$ and $\lambda_1(0) = 1.53101$. The resulting optimal control $w(t)$ is shown in Figure 3, together with the optimal parameterised control obtained by appyling the direct multiple shooting method. Figure 4 shows the corresponding state trajectories. The minimum deviations obtained by these controls are $\Phi = 1.34408$ for the indirect method and $\Phi = 1.34466$ for the parameterised approximation (that takes into account $\Psi = \{\tau_1, \tau_2, \ldots, \tau_{60}\}$).

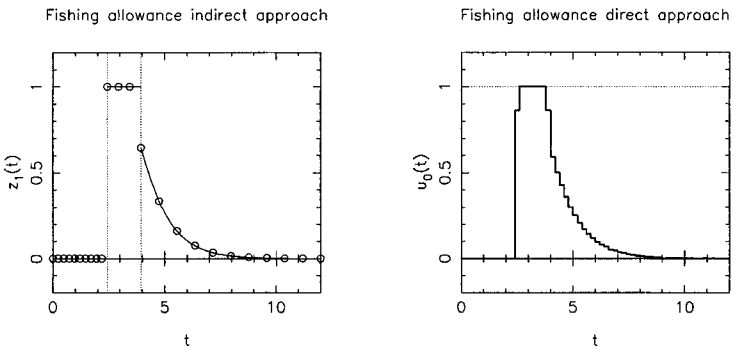


Fig. 3. Optimal controls for the relaxed problem. Left: indirect approach. Right: direct multiple shooting with 60 intervals.

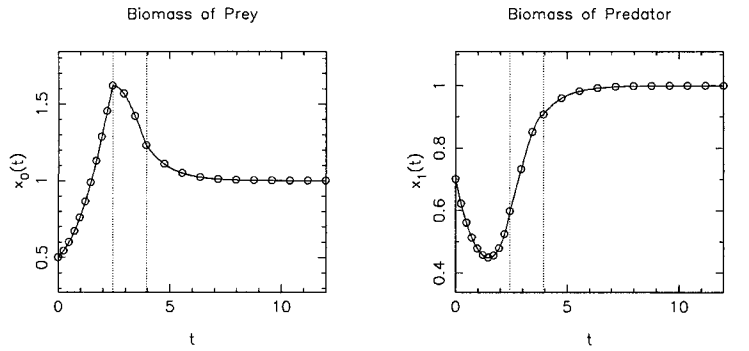


Fig. 4. Optimal states for the relaxed problem, obtained with the indirect method.

5 Numerical Results

In this section we want to show and discuss results obtained from the application of the methods described in Section 3 to the optimal control problem presented in Section 4. In 5.1 we will show how certain heuristics perform, while 5.2 gives results for the global Branch and Bound approach. In 5.3 we discuss a possible extension to a continuous optimisation of the switching points.

In all problems that had to be solved, the control was initialised with $w(t) = 0\ \forall t \in [t_0, T]$ and the initial multiple shooting node values were obtained by integration with this fixed control. As a measurement for computational effort we consider the number of QPs to be solved (more precise would be SQP iterations as there is additional effort such as solving the ODE and linearising).

5.1 Heuristics

Heuristic solutions for problem (18) can be obtained in a number of ways. In this discussion we will focus on the rounding heuristics and the penalty term homotopy described in Section 3.1. Table 1 shows how many quadratic programs (QPs) had to be solved and what objective values were obtained. $w(t) = 0$ and $w(t) = 1$ correspond to the cases where the control is fixed to one value for the whole time horizon, thus no fishing at all or fishing all the time. The plots in Figures 1 and 2 show the respective states for these controls. They both do not require any QP to be solved, as an integration of the system with fixed controls suffices to obtain the solution. The rounding heuristics require a relaxed solution $\bar{w}(t)$ of problem (17), which takes 23 iterations. The penalty parameters for the penalty homotopy (see 3.1) are

Heuristics	# QPs	Objective value
$w(t) = 0$	0	6.06187
$w(t) = 1$	0	9.40231
Rounding 1	23	1.51101
Rounding 2	23	1.45149
Rounding 3	23	1.34996
Penalty homotopy	89	1.34898

Table 1. Number of QPs to be solved and obtained objective value for several heuristics to get a feasible solution.

chosen exponentially increasing as

$$\epsilon_i^k = \epsilon_{\text{init}} * \epsilon_{\text{inc}}^{k-1}$$

for $k \geq 1$ and $\epsilon_i^0 = 0$ to get the relaxed parameterised solution as starting point for the homotopy. A choice of $\epsilon_{\text{init}} = 10^{-4}$ and $\epsilon_{\text{inc}} = 2.1$ showed good results. A faster increase in the penalty parameters is getting closer to simple rounding, while a slower increase leads to an augmentation in the number of QPs to be solved. All problems of the homotopy were solved to an accuracy of 10^{-4}, while all other problems in this paper were solved up to 10^{-6}.

Table 1 shows that the proposed homotopy delivers a solution with an objective value of 1.34898 closer to the objective value 1.34466 of the parameterised relaxed model (17) than the rounding heuristics. As the optimal solution of (17) is a lower bound on the optimal integer solution of (18), the difference gives an indication about how good our heuristic solution is. As the relative gap of about 0.3% is known at runtime, one can decide whether the obtained solution suffices, otherwise one has to turn to global methods, where it can be used as an upper bound. This will be the topic of the next section.

5.2 Branch and Bound

Before applying a Branch and Bound approach to problem (18) we have to investigate whether the optimal control problem is convex. The feasible set is the hypercube in $\mathbb{R}^{n_w}$ and thus convex. We do not show analytically that the objective function is convex as well, but refer to Figure 5 that shows the behaviour of the objective function in the vicinity of the optimal relaxed solution – on 59 stages $w(t)$ is fixed and on one stage the value is changing. In

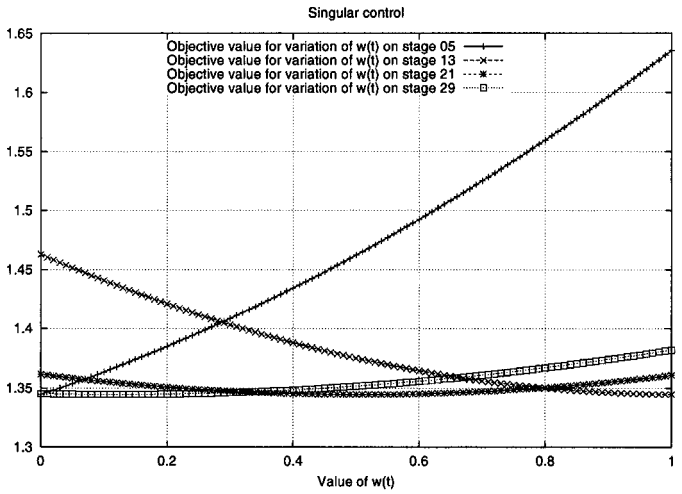


Fig. 5. The objective function in the vicinity of the optimal solution: for four selected stages the control on this stage is changing its value while the other 59 values are fixed. The trajectories give an indication for the convexity of the objective function over the feasible set.

the following, we do assume that the objective function is indeed convex for the given data. Thus we can apply a Branch and Bound approach as presented in Section 3.2. Figure 6 shows the optimal controls obtained by the Branch and Bound approach and, for comparison, a rounded control. The optimal solution on $\Psi = \{\tau_1, \tau_2, \ldots, \tau_{60}\}$ is

$$w(t) = \begin{cases} 0 & t \in [\tau_i, \tau_{i+1}] \text{ and } i \in I_{\text{off}} \\ 1 & t \in [\tau_i, \tau_{i+1}] \text{ and } i \in I_{\text{on}} \end{cases}$$

with

$$I_{\text{on}} = \{13, 14, \ldots, 20, 22, 25, 28\}$$

$$I_{\text{off}} = \{1, 2, \ldots, 60\} \setminus I_{\text{on}}$$

Figure 7 shows the state trajectories of the biomasses that correspond to

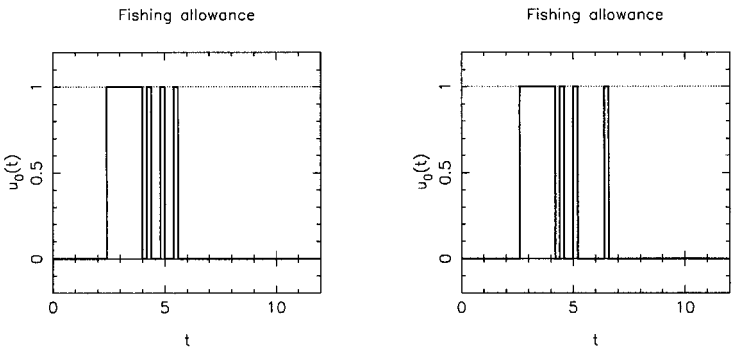


Fig. 6. Left: global optimal control obtained by Branch and Bound. Right: optimal control obtained by rounding strategy 2.

the optimal integer solution obtained by Branch and Bound. Note the non-differentiabilities in the states caused by the switching of the control function.

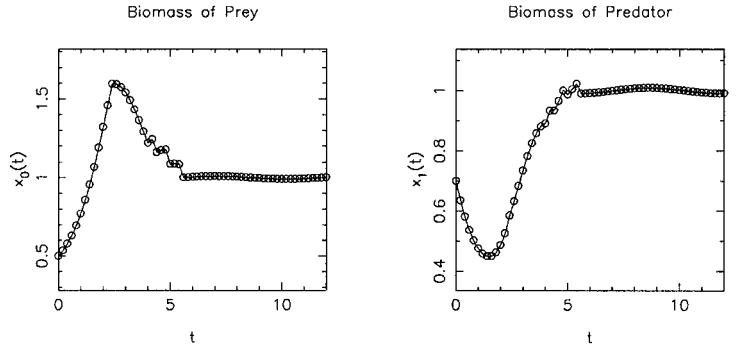


Fig. 7. State trajectories corresponding to the optimal integer solution.

The behaviour of the Branch and Bound method depends very strongly on the availability of an upper bound. Table 2 gives performance data for different heuristics to get such an a priori upper bound. The first integer solution and therefore upper bound is found after branching on 33 variables, if no nodes are fathomed. The objective value of this feasible solution is 1.36614 – thus it is clear, that no heuristics will help to reduce the size of the Branch and Bound tree that delivers an upper bound above this value and it explains why two rounding heuristics perform as bad as the Branch and Bound without any upper bounding heuristics at all. Rounding strategy 3 gives a result

Start heuristics	# Nodes	# QPs	Opt. in iter	First upper bound
None	1634	15720	1064	∞
Rounding 1	1634	15720	1064	1.51101
Rounding 2	1634	15720	1064	1.45149
Rounding 3	906	9210	336	1.34996
Penalty homotopy	757	7769	1	1.34898

Table 2. Number of nodes in the Branch and Bound tree, overall number of QPs to be solved, the node iteration when the optimal solution is found and the start upper bound for several heuristics to produce an upper bound.

close to the optimal solution, differing only on intervals $22, 23$ and $27, 28$. The solution obtained by the penalty homotopy turns out to be the global solution, although further 756 nodes have to be visited and 7680 SQP iterations are needed to verify this in our Branch and Bound implementation.

5.3 Switching Time Approach

Although we have the additional Ψ-restriction on $w(t) \in \Omega(\Psi)$ resp. $w(t) \in \bar{\Omega}(\Psi)$, it is interesting to investigate how much we could improve an obtained solution by giving additional degrees of freedom. Table 3 lists results obtained

Method	n_{sw}	# QPs	Objective value
ST after Rounding 2	8	35	1.34541
ST after B&B	8	7852	1.34604
ST after Penalty homotopy	8	171	1.34604
ST after Rounding 3	8	35	1.34616
ST after Rounding 1	2	38	1.38273
ST after Initialisation by hand	8	142	1.38273

Table 3. Switching time optimisation results.

by a switching time (ST) approach. n_{sw} and initial values for $\hat{t}_1, \ldots, \hat{t}_{n_{\text{sw}}}$ are obtained by a transformation of a solution $w(t)$ of problem (18). To get this, methods investigated in 5.1 or 5.2 are used. After the transformation the controls $w(t)$ and $w(t; \hat{t}, n_{\text{sw}})$ are identical. With this start initialisation the switching times $\hat{t}$ are optimised, n_{sw} is kept constant. *Initialisation by hand* is a solution set up arbitrarily in the following way:

$$n_{\text{sw}} = 8, \hat{t}_0 = 0.0, \hat{t}_1 = 2.8, \hat{t}_2 = 4.0, \hat{t}_3 = 7.0, \hat{t}_4 = 7.4,$$
$$\hat{t}_5 = 8.0, \hat{t}_6 = 8.2, \hat{t}_7 = 10.0, \hat{t}_8 = 10.2, \hat{t}_9 = 12.0$$

Figure 8 shows this initialisation and the control obtained by an optimisation of $\hat{t}$, Figure 9 shows the corresponding states. Although n_{sw} is chosen

such that eight switches are allowed, the optimisation procedure reduced three intervals to size zero and ends in the local solution also found by optimisation after initialisation with rounding strategy 1 (with $n_{\text{sw}} = 2$). This makes clear that it is not enough to simply increase n_{sw} without suppyling good initial values for a switching time approach. Another interesting point is that initialisation with rounding strategy 2 gives the best solution

$$\begin{aligned} n_{\text{sw}} = 8, \hat{t}_0 = 0.00000, \hat{t}_1 = 2.44093, \hat{t}_2 = 4.07798, \\ \hat{t}_3 = 4.29155, \hat{t}_4 = 4.50443, \hat{t}_5 = 4.90853 \\ \hat{t}_6 = 5.12223, \hat{t}_7 = 6.15604, \hat{t}_8 = 6.28131, \hat{t}_9 = 12.0 \end{aligned} \tag{21}$$

although it has a higher objective value than other initialisations. This is also due to the many local minima in the switching time formulation. Table 4 gives a final comparing overview over all obtained solutions in this study.

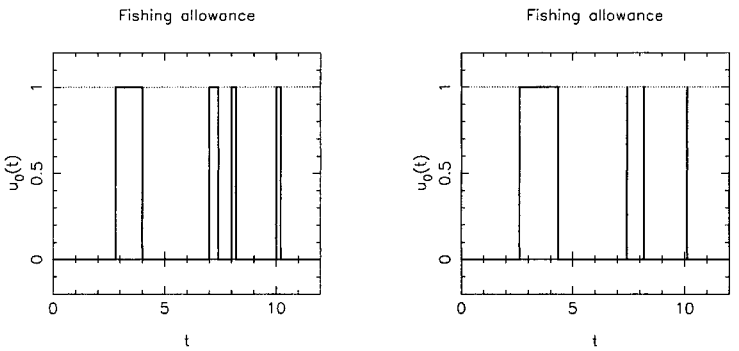


Fig. 8. Control initialisation set up by hand and resulting control after optimisation of switching points. Three intervals shrink to length zero.

Remark 6. We do know from the maximum principle that a bang-bang control exists with the same objective function value as the solution including the singular arc, thus solution (21) is still suboptimal, n_{sw} needs to be increased (probably to infinity). Here we are content with the feasible integer solution (21), being closer than 10^{-3} to the relaxed parameterised solution.

6 Conclusion

We have presented a benchmark problem for optimal control with $0-1$ controls that can be extended in a straightforward way to several species, other parameters or discretisations. Several heuristics and a global approach, namely a Branch and Bound strategy, have been described and applied successfully.

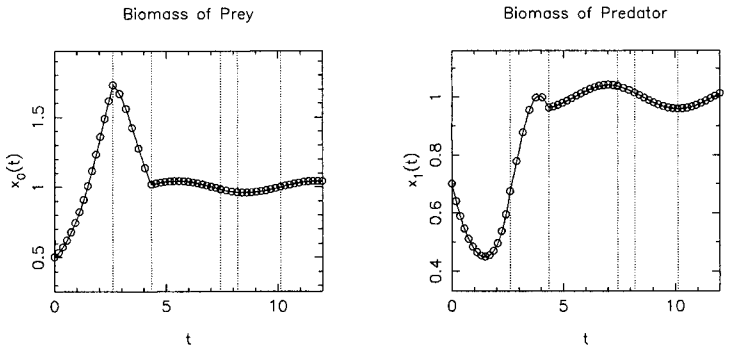


Fig. 9. State trajectories corresponding to control of Figure 8.

Method	# QPs	Objective value
Relaxed indirect	28	1.34408
Relaxed parameterised	23	1.34466
ST after Rounding 2	35	1.34541
ST after B&B	7852	1.34604
ST after Penalty homotopy	171	1.34604
ST after Rounding 3	35	1.34616
B&B	7769	1.34898
Penalty homotopy	89	1.34898
Rounding 3	23	1.34996
ST after Rounding 1	38	1.38273
ST after Initialisation by hand	142	1.38273
Rounding 2	23	1.45149
Rounding 1	23	1.51101
$w(t) = 0$	0	6.06187
$w(t) = 1$	0	9.40231

Table 4. Overview: number of QPs to be solved and obtained objective value for different approaches. Parameterisation was done with 60 multiple shooting intervals.

Numerical results have been given that show the potential of a penalty term homotopy. In the special problem considered in this study it delivered the global (under a convexity assumption) optimal solution for a fixed time grid. Furthermore we showed how these methods may be used to initialise parameters in a switching time approach to deal with problems without fixed time grid.

The methods described in this paper, several heuristics and Branch and Bound, are implemented in a software package based on the direct multiple shooting method and advanced algorithms also implemented in `MUSCOD-II` and may be applied to larger-scale optimal control problems with $0-1$ controls in the future.

Future research will focus on a globalisation of the penalty term homotopy and further applications in cell biology and chemical engineering.

Acknowledgements. We thank the anonymous referee for helpful remarks and suggestions.

References

1. T. Binder, L. Blank, H.G. Bock, R. Bulirsch, W. Dahmen, M. Diehl, T. Kronseder, W. Marquardt, J.P. Schlöder, and O.v. Stryk. Introduction to model based optimization of chemical processes on moving horizons. In M. Grötschel, S.O. Krumke, and J. Rambau, editors, Online Optimization of Large Scale Systems: State of the Art, pp. 295–340. Springer, 2001.
2. H.G. Bock, E. Eich, and J.P. Schlöder. Numerical solution of constrained least squares boundary value problems in differential-algebraic equations. In K. Strehmel, editor, Numerical Treatment of Differential Equations. Teubner, Leipzig, 1988.
3. H.G. Bock and R.W. Longman. Computation of optimal controls on disjoint control sets for minimum energy subway operation. In Proc. Amer. Astronomical Soc., Symposium on Engineering Science and Mechanics, Taiwan, 1982.
4. H.G. Bock and K.J. Plitt. A multiple shooting algorithm for direct solution of optimal control problems. In Proc. 9th IFAC World Congress Budapest, pp. 243–247. Pergamon Press, 1984.
5. B. Borchers and J.E. Mitchell. An improved branch and bound algorithm for mixed integer nonlinear programming. Computers and Oper. Res., 21(4): 359–367, 1994.
6. U. Brandt-Pollmann. Numerical solution of optimal control problems with implicitly defined discontinuities with applications in engineering. PhD Thesis, IWR, Univ. of Heidelberg, 2004.
7. J. Burgschweiger, B. Gnädig, and M.C. Steinbach. Optimization models for operative planning in drinking water networks. Tech. Rep. ZR-04-48, ZIB, 2004.
8. M. Buss, M. Glocker, M. Hardt, O. v. Stryk, R. Bulirsch, and G. Schmidt. Nonlinear hybrid dynamical systems: Modelling, optimal control, and applications. In S. Engell, G. Frehse, and E. Schnieder, editors, Modelling, Analysis and Design of Hybrid Systems, Lect. Notes in Control and Information Science, Vol. 279, pp. 311–335, Heidelberg, Springer-Verlag, 2002
9. M. Diehl, D.B. Leineweber, and A.A.S. Schäfer. MUSCOD-II Users' Manual. IWR-Preprint 2001-25, Univ. of Heidelberg, 2001.
10. I.E. Grossmann and Z. Kravanja. Mixed-integer nonlinear programming: A survey of algorithms and applications. In Biegler et al., editors, Large-Scale Optimization with Applications. Part II: Optimal Design and Control, Vol.93 of The IMA Volumes in Math. and its Appl., Springer Verlag, 1997.
11. O.K. Gupta and A. Ravindran. Branch and bound experiments in convex nonlinear integer programming. Manag. Science, 31: 1533–1546, 1985.
12. C.Y. Kaya and J.L. Noakes. Computations and time-optimal controls. Optimal Control Appl. and Methods, 17: 171–185, 1996.
13. C.Y. Kaya and J.L. Noakes. A computational method for time-optimal control. J. Optim. Theory Appl., 117: 69–92, 2003.

14. D. Lebiedz and U. Brandt-Pollmann. Manipulation of self-aggregation patterns and waves in a reaction-diffusion system by optimal boundary control strategies. Phys. Rev. Lett., 91(20), 2003.
15. H.W.J. Lee, K.L. Teo, L.S. Jennings, and V. Rehbock. Control parametrization enhancing technique for optimal discrete-valued control problems. Automatica, 35(8): 1401–1407, 1999.
16. D.B. Leineweber. Efficient reduced SQP methods for the optimization of chemical processes described by large sparse DAE models, Vol. 613 of Fortschr.-Ber. VDI Reihe 3, Verfahrenstechnik. VDI Verlag, Düsseldorf, 1999.
17. D.B. Leineweber, I. Bauer, H.G. Bock, and J.P. Schlöder. An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part I: Theoretical aspects. Comp. and Chemical Eng., 27: 157–166, 2003.
18. S. Leyffer. Integrating SQP and branch-and-bound for mixed integer nonlinear programming. Computational Optim. and Appl., 18(3): 295–309, 2001.
19. H. Maurer, C. Büskens, J.H.R. Kim, and Y. Kaya. Optimization methods for the verification of second-order sufficient conditions for bang-bang controls. Optimal Control Meth. and Appl., 2004. submitted.
20. J. Oldenburg, W. Marquardt, D. Heinz, and D.B. Leineweber. Mixed logic dynamic optimization applied to batch distillation process design. AIChE J., 49(11): 2900–2917, 2003.
21. I. Papamichail and C.S. Adjiman. Global optimization of dynamic systems. Computers and Chemical Eng., 28: 403–415, 2004.
22. V. Rehbock and L. Caccetta. Two defence applications involving discrete valued optimal control. ANZIAM J., 44(E): E33–E54, 2002.
23. V.H. Schulz. Reduced SQP methods for large-scale optimal control problems in DAE with application to path planning problems for satellite mounted robots. PhD Thesis, Univ. of Heidelberg, 1996.
24. M.S. Shaikh. Optimal control of hybrid systems: theory and algorithms. PhD Thesis, Dep. Elect. and Computer Eng., McGill Univ., Montréal, Canada, 2004.
25. O. Stein, J. Oldenburg, and W. Marquardt. Continuous reformulations of discrete-continuous optimization problems. Computers and Chemical Eng., 28(10): 3672–3684, 2004.
26. O. von Stryk and M. Glocker. Decomposition of mixed-integer optimal control problems using branch and bound and sparse direct collocation. In Proc. ADPM 2000 – The 4th Int. Conf. on Automatisation of Mixed Processes: Hybrid Dynamical Systems, pp. 99–104, 2000.
27. Y. Takeuchi. Global Dynamical Properties of Lotka-Volterra Systems. World Scientific Publishing, 1996.
28. S. Terwen, M. Back, and V. Krebs. Predictive powertrain control for heavy duty trucks. In Proc. IFAC Symp. in Advances in Automotive Control, Salerno, Italy, 2004.
29. T.H. Tsang, D.M. Himmelblau, and T.F. Edgar. Optimal control via collocation and non-linear programming. Int. J. on Control, 1975.
30. V. Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem. R. Accad. Naz. dei Lincei., VI-2, 1926.

Part III

Game Theory

Some Characterizations of Convex Games *

Juan Enrique Martínez-Legaz

Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. juanenrique.martinez@uab.es

Summary. Several characterizations of convexity for totally balanced games are presented. As a preliminary result, it is first shown that the core of any subgame of a nonnegative totally balanced game can be easily obtained from the maximum average value (MAV) function of the game. This result is then used to get a characterization of convex games in terms of MAV functions. It is also proved that a game is convex if and only if all of its marginal games are totally balanced.

1 Introduction

This paper contains some characterizations of convexity for totally balanced games. Totally balancedness was defined by Shapley and Shubik [8] as the property of having all subgames with nonempty core. These authors proved that totally balanced games coincide with market games generated by exchange economies whose traders have continuous concave utility functions. Another characterization of totally balanced games, namely, as flow games, was provided by Kalai and Zemel [3]. A flow game arises from a directed network each of whose arcs has a given capacity and belongs to a unique player; the worth of a coalition is the maximum flow that can be sent from the source to the sink by using only the arcs owned by its members. The totally balanced character of flow games is a consequence of the max flow-min cut theorem of Ford and Fulkerson [2], according to which the maximum source to sink flow equals the minimum capacity of a cut (i.e., of a set of arcs such that, when removed from the network, nothing can be sent from the source to the sink). Nonnegative totally balanced games are also known to be equivalent to linear production games in the sense of Owen [6]. Indeed, to any nonnegative

* This work has been supported by the Ministerio de Ciencia y Tecnología (Spain) and the FEDER, project BEC2002-00642, and by the Departament d'Universitats, Recerca i Societat de la Informació, Direcció General de Recerca de la Generalitat de Catalunya, project 2001SGR-00162. The author thanks the support of the Barcelona Economics Program of CREA.

game one can associate the linear production game in which the resources are the players, each of which owns only one unit of himself, the goods are the nonempty coalitions, each of which can be sold at a price equal to its worth, and to produce one unit of a given coalition one requires one unit of each of its members. One can easily show that the linear production game so defined is precisely the totally balanced cover of the initial game (i.e., its smallest totally balanced majorant). Note that this linear production representation of a nonnegative totally balanced game needs n resources (n being the number of players) and $2^n - 1$ goods. An alternative linear production representation, requiring just one good and at most $2^n - 1$ resources, can be deduced from the observation, due to Kalai and Zemel [3], that the class of totally balanced games is the span of the additive games by the minimum operation.

Section 2 deals with nonnegative totally balanced games. For these games, a duality theory has been proposed by Martínez-Legaz [5], relating them to a special class of convex functions. To each nontrivial nonnegative game, one associates its maximum average value (MAV) function, which is convex and contains all the information on the game provided that it is totally balanced. Since totally balanced games have all subgames with nonempty core, the natural question arises how to compute these cores from the MAV function of the game. A simple answer to this question is given in Section 3, where it is shown that the computation of the core of a subgame reduces to minimizing the MAV function of the game subject to a simple linear constraint. In sections 4 and 5 we consider a very special class of totally balanced games, namely, that of convex games. Both sections have in common that they analyze convexity from the point of view of totally balancedness. In Section 4, this analysis is made by means of MAV functions: We characterize convex games in terms of the optimal solutions to the optimization problems that arise in the computation of the cores of the subgames. Section 5 analyzes convex games by means of their marginal games; the main result in this section establishes that convex games are precisely those games all of whose marginal games are totally balanced.

We shall use some basic notions of convex analysis (in particular, the concept of subdifferential), for which we refer to the classical book by Rockafellar [7].

2 The MAV Function of a TU Game

A TU game is a pair $\Gamma = (N, v)$, where N is a finite set of players, and $v : 2^N \longrightarrow \mathbb{R}$ is a function, called the characteristic function of the game, defined on the power set of N and satisfying the condition $v(\emptyset) = 0$. In this section we will only consider nontrivial nonnegative games, i.e., those whose characteristic function satisfies $v(S) \geq 0$ for all $S \in 2^N$ and is not identically zero. As is well known, there is no loss of generality in assuming that a totally balanced game is nonnegative, since one can replace the original

game by another strategically equivalent 0-normalized game, which is totally balanced and nonnegative. For such games, the following duality theory has been developed by Martínez-Legaz [5]. One defines $\mu : \mathbb{R}_+^N \setminus \{0\} \longrightarrow \mathbb{R}_{++} \cup \{+\infty\}$, the maximum average value (MAV) function of Γ, by

$$\mu(w) = \max_{S \subset N} \frac{v(S)}{w(S)} \qquad \left(w = (w_i)_{i \in N} \in \mathbb{R}_+^N \setminus \{0\}\right)$$

(with the conventions $\frac{\alpha}{0} = +\infty$ for any $\alpha > 0$ and $\frac{0}{0} = 0$), where $w(S) = \sum_{i \in S} w_i$. This function admits the following economic interpretation: if the components of w represent the salaries demanded by the players and $v(S)$ is the total amount of output produced to an employer by a set S of players when they use his resources, then $\mu(w)$ is the maximum amount of output per unit of money spent that the employer can obtain by hiring a coalition. In order to make this paper self-contained, we restate here the main results (Theorem 2.1 and Corollary 2.2) in Martínez-Legaz [5]:

Theorem 1. *The MAV function $\mu : \mathbb{R}_+^N \setminus \{0\} \longrightarrow \mathbb{R}_{++} \cup \{+\infty\}$ of any nontrivial nonnegative TU game $\Gamma = (N, v)$ is a positively homogeneous of degree -1 continuous convex function, finite valued on $\mathbb{R}_{++}^N$, such that, at each point where the gradient exists, all of its nonzero components are the same. Conversely, if $\mu : \mathbb{R}_+^N \setminus \{0\} \longrightarrow \mathbb{R}_{++} \cup \{+\infty\}$ satisfies these conditions then there exists a unique nontrivial totally balanced nonnegative TU game $\Gamma = (N, v)$ having μ as its MAV function; its characteristic function v is given by*

$$v(S) = \min_{w \in \mathbb{R}_+^N \setminus \{0\}} \mu(w) w(S) \qquad \forall S \subset N \tag{1}$$

(with the convention $(+\infty) \cdot 0 = +\infty$).

Corollary 1. *Let $\Gamma = (N, v)$ be a nontrivial nonnegative TU game with MAV function μ and let $\tilde{v} : 2^N \longrightarrow \mathbb{R}$ be defined by*

$$\tilde{v}(S) = \min_{w \in \mathbb{R}_+^N \setminus \{0\}} \mu(w) w(S) \qquad \forall S \subset N. \tag{2}$$

Then $\tilde{\Gamma} = (N, \tilde{v})$ is the totally balanced cover of Γ, i.e., $\tilde{v}$ is the smallest majorant of v that defines a totally balanced game.

Corollary 2. *The MAV function of any nontrivial nonnegative TU game coincides with that of its totally balanced cover.*

Proof. According to Theorem 1, for any nontrivial nonnegative n-person TU game there is a unique totally balanced game with the same MAV function; by Corollary 1, this totally balanced game is precisely the totally balanced cover of the initial game. □

To illustrate Corollary 1, consider the game (N, v) with $N = \{1, 2, 3\}$ and v defined by

$$v(S) = \begin{cases} 0 & \text{if } |S| \leq 1, \\ 1 & \text{if } |S| \geq 2. \end{cases}$$

One can easily check that the MAV function μ of this game is given by

$$\mu(w_1, w_2, w_3) = \frac{1}{\min\{w_1 + w_2, w_1 + w_3, w_2 + w_3\}}. \tag{3}$$

Thus, according to (2), the characteristic function of the totally balanced cover $(N, \tilde{v})$ of (N, v) is given by

$$\tilde{v}(S) = \begin{cases} \min_{w \in \mathbb{R}^3_+ \setminus \{0\}} \mu(w_1, w_2, w_3) w_i = 0 & \text{if } S = \{i\}, \\ \min_{w \in \mathbb{R}^3_+ \setminus \{0\}} \mu(w_1, w_2, w_3)(w_i + w_j) = 1 & \text{if } S = \{i, j\}, \text{ with } i \neq j, \\ \min_{w \in \mathbb{R}^3_+ \setminus \{0\}} \mu(w_1, w_2, w_3)(w_1 + w_2 + w_3) = 3/2 & \text{if } S = N. \end{cases}$$

Indeed, the minima in this formula are attained, e.g., at the points (w_1, w_2, w_3) given by $w_i = 0$ and $w_j = 1$ for $j \neq i$ in the first case, and at $(1, 1, 1)$ in the other two cases. Notice also that, by Corollary 2, the MAV function of $(N, \tilde{v})$ is μ.

3 Computing the Core of a Subgame

From Theorem 1, it follows that the characteristic function of a nontrivial nonnegative totally balanced game Γ can be recovered from its MAV function by means of (1). It turns out that, in this case, μ contains all the information on the game. Therefore, it is in principle possible to compute the cores of the subgames of Γ (which are nonempty as Γ is totally balanced) directly from μ. A way for doing it is suggested by the following theorem.

Theorem 2. *Let $\Gamma = (N, v)$ be a nontrivial nonnegative totally balanced TU game with MAV function μ and let $T \subset N$ be such that $v(T) > 0$. For any $x \in \mathbb{R}^T_+ \setminus \{0\}$, the following statements are equivalent:*

(1) The point x belongs to the core of the subgame $\Gamma_T = (T, v_{|2^T})$.

(2) There exists $\overline{w} \in \mathbb{R}^N_+ \setminus \{0\}$ such that $x = \overline{w}_T := (\overline{w}_i)_{i \in T}$ and $\mu(\overline{w}) = 1$; for every $\overline{w} \in \mathbb{R}^N_+ \setminus \{0\}$ satisfying these conditions, $\dfrac{\overline{w}}{\overline{x}(T)}$ is an optimal solution of

$$(\mathcal{P}_T) \quad \begin{array}{l} \text{minimize } \mu(w) \\ \text{subject to } w(T) = 1. \end{array}$$

(3) There exists $\overline{w} \in \mathbb{R}^N_+ \setminus \{0\}$ such that $x = \overline{w}_T$, $\mu(\overline{w}) = 1$ and $\dfrac{\overline{w}}{\overline{x}(T)}$ is an optimal solution of $(\mathcal{P}_T)$.

Proof. To prove the implication (1) $\Rightarrow$ (2), let x be a core element of Γ_T and take any $\overline{w} \in \mathbb{R}_+^N \setminus \{0\}$ such that $\overline{w}_T = x$ and $v(S) \leq \overline{w}(S)$ for all $S \not\subset T$ (this condition can be achieved by giving sufficiently high values to $\overline{w}_i$ for $i \notin T$). Since we also have $v(S) \leq \overline{w}(S)$ for all $S \subset T$ (as $\overline{w}_T = x$ is in the core of Γ_T), it follows that $\mu(\overline{w}) \leq 1$. But we actually have $\mu(\overline{w}) = 1$, as a consequence of

$$\mu(\overline{w}) \geq \frac{v(T)}{\overline{w}(T)} = \frac{v(T)}{x(T)} = 1.$$

Let $\overline{w} \in \mathbb{R}_+^N \setminus \{0\}$ be any point satisfying $x = \overline{w}_T$ and $\mu(\overline{w}) = 1$. By $\overline{w}_T = x$, the point $\overline{w}/x(T)$ is a feasible solution to problem $(\mathcal{P}_T)$. To show that it is optimal, it suffices to observe that, for each feasible $w \in \mathbb{R}_+^N \setminus \{0\}$, one has

$$\mu(w) \geq \frac{v(T)}{w(T)} = v(T) = x(T) = \mu(\overline{w})x(T) = \mu\left(\frac{\overline{w}}{x(T)}\right).$$

Implication (2) $\Longrightarrow$ (3) is obvious. Let us now prove (3) $\Longrightarrow$ (1). Given $\overline{w}$ as in (3) and any $S \subset T$, we have

$$v(S) \leq \mu(\overline{w})\overline{w}(S) = \overline{w}(S) = x(S).$$

Take $w \in \mathbb{R}_+^N \setminus \{0\}$ such that $\mu(w)w(T) = v(T)$ (the existence of w follows from Corollary 1). From the optimality of $\overline{w}/x(T)$, we deduce that

$$x(T) = \mu(\overline{w})x(T) = \mu\left(\frac{\overline{w}}{x(T)}\right) \leq \mu\left(\frac{w}{w(T)}\right) = \mu(w)w(T) = v(T);$$

hence $x(T) \leq v(T)$. Since the opposite inequality also holds, we conclude that x belongs to the core of Γ_T. □

As a particular case of Theorem 2, the next result characterizes the core of the game itself.

Corollary 3. *Let Γ be as in Theorem 2 with $v(N) > 0$. For any $x \in \mathbb{R}_+^N \setminus \{0\}$, the following statements are equivalent:*

(1) x belongs to the core of Γ.
(2) $\mu(x) = 1$ and $\frac{x}{x(N)}$ is an optimal solution of $(\mathcal{P}_N)$.

Theorem 2 shows that each point belonging to the core of a subgame Γ_T induces an optimal solution of the associated optimization problem $(\mathcal{P}_T)$. In the opposite direction, we have

Corollary 4. *Let Γ and T be as in Theorem 2. For any $w \in \mathbb{R}_+^N \setminus \{0\}$, the following statements are equivalent:*

(1) w is an optimal solution of $(\mathcal{P}_T)$.
(2) $w(T) = 1$ and $\mu(w)w_T$ belongs to the core of Γ_T.

Proof. Let $x = \mu(w)w_T$. If (1) holds then $\overline{w} := \mu(w)w$ satisfies (3) of Theorem 2, hence (2) follows from the implication (3) $\Rightarrow$ (1) in that theorem. Conversely, if (2) holds then $\overline{w} := \mu(w)w$ satisfies $x = \overline{w}_T$ and $\mu(\overline{w}) = 1$; hence, by (1) $\Rightarrow$ (2) in Theorem 2, we obtain (1). □

The preceding results allow us to interpret Problem $(\mathcal{P}_T)$ in economic terms as a mathematical formulation of the following question: Given the total amount $w(T) = 1$ of the salaries received by the members of T, which amount of output $\mu(w)$ per unit of money spent the employer will obtain in the worst case (i.e., under the least favorable distribution of those salaries)? In other words, which is the guaranteed return per unit of money spent to the employer of an investment of which one money unit is assigned to paying salaries to the members of T? By Corollary 4, an optimal solution w of $(\mathcal{P}_T)$ satisfies $\mu(w) = \mu(w)w(T) = \mu(w)w_T(T) = v(T)$, so that the optimal value of $(\mathcal{P}_T)$ (i.e., the guaranteed return considered above) is precisely $v(T)$. Following (2) of Corollary 4, the optimal solution w gives us the weights according to which the payoff $\mu(w) = v(T)$ should be distributed among the members of T.

In view of Theorem 2 and Corollary 4, to compute the core of a (nontrivial) subgame Γ_T one can apply the following method: find all optimal solutions $\overline{w}$ to the problem $(\mathcal{P}_T)$; the elements in the core of Γ_T are just those of the form $\mu(\overline{w})\overline{w}_T$. Indeed, by Corollary 4, each $\mu(\overline{w})\overline{w}_T$ belongs to the core of Γ_T. Conversely, each element x in the core of Γ_T can be obtained in this way. To see this, take $\overline{w}$ as in (3) of Theorem 2. Then $\overline{w}/x(T)$ is an optimal solution of $(\mathcal{P}_T)$ and, as $\mu(\overline{w}) = 1$, one has

$$x = \overline{w}_T = \mu(\overline{w})\overline{w}_T = \mu\left(\frac{\overline{w}}{x(T)}\right)\frac{\overline{w}_T}{x(T)}.$$

One can illustrate this method by computing the core of the unanimity game $\Gamma^P = (N, v^P)$ associated to a nonempty coalition $P \subset N$, whose characteristic function is given by

$$v^P(S) = \begin{cases} 1 \text{ if } S \supset P \\ 0 \text{ otherwise.} \end{cases}$$

As shown in Martínez-Legaz [5], the MAV function μ^P of Γ^P is simply $\mu^P(w) = \dfrac{1}{w(P)}$. Therefore, the minimizers of $\mu^P(w)$ under the constraint $w(N) = 1$ are those $\overline{w} \in \mathbb{R}^n_+ \setminus \{0\}$ such that $\overline{w}(P) = 1$ and $\overline{w}_{N\setminus P} = 0$. Since these points satisfy $\mu^P(\overline{w}) = 1$, it follows that they are the core elements of Γ^P.

As a second example, consider the game $\Gamma = (N, v)$ with $N = \{1, 2, 3\}$ and v defined by

$$v(S) = \begin{cases} 0 & \text{if } S = \{i\} \\ 1 & \text{if } S = \{i, j\}, \text{ with } i \neq j, \\ \frac{3}{2} & \text{if } S = N. \end{cases}$$

As shown above, the MAV function μ of this game is as in (3). To find the core elements of Γ one therefore has to look for the minimizers of (3) under the constraints $w_1+w_2+w_3=1,\ w_i \geq 0 \quad (i=1,2,3)$. Since, by the first constraint, the right hand side of (3) is equal to $\frac{1}{1-\max\{w_1,w_2,w_3\}}$, this is equivalent to minimizing $\max\{w_1,w_2,w_3\}$ under the same constraints. This problem has a unique optimal solution, namely, the point $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$. As $\mu(\frac{1}{3},\frac{1}{3},\frac{1}{3})=\frac{3}{2}$, it turns out that the core of Γ is $\left\{\left(\frac{1}{2},\frac{1}{2},\frac{1}{2}\right)\right\}$.

To summarize, our results show that the computation of the core of any subgame of a nonnegative totally balanced game reduces to the minimization of a convex function (of nonnegative variables) under one linear constraint. Although this can be regarded as an easy problem, one should keep in mind that to use this method requires first computing the MAV function of the game, which is, in general, a hard task. So we do not claim that our method has any advantage upon the standard one consisting in solving the inequality system that defines the core (except when the MAV function is known or easy to compute); however, it allows one to express easily the core of a nontrivial nonnegative totally balanced game directly in terms of its MAV function. The importance of this fact lies in that the MAV function provides an alternative representation of the game, but such representation would not be of much use if one could not express standard concepts, like the core, in terms of it in an easy way.

4 Characterizing Convex Games in Terms of Their MAV Functions

A very important class of totally balanced games is that of convex games. One says that $\Gamma=(N,v)$ is convex if for every two coalitions S and T one has

$$v(S)+v(T)\leq v(S\cup T)+v(S\cap T).$$

The term "convex" is due to the property of "increasing returns" enjoyed by these games. Indeed, it is well-known that Γ is convex if and only if it satisfies

$$v(S\cup\{i\})-v(S)\leq v(T\cup\{i\})-v(T)$$

for each $i\in N$ and every coalitions S,T such that $S\subset T\subset N\backslash\{i\}$. An example of convex games is provided by unanimity games (see Section 3).

In this section we give a necessary and sufficient condition for a nonnegative totally balanced game to be convex, in terms of its MAV function. This condition will be based upon the following characterization of convex games, due to Einy and Shitovitz [1, Props. 3.8 and 4.2]:

Proposition 1. *Let Γ be a totally balanced TU game. The following statements are equivalent:*

(1) Γ is convex.
(2) For every $S,T \subset N$ with $S \subset T$ and every core element x of Γ_S there is a core element w of Γ_T such that $w_S = x$.

Theorem 3. *Let Γ be as in Theorem 2. The following statements are equivalent:*

(1) Γ is convex.
(2) For every $S,T \subset N$ with $S \subset T$ and $v(S) > 0$ and every optimal solution $\overline{w}$ of $(\mathcal{P}_S)$ there is an optimal solution $\overline{\overline{w}}$ of $(\mathcal{P}_T)$ and $\lambda > 0$ such that $\overline{\overline{w}}_S = \lambda \overline{w}_S$ and $\overline{\overline{w}}/\overline{\overline{w}}(S)$ is an optimal solution of $(\mathcal{P}_S)$.

Proof. Let us first recall that a totally balanced game is nonnegative if and only if it is monotonic (see Martínez-Legaz [5, Prop. 2.3]). Hence, if S, T and N are as in (2) then $v(T) > 0$.

(1) $\Longrightarrow$ (2). If $S \subset T \subset N$, $v(S) > 0$ and $\overline{w}$ is an optimal solution of $(\mathcal{P}_S)$ then, by Corollary 4, $\overline{w}(S) = 1$ and $\mu(\overline{w})\,\overline{w}_S$ belongs to the core of Γ_S. According to Proposition 1, there exists a core element w' of Γ_T such that $w'_S = \mu(\overline{w})\,\overline{w}_S$. By Theorem 2, there exists $w'' \in \mathbb{R}^N_+ \setminus \{0\}$ such that $w''_T = w'$, $\mu(w'') = 1$ and $w''/w'(T)$ is an optimal solution of $(\mathcal{P}_T)$. Then, for $\overline{\overline{w}} := w''/w'(T)$ and $\lambda := \mu(\overline{w})/w'(T)$, one has

$$\overline{\overline{w}}_S = \frac{w''_S}{w'(T)} = \frac{(w''_T)_S}{w'(T)} = \frac{w'_S}{w'(T)} = \frac{\mu(\overline{w})\,\overline{w}_S}{w'(T)} = \lambda \overline{w}_S.$$

Moreover $\overline{\overline{w}}/\overline{\overline{w}}(S)$ is an optimal solution of $(\mathcal{P}_S)$, since it is a feasible point and has the same objective function value as the optimal solution $\overline{w}$:

$$\begin{aligned}\mu\left(\frac{\overline{\overline{w}}}{\overline{\overline{w}}(S)}\right) = \mu\left(\frac{w''}{w''(S)}\right) &= w''(S)\,\mu(w'') = w''(S) = w''_T(S) = w'(S) \\ &= w'_S(S) = \mu(\overline{w})\,\overline{w}_S(S) = \mu(\overline{w})\,\overline{w}(S) = \mu(\overline{w}).\end{aligned}$$

(2) $\Longrightarrow$ (1). We shall prove that condition (2) above implies condition (2) of Proposition 1. Let $S \subset T \subset N$ and x be a core element of Γ_S. If $v(S) = 0$, from the monotonicity of v it follows that $x = 0$ and v vanishes at each subcoalition of S. Therefore one can easily check that, taking any core element $y = (y_i)_{i \in T \setminus S}$ of $\Gamma_{T \setminus S}$, the vector $w = (w_i)_{i \in T}$ defined by

$$w_i := 0 \text{ if } i \in S, \quad w_i := y_i + \frac{v(T) - v(T \setminus S)}{|T| - |S|} \text{ if } i \in T \setminus S,$$

belongs to the core of Γ_T and satisfies $w_S = 0 = x$. If $v(S) > 0$ then, by Theorem 2, there exists $\overline{w} \in \mathbb{R}^N_+ \setminus \{0\}$ such that $x = \overline{w}_S$, $\mu(\overline{w}) = 1$ and $\overline{w}/x(S)$ is an optimal solution of $(\mathcal{P}_S)$. According to condition (2), there are an optimal solution $\overline{\overline{w}}$ of $(\mathcal{P}_T)$ and $\lambda > 0$ such that $\overline{\overline{w}}_S = \lambda \overline{w}_S/x(S)$ and $\overline{\overline{w}}/\overline{\overline{w}}(S)$ is an optimal solution of $(\mathcal{P}_S)$; by $\overline{w}_S = x$, one has $\lambda = \overline{\overline{w}}_S(S) = \overline{\overline{w}}(S)$, so that

$$\overline{\overline{w}}_S = \frac{\overline{\overline{w}}(S)}{x(S)}\overline{w}_S = \frac{\overline{\overline{w}}(S)}{x(S)}x.$$

Let $w' := \frac{x(S)}{\overline{\overline{w}}(S)}\overline{\overline{w}}$. Since both $\frac{\overline{\overline{w}}}{\overline{\overline{w}}(S)}$ and $\frac{\overline{w}}{x(S)}$ are optimal solutions of $(\mathcal{P}_S)$, we have

$$\begin{aligned}\mu(w') &= \mu\left(\frac{x(S)}{\overline{\overline{w}}(S)}\overline{\overline{w}}\right) = \frac{1}{x(S)}\mu\left(\frac{\overline{\overline{w}}}{\overline{\overline{w}}(S)}\right)\\ &= \frac{1}{x(S)}\mu\left(\frac{\overline{w}}{x(S)}\right) = \mu(\overline{w}) = 1.\end{aligned}$$

On the other hand,

$$\frac{w'}{w'(T)} = \frac{\overline{\overline{w}}}{\overline{\overline{w}}(T)} = \overline{\overline{w}}$$

is an optimal solution of $(\mathcal{P}_T)$. Therefore, by Theorem 2, $w := w'_T$ belongs to the core of Γ_T; moreover, it satisfies

$$w_S = (w'_T)_S = w'_S = \frac{x(S)}{\overline{\overline{w}}(S)}\overline{\overline{w}}_S = x. \quad \square$$

5 Characterizing Convex Games in Terms of Their Marginals

Since totally balancedness is not a sufficient condition for a game $\Gamma = (N, v)$ to be convex, a natural question to ask is which additional conditions imposed on a totally balanced game ensure its convexity. The answer is given by the following theorem, which says that the required conditions are the totally balancedness of the marginal games as well. By the marginal game relative to coalition $T \subset N$, we mean the game $\Gamma'_T = (N\backslash T, v'_T)$ whose characteristic function is defined by $v'_T(S) = v(T \cup S) - v(T)$.

Theorem 4. *Let $\Gamma = (N, v)$ be a TU game. The following statements are equivalent:*

(1) Γ is convex.
(2) Γ'_T is convex for every $T \subset N$.
(3) Γ'_T is totally balanced for every $T \subset N$.
(4) Γ'_T is superadditive for every $T \subset N$.

Proof. To prove (1) $\Longrightarrow$ (2), let $T \subset N$ and $S_1, S_2 \subset N\backslash T$. Since Γ is convex, we have

$$\begin{aligned}v'_T(S_1) + v'_T(S_2) &= v(T \cup S_1) + v(T \cup S_2) - 2v(T) \leq\\ &\leq v(T \cup S_1 \cup S_2) + v(T \cup (S_1 \cap S_2)) - 2v(T) =\\ &= v'_T(S_1 \cup S_2) + v'_T(S_1 \cap S_2),\end{aligned}$$

which shows that v'_T is convex. Implications (2) $\Longrightarrow$ (3) $\Longrightarrow$ (4) follow from the well-known facts that all convex games are totally balanced and that the latter are superadditive. So, it only remains to prove (4) $\Longrightarrow$ (1); to this aim, it suffices to observe that, for each $S_1, S_2 \subset N$, one has

$$\begin{aligned} v(S_1) + v(S_2) &= v'_{S_1 \cap S_2}(S_1 \backslash S_2) + v'_{S_1 \cap S_2}(S_2 \backslash S_1) + 2v(S_1 \cap S_2) \leq \\ &\leq v'_{S_1 \cap S_2}\left((S_1 \cup S_2) \setminus (S_1 \cap S_2)\right) + 2v(S_1 \cap S_2) = \\ &= v(S_1 \cup S_2) + v(S_1 \cap S_2), \end{aligned}$$

where the inequality follows from the superadditivity of $v'_{S_1 \cap S_2}$. □

The equivalence between statements (1) and (4) of the preceding theorem was implicitly used in Martínez-Legaz [4] to prove Proposition 20 on a characterization of convex games in terms of indirect functions. Based on the equivalence (1) $\Longleftrightarrow$ (2), we will next present an alternative characterization of convex games, similar to that of totally balanced games in terms of balanced sets of coalitions (cf., e.g., Shapley and Shubik [8]). To this aim, we need to introduce the following notion:

Definition 1. *A collection $\mathcal{B}$ of subsets of $P \subset N$ is marginally $P-$balanced if $\bigcap_{S \in \mathcal{B}} S \notin \mathcal{B}$ and there exist positive weights $\{\gamma_S\}_{S \in \mathcal{B}}$ such that for each $i \in P \setminus \left(\bigcap_{S \in \mathcal{B}} S\right)$ one has $\sum_{\substack{S \in \mathcal{B} \\ S \ni i}} \gamma_S = 1$.*

Corollary 5. *A TU game $\Gamma = (N, v)$ is convex if and only if*

$$v(P) \geq \sum_{S \in \mathcal{B}} \gamma_S v(S) - \left(\sum_{S \in \mathcal{B}} \gamma_S - 1\right) v\left(\bigcap_{S \in \mathcal{B}} S\right)$$

for every $P \subset N$ and every marginally P-balanced collection $\mathcal{B}$ with weights $\{\gamma_S\}_{S \in \mathcal{B}}$.

Proof. The "only if" part follows from the totally balancedness of v'_T, with $T = \cap_{S \in \mathcal{B}} S$, and the fact that the marginal P-balancedness of $\mathcal{B}$ is equivalent to the balancedness of $\{S \backslash T\}_{S \in \mathcal{B}}$ as a collection of subsets of $P \backslash T$, associating to each $S \backslash T$ the weight γ_S. To prove the converse, given $S, T \subset N$ with $S \not\subset T$ and $T \not\subset S$, let $P = S \cup T$. Then $\{S, T\}$ is marginally P-balanced with $\gamma_S = \gamma_T = 1$. Thus, the assumed inequality reduces to

$$v(S \cup T) \geq v(S) + v(T) - v(S \cap T). \quad \square$$

The interest of Corollary 5 lies in that it allows for an easy comparison between convex games and totally balanced games. Notice that the condition stated in Corollary 5 reduces to that of totally balancedness when restricted to collections $\mathcal{B}$ having an empty intersection. Moreover, it admits the following interpretation. If a fraction γ_S of coalition S forms (in the sense, e.g., that

coalition S works during γ_S units of time), thus yielding an output $\gamma_S v(S)$, the total output that P can obtain is at least the sum of all these outputs minus that paid, by their extra effort, to the subcoalition consisting of those players who contributed $\sum_{S\in\mathcal{B}}\gamma_S$ (greater than 1) units of themselves (i.e., those players who worked during more than one unit of time). This payment is the output they would be able to obtain by themselves with this extra effort. Note that, as $\mathcal{B}$ is marginally P-balanced, the other players contribute exactly one unit of themselves.

Acknowledgments. I am grateful to Carles Rafels for his stimulating comments on convex games and for some useful information about the existing literature, and to Michael Maschler for suggestions on improvements and extensions of an earlier version. I also thank two anonymous referees for their careful reading and helpful corrections.

References

1. E. Einy and B. Shitovitz. Convex games and stable sets. Games and Economic Behavior16:192-201, 1996.
2. L.R. Ford and D.R. Fulkerson. Flows in Networks. Princeton University Press, Princeton, 1962.
3. E. Kalai and E. Zemel. Totally balanced games and games of flow. Math. Oper. Res., 7:476-478, 1982.
4. J.E. Martínez-Legaz. Dual representation of cooperative games based on Fenchel-Moreau conjugation. Optimization 36:291-319, 1996.
5. J.E. Martínez-Legaz. A duality theory for totally balanced games. In: C. García, C. Olivé, and M. Sanromà (eds.), Proc. IV Catalan Days of Applied Mathematics, Tarragona, Universitat Rovira i Virgili, pp. 151-161, 1998.
6. G. Owen. On the core of linear production games. Math. Programming 9:358-370, 1975.
7. R.T. Rockafellar. Convex Analysis. Princeton University Press, Princeton, 1970.
8. L.S. Shapley and M. Shubik. On market games. J. Economic Theory 1:9-25, 1969.
9. V.M. Tikhomirov. Convex Analysis. In: R.V. Gamkrelidze (ed.), Analysis II, Encyclopaedia of Mathematical Sciences, Volume 14, Berlin, Springer-Verlag, 1990.

The Bird Core for Minimum Cost Spanning Tree Problems Revisited: Monotonicity and Additivity Aspects

Stef Tijs[1], Stefano Moretti[2], Rodica Branzei[3], and Henk Norde[4]

[1] Department of Mathematics, University of Genoa, Italy and CentER and Department of Econometrics and Operations Research, Tilburg University, The Netherlands. S.h.tijs@uvt.nl
[2] Department of Mathematics, University of Genoa and Unit of Molecular Epidemiology, National Cancer Research Institute of Genoa, Italy. moretti@dima.unige.it
[3] Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iasi, Romania. branzeir@info.uaic.ro
[4] CentER and Department of Econometrics and Operations Research, Tilburg University, The Netherlands. h.norde@uvt.nl

Summary. A new way is presented to define for minimum cost spanning tree (mcst) games the irreducible core, which is introduced by Bird in 1976. The Bird core correspondence turns out to have interesting monotonicity and additivity properties and each stable cost monotonic allocation rule for mcst-problems is a selection of the Bird core correspondence. Using the additivity property an axiomatic characterization of the Bird core correspondence is obtained.

1 Introduction

One of the classical problems in Operations Research is the problem of finding a minimum cost spanning tree (mcst) in a connected network. For algorithms solving this problem see [16] and [20]. Claus and Kleitman [8] discuss the problem of allocating costs among users in a minimum cost spanning tree. This inspired independently [3] and [13] to construct and use a cooperative game to tackle this cost allocation problem.

In the seminal paper of Bird [3] a method is indicated how to find a core element of the minimum cost spanning tree game (mcst game) when a minimum cost spanning tree is given. Further he has introduced, using a fixed mcst, the irreducible core of an mcst game, which is a subset of the core of the game, and which we will call in this paper the Bird core. The Bird core is central in this paper. First, we will give a new "tree free" way to introduce the Bird core by constructing for each mcst-problem a related

problem, where the weight function is a non-Archimedean semimetric. The Bird core correspondence turns out to be a crucial correspondence if one is interested in stable cost monotonic allocation rules for mcst-problems. In fact, the Bird core is the "largest" among the correspondences which are cost monotonic and stable.

The question of the existence of cost allocation rules which are cost monotonic is central in applied economic frameworks where connection costs may increase or decrease in time. In the paper of Dutta and Kar [10], cost monotonic allocation rules have been studied, where cost monotonicity means that an agent i does not pay more if the cost of a link involving i goes down, nothing else changing in the network.

Actually, our concept of cost monotonicity is stronger than the concept of cost monotonicity introduced in [10], because we simply impose that if some connection costs go down, then no agents will pay more (as in the strong cost monotonicity property used by [2]). Moreover, we introduce a related concept of cost monotonicity for multisolutions in mcst situations which generalize our concept of cost monotonicity for mcst solutions.

The Bird core has also an interesting additivity property i.e. the Bird core correspondence is additive on each Kruskal cone in the space of mcst-problems with a fixed number of users. The additivity on Kruskal cones can be used to find an axiomatic characterization of the Bird core correspondence.

The outline of the paper is as follows. Section 2 settles notions and notations. In Section 3 the non-Archimedean semimetric is introduced and used to define in a canonical (tree independent) way the reduced game and the Bird core. The relations between stable cost monotonic rules and the Bird core are discussed in Section 4. An axiomatic characterization of the Bird core is given in Section 5. Section 6 concludes.

2 Preliminaries and Notations

An (undirected) *graph* is a pair $< V, E >$, where V is a set of vertices or nodes and E is a set of edges e of the form $\{i, j\}$ with $i, j \in V$, $i \neq j$. The *complete graph* on a set V of vertices is the graph $< V, E_V >$, where $E_V = \{\{i, j\} | i, j \in V \text{ and } i \neq j\}$. A *path* between i and j in a graph $< V, E >$ is a sequence of nodes $(i_0, i_1, \ldots, i_k)$, where $i = i_0$ and $j = i_k$, $k \geq 1$, and such that $\{i_s, i_{s+1}\} \in E$ for each $s \in \{0, \ldots, k-1\}$. A *cycle* in $< V, E >$ is a path with all distinct edges from i to i for some $i \in V$. A path $(i_0, i_1, \ldots, i_k)$ is *without cycles* if there do not exist $a, b \in \{0, 1, \ldots, k\}$, $a \neq b$, such that $i_a = i_b$.

Two nodes $i, j \in V$ are connected in $< V, E >$ if $i = j$ or if there exists a path between i and j in $< V, E >$. A *connected component* of V in $< V, E >$ is a maximal subset of V with the property that any two nodes in this subset are connected in $< V, E >$. Given a path $P = (i_0, i_1, \ldots, i_k)$ between i and j in a graph $< V, E >$, $k \geq 1$, we say that $v \in V$ is a node in P if $v = i_m$

for some $m \in \{0, \ldots, k\}$; we say that an edge $\{r, t\} \in E$ is on the path P or, equivalently, that i is connected to j via the edge $\{r, t\}$ in the path P, if there exists $m \in \{0, \ldots, k-1\}$ such that $r = i_m$ and $t = i_{m+1}$ or $t = i_m$ and $r = i_{m+1}$.

Now, we consider *minimum cost spanning tree* (mcst) *situations.* In an mcst situation a set $N = \{1, \ldots, n\}$ of agents is involved willing to be connected as cheap as possible to a source (i.e. a supplier of a service) denoted by 0. In the sequel we use the notation N' for $N \cup \{0\}$. An mcst situation can be represented by a tuple $< N', E_{N'}, w >$, where $< N', E_{N'} >$ is the complete graph on the set N' of nodes or vertices, and $w : E_{N'} \to \mathbb{R}_+$ is a map which assigns to each edge $e \in E_{N'}$ a nonnegative number $w(e)$ representing the weight or cost of edge e. We call w a *weight function.* If $w(e) \in \{0, 1\}$ for every $e \in E_{N'}$, the weight function w is called a *simple weight function*, and we refer then to $< N', E_{N'}, w >$ as a *simple mcst situation.* Since in our paper the graph of possible edges is always the complete graph, we simply denote an mcst situation with the set of users N, source 0, and weight function w by $< N', w >$. Often we identify an mcst situation $< N', w >$ with the corresponding weight function w. We denote by $\mathcal{W}^{N'}$ the set of all mcst situations $< N', w >$ (or w) with node set N'. For each $S \subseteq N$ one can consider the mcst subsituation $< S', w_{|S'} >$, where $S' = S \cup \{0\}$ and $w_{|S'} : E_{S'} \to \mathbb{R}_+$ is the restriction of the weight function w to $E_{S'} \subseteq E_{N'}$, i.e. $w_{|S'}(e) = w(e)$ for each $e \in E_{S'}$.

Let $< N', w >$ be an mcst situation. Two nodes i and j are called (w, N')-*connected* if $i = j$ or if there exists a path $(i_0, \ldots, i_k)$ from i to j, with $w(\{i_s, i_{s+1}\}) = 0$ for every $s \in \{0, \ldots, k-1\}$. A (w, N')-*component* of N' is a maximal subset of N' with the property that any two nodes in this subset are (w, N')-connected. We denote by $C_i(w)$ the (w, N')-component to which i belongs and by $\mathcal{C}(w)$ the set of all the (w, N')-components of N'. Clearly, the collection of (w, N')-components forms a partition of N'.

We define the set $\Sigma_{E_{N'}}$ of *linear orders* on $E_{N'}$ as the set of all bijections $\sigma : \{1, \ldots, |E_{N'}|\} \to E_{N'}$, where $|E_{N'}|$ is the cardinality of the set $E_{N'}$. For each mcst situation $< N', w >$ there exists at least one linear order $\sigma \in \Sigma_{E_{N'}}$ such that $w(\sigma(1)) \leq w(\sigma(2)) \leq \ldots \leq w(\sigma(|E_{N'}|))$. We denote by w^σ the column vector $\big(w(\sigma(1)), w(\sigma(2)), \ldots, w(\sigma(|E_{N'}|))\big)^t$.

For any $\sigma \in \Sigma_{E_{N'}}$ we define the set

$$K^\sigma = \{w \in \mathbb{R}_+^{E_{N'}} \mid w(\sigma(1)) \leq w(\sigma(2)) \leq \ldots \leq w(\sigma(|E_{N'}|))\},$$

which we call the *Kruskal cone with respect to* σ. One can easily see that $\bigcup_{\sigma \in \Sigma_{E_{N'}}} K^\sigma = \mathbb{R}_+^{E_{N'}}$. For each $\sigma \in \Sigma_{E_{N'}}$ the cone K^σ is a simplicial cone with generators $e^{\sigma,k} \in K^\sigma$, $k \in \{1, 2, \ldots, |E_{N'}|\}$, where

$$\begin{gathered} e^{\sigma,k}(\sigma(1)) = e^{\sigma,k}(\sigma(2)) = \ldots = e^{\sigma,k}(\sigma(k-1)) = 0 \\ \text{and} \\ e^{\sigma,k}(\sigma(k)) = e^{\sigma,k}(\sigma(k+1)) = \ldots = e^{\sigma,k}(\sigma(|E_{N'}|)) = 1. \end{gathered} \tag{1}$$

[Note that $e^{\sigma,1}(\sigma(k)) = 1$ for all $k \in \{1, 2, \ldots, |E_{N'}|\}$]. This implies that each $w \in K^\sigma$ can be written in a unique way as non-negative linear combination of these generators. To be more concrete, for $w \in K^\sigma$ we have

$$w = w(\sigma(1))e^{\sigma,1} + \sum_{k=2}^{|E_{N'}|} \left(w(\sigma(k)) - w(\sigma(k-1))\right) e^{\sigma,k}. \tag{2}$$

Clearly, we can also write $\mathcal{W}^{N'} = \bigcup_{\sigma \in \Sigma_{E_{N'}}} K^\sigma$, if we identify an mcst situation $< N', w >$ with w.

Any mcst situation $w \in \mathcal{W}^{N'}$ gives rise to two problems: the construction of a network $\Gamma \subseteq E_{N'}$ of minimal cost connecting all users to the source, and a cost sharing problem of distributing this cost in a fair way among users. The cost of a network Γ is $w(\Gamma) = \sum_{e \in \Gamma} w(e)$. A network Γ is a *spanning network* on $S' \subseteq N'$ if for every $e \in \Gamma$ we have $e \in E_{S'}$ and for every $i \in S$ there is a path in Γ from i to the source. Given a spanning network Γ on N' we define the set of edges of Γ with nodes in $S' \subseteq N'$ as the set $E^\Gamma_{S'} = \{\{i,j\} | \{i,j\} \in \Gamma \text{ and } i,j \in S'\}$.

For any mcst situation $w \in \mathcal{W}^{N'}$ it is possible to determine at least one *spanning tree* on N', i.e. a spanning network without cycles on N', of minimum cost; each spanning tree of minimum cost is called an mcst for N' in w or, shorter, an mcst for w. Two famous algorithms for the determination of minimum cost spanning trees are the algorithm of Prim ([20]) and the algorithm of Kruskal ([16]). The cost of a minimum cost spanning network Γ on N' in a simple mcst situation w equals $|\mathcal{C}(w)| - 1$ (see Lemma 2 in [19]).

Now, let us introduce some basic game theoretical notations. A *cooperative cost game* is a pair (N, c) where $N = \{1, \ldots, n\}$ is a finite (*player-*)set and the *characteristic function* $c : 2^N \to \mathbb{R}$ assigns to each subset $S \in 2^N$, called a *coalition*, a real number $c(S)$, called the *cost of coalition* S, where 2^N stands for the power set of the player set N, and $c(\emptyset) = 0$. The *core* of a game (N, c) is the set of payoff vectors for which no coalition has an incentive to leave the grand coalition N, i.e.

$$\mathcal{C}(c) = \{x \in \mathbb{R}^N | \sum_{i \in S} x_i \leq c(S) \ \forall S \in 2^N \setminus \{\emptyset\}; \sum_{i \in N} x_i = c(N)\}.$$

Note that the core of a game can be empty. A game (N, c) is called a *concave game* if the marginal contribution of any player to any coalition is more than his marginal contribution to a larger coalition, i.e. if it holds that

$$c(S \cup \{i\}) - c(S) \geq c(T \cup \{i\}) - c(T) \tag{3}$$

for all $i \in N$ and all $S \subseteq T \subseteq N \setminus \{i\}$.

An *order* τ of N is a bijection $\tau : \{1, \ldots, |N|\} \to N$. This order is denoted by $\tau(1), \ldots, \tau(n)$, where $\tau(i) = j$ means that with respect to τ, player j is in the i-th position. We denote by Σ_N the set of possible orders on the set N.

Let (N, c) be a cooperative cost game. For $\tau \in \Sigma_N$, the *marginal vector* $m^\tau(c)$ is defined by

$$m_i^\tau(c) = c([i,\tau]) - c((i,\tau)) \text{ for all } i \in N,$$

where $[i,\tau] = \{j \in N : \tau^{-1}(j) \leq \tau^{-1}(i)\}$ is the set of predecessors of i with respect to τ including i, and $(i,\tau) = \{j \in N : \tau^{-1}(j) < \tau^{-1}(i)\}$ is the set of predecessors of i with respect to τ excluding i. In a coherent way with respect to previous notations, we will indicate the set $[i,\tau] \cup \{0\}$ and $(i,\tau) \cup \{0\}$ as $[i,\tau]'$ and $(i,\tau)'$, respectively. For instance, for each $k \in \{1,\ldots,|N|\}$ and for each $l \in \{2,\ldots,|N|\}$, the set $[\tau(k),\tau]' = \{0,\tau(1),\ldots,\tau(k)\}$ and $(\tau(l),\tau)' = \{0,\tau(1),\ldots,\tau(l-1)\}$, which will be denoted shorter as $[\tau(k)]'$ and $(\tau(l))'$, respectively.

Let $< N', w >$ be an mcst situation. The *minimum cost spanning tree game* (N, c_w) (or simply c_w), corresponding to $< N', w >$, is defined by

$$c_w(S) = \min\{w(\Gamma) | \Gamma \text{ is a spanning network on } S'\}$$

for every $S \in 2^N \backslash \{\emptyset\}$, with the convention that $c_w(\emptyset) = 0$.

We denote by $\mathcal{MCST}^N$ the class of all mcst games corresponding to mcst situations in $\mathcal{W}^{N'}$. For each $\sigma \in \Sigma_{E_{N'}}$, we denote by $\mathcal{G}^\sigma$ the set $\{c_w \mid w \in K^\sigma\}$ which is a cone. We can express $\mathcal{MCST}^N$ as the union of all cones $\mathcal{G}^\sigma$, i.e. $\mathcal{MCST}^N = \bigcup_{\sigma \in \Sigma_{E_{N'}}} \mathcal{G}^\sigma$, and we would like to point out that $\mathcal{MCST}^N$ itself is not a cone if $|N| \geq 2$.

The core $\mathcal{C}(c_w)$ of an mcst game $c_w \in \mathcal{MCST}^N$ is nonempty ([14], [3]) and, given an mcst Γ (with no cycles) for N' in mcst situation w, one can easily find an element in the core looking at the *Bird allocation* in w corresponding to Γ, i.e. the cost allocation where each player $i \in N$ pays the edge in Γ which connects him with his *immediate predecessor* in $< N', \Gamma >$.

We call a map $F : \mathcal{W}^{N'} \to \mathbb{R}^N$ assigning to every mcst situation w a unique cost allocation in $\mathbb{R}^N$ a *solution*. A solution F is *efficient* if for each $w \in \mathcal{W}^{N'}$

$$\sum_{i \in N} F_i(w) = w(\Gamma),$$

where Γ is a minimum cost spanning network on N' for w.

3 The Non-Archimedean Semimetric Corresponding to an MCST Situation

Let $w \in \mathcal{W}^{N'}$. For each path $P = (i_0, i_1, \ldots, i_k)$ from i to j in the graph $< N', E_{N'} >$ we denote the set of its edges by $E(P)$, that is $E(P) = \{\{i_0, i_1\}, \{i_1, i_2\}, \ldots, \{i_{k-1}, i_k\}\}$. Moreover, we call $\max_{e \in E(P)} w(e)$ the *top of the path* P and denote it by $t(P)$. We denote by $\mathcal{P}_{ij}^{N'}$ the set of all paths without cycles from i to j in the graph $< N', E_{N'} >$.

Now we define the key concept of this section, namely the reduced weight function.

Definition 1. *Let $w \in \mathcal{W}^{N'}$. The* reduced weight function $\bar{w}$ *is given by*

$$\bar{w}(i,j) = \min_{P \in \mathcal{P}_{ij}^{N'}} \max_{e \in E(P)} w(e) = \min_{P \in \mathcal{P}_{ij}^{N'}} t(P) \tag{4}$$

for each $i, j \in N'$, $i \neq j$.

Now, extending $\bar{w}$ by putting $\bar{w}(i,i) = 0$ for each $i \in N'$, we obtain a nonnegative function on the set of all pairs of elements in N'. The obtained reduced weight function $\bar{w}$ is a semimetric on N' with the sharp triangle inequality, i.e. a *non-Archimedean* (NA-)*semimetric*. In formula, for each $i, j, k \in N'$

$\bar{w}(i,j) \geq 0$ and $\bar{w}(i,i) = 0$ (non-negativity);
$\bar{w}(i,j) = \bar{w}(j,i)$ (symmetry);
$\bar{w}(i,k) \leq \max\{\bar{w}(i,j), \bar{w}(j,k)\}$ (sharp triangle inequality).

The proof is left to the reader. If $w > 0$, then $\bar{w}$ is a non-Archimedean metric on the set N'.

For the reduced weight function $\bar{w}$ we have a special property related to triangles (the isoscele triangle property), as the next proposition shows.

Proposition 1. *Let $\bar{w}$ be the reduced weight function corresponding to $w \in \mathcal{W}^{N'}$ and $i, j, k \in N'$ such that $\bar{w}(i,j) \leq \bar{w}(i,k)$ and $\bar{w}(i,j) \leq \bar{w}(k,j)$. Then $\bar{w}(i,k) = \bar{w}(j,k)$.*

Proof. By the sharp triangle inequality $\bar{w}(i,k) \leq \max\{\bar{w}(i,j), \bar{w}(j,k)\} = \bar{w}(j,k)$ and $\bar{w}(j,k) \leq \max\{\bar{w}(j,i), \bar{w}(i,k)\} = \bar{w}(i,k)$. So $\bar{w}(i,k) = \bar{w}(j,k)$. □

This property for NA-semimetrics will be useful in proving that there are many minimum cost spanning trees for $(N', \bar{w})$, as we see in Theorem 1.

In the sequel we simply refer to $\bar{w}$ as the mcst situation which assigns to each edge $\{i,j\} \in E_{N'}$ the reduced weight value as defined in equality (4). Further, we will denote by $\bar{\mathcal{W}}^{N'} \subset \mathcal{W}^{N'}$ the set of all NA-semimetric mcst situations which assign to each edge $\{i,j\} \in E_{N'}$ the distance $\bar{w}(i,j)$ provided by a NA-semimetric $\bar{w}$ on N'.

Example 1. Consider the mcst situation $< N', w >$ with $N' = \{0,1,2,3\}$ and w as depicted in Figure 1. Note that $w \in K^{\sigma}$, with $\sigma(1) = \{1,2\}$, $\sigma(2) = \{1,0\}$, $\sigma(3) = \{1,3\}$, $\sigma(4) = \{3,0\}$, $\sigma(5) = \{2,0\}$, $\sigma(6) = \{2,3\}$. The corresponding mcst situation $\bar{w}$ is depicted in Figure 2.

One main result in this section, Proposition 2, concerns an interesting relation which can be established between the mcst situation $\bar{w}$ and a *minimal mcst situation* w^{Γ} as defined by Bird [3], where Γ is an mcst for N' in w. Recall

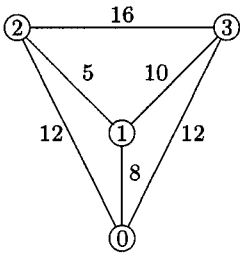


Fig. 1. An mcst situation with three agents.

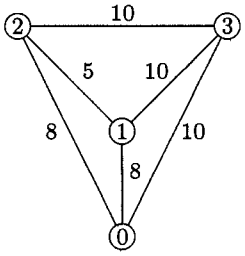


Fig. 2. The mcst situation $\bar{w}$ corresponding to w.

that given an mcst situation $w \in \mathcal{W}^{N'}$ and an mcst Γ for N' in w, the minimal mcst situation w^Γ is defined (cf. Bird, 1976) by

$$w^\Gamma(\{i,j\}) = \max_{e \in P_{ij}^\Gamma} w(e) = t(P_{ij}^\Gamma), \tag{5}$$

where $P_{ij}^\Gamma \in \mathcal{P}_{ij}^{N'}$ is the unique path in Γ from i to j.

Proposition 2. *Let $w \in \mathcal{W}^{N'}$ and $i, j \in N'$. Let Γ be an mcst for N' in w and P_{ij}^Γ be the unique path in Γ from i to j. Then*

$$t(P_{ij}^\Gamma) = \min_{P \in \mathcal{P}_{ij}^{N'}} t(P). \tag{6}$$

Proof. Let $P^* \in \arg\min_{P \in \mathcal{P}_{ij}^{N'}} t(P)$ and let e^* be an edge on P^* such that $t(P^*) = w(e^*)$. Let $\hat{e} = \{m, n\}$ be an edge on P_{ij}^Γ with $w(\hat{e}) = t(P_{ij}^\Gamma)$. We have to prove that $w(\hat{e}) = w(e^*)$. If so, then it follows immediately that $\min_{P \in \mathcal{P}_{ij}^{N'}} t(P) = w(e^*) = w(\hat{e}) = t(P_{ij}^\Gamma)$. If $e^* = \hat{e}$ then of course $w(e^*) = w(\hat{e})$. Otherwise, first note that by definition of e^*

$$w(\hat{e}) \geq w(e^*). \tag{7}$$

Let S_m be the set of all nodes $r \in N'$ such that n is not on the path from r to m in $< N', \Gamma >$; let S_n be the set of nodes $r \in N'$ such that m is not on the path from r to n in $< N', \Gamma >$, i.e.

$$S_m = \{r \in N' | n \notin P^{\Gamma}_{mr}\} \quad \text{and} \quad S_n = \{r \in N' | m \notin P^{\Gamma}_{nr}\}.$$

Note that $\{S_n, S_m\}$ is a partition of N' and nodes in S_n are connected in $< N', \Gamma >$ to nodes in S_m via edge $\{m, n\}$. Moreover, by the definition of a path without cycles, i, j must belong to different sets of the partition $\{S_n, S_m\}$. So without loss of generality we suppose that $i \in S_m$ and $j \in S_n$. Consider the set of edges $E^+ = \{\{t, v\} | t \in S_m, v \in S_n\}$. Then,

$$w(\{m, n\}) = w(\hat{e}) \leq w(e), \text{ for each } e \in E^+. \tag{8}$$

In order to prove inequality (8), suppose on the contrary that $w(\{m, n\}) > w(e)$ for some $e \in E^+$. Then the graph $\Gamma^+ = (\Gamma \setminus \{\hat{e}\}) \cup \{e\}$ would be a spanning network in N' cheaper than Γ, which yields a contradiction. By the definition of a path, for each $P \in \mathcal{P}^{N'}_{ij}$ there exists at least one edge $e \in E^+$ such that e is on the path P. By inequality (8), it follows that $t(P) \geq w(e) \geq w(\hat{e})$. This implies that $w(e^*) = \min_{P \in \mathcal{P}^{N'}_{ij}} t(P) \geq w(\hat{e})$. Together with inequality (7) we have finally $w(e^*) = w(\hat{e})$. □

As a direct consequence of Proposition 2 we have that the mcst situation $\bar{w}$ coincides, for each mcst Γ for w, with the minimal mcst situation w^{Γ} introduced by [3]. So $w^{\Gamma} = w^{\Gamma'}$ for each pair of mcst Γ, Γ', a fact which is already known (cf. [1, 11, 12]), but with a complicated proof. Let $w \in \mathcal{W}^{N'}$ and let Γ be an mcst for w. Let $\tau \in \Sigma_N$. We say that Γ and τ *fit* (or, also, that τ fits with Γ) if $E^{\Gamma}_{[\tau(1)]'}$, $E^{\Gamma}_{[\tau(2)]'}, \ldots, E^{\Gamma}_{[\tau(|N|)]'}$ are spanning networks on sets of nodes $[\tau(1)]'$, $[\tau(2)]', \ldots, [\tau(|N|)]'$, respectively.

Example 2. In Figure 3 is depicted an mcst, denoted by Γ, for the mcst situation $\bar{w}$ of Figure 2. Consider $\tau_1, \tau_2 \in \Sigma_N$ such that $\tau_1(1) = 1$, $\tau_1(2) = 2$, $\tau_1(3) = 3$ and $\tau_2(1) = 1$, $\tau_2(2) = 3$, $\tau_2(3) = 2$. Note that both τ_1 and τ_2 fit with Γ but none of the other four elements of Σ_N fits with Γ.

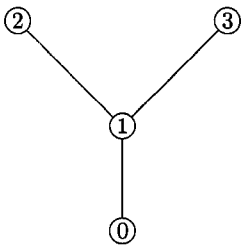


Fig. 3. An mcst Γ for the mcst situation $\bar{w}$ of Figure 2.

Remark 1. Let $w \in \mathcal{W}^{N'}$, let Γ be an mcst for w and let $\tau \in \Sigma_N$ be an order such that Γ and τ fit. Then,

$$\sum_{e \in E^{\Gamma}_{[\tau(r)]'}} w(e) = c_w([\tau(r)]) \tag{9}$$

for each $r \in \{1, \ldots, |N|\}$. So $E^{\Gamma}_{[\tau(r)]'}$ is an mcst for the mcst situation $< [\tau(r)]', w_{|[\tau(r)]'} >$.

Remark 2. Let $w \in \mathcal{W}^{N'}$, let Γ be an mcst for w and let $\tau \in \Sigma_N$ be an order such that Γ and τ fit. The marginal vector $m^{\tau}(c_w)$ of the mcst game c_w coincides with the Bird allocation in w corresponding to Γ and therefore $m^{\tau}(c_w) \in \mathcal{C}(c_w)$, as is proved in [14].

Remark 3. For each $\sigma \in \Sigma_{E_{N'}}$ there exists a tree Γ which is an mcst for every $w \in K^{\sigma}$; further, there exists a $\tau \in \Sigma_N$ such that Γ and τ fit.

The considerations in Remarks 1-3 together with the next lemma prelude to Theorem 1.

Lemma 1. *Let $w \in \bar{\mathcal{W}}^{N'}$, let Γ be an mcst for w and let $\tau \in \Sigma_N$ be such that Γ and τ fit. Let $r \in \{1, \ldots, |N| - 1\}$ and let $\tau' \in \Sigma_N$ be such that $\tau'(r) = \tau(r+1)$, $\tau'(r+1) = \tau(r)$ and $\tau'(i) = \tau(i)$ for each $i \in \{1, \ldots, |N|\} \setminus \{r, r+1\}$ (i.e. τ' is obtained from τ by a neighbor switch of $\tau(r)$ and $\tau(r+1)$). Then there is an mcst Γ' for w such that τ' and Γ' fit.*

Proof. If $\tau(r)$ is not the immediate predecessor of $\tau(r+1)$ in Γ then take $\Gamma' = \Gamma$ and then τ' and Γ' fit. If $\tau(r)$ is the immediate predecessor of $\tau(r+1)$ in Γ, then let $k \in [\tau(r-1)]'$ be the immediate predecessor of $\tau(r)$ in Γ. First, note that

$$w(\{k, \tau(r+1)\}) \geq w(\{k, \tau(r)\}) \tag{10}$$

and

$$w(\{k, \tau(r+1)\}) \geq w(\{\tau(r), \tau(r+1)\}) \tag{11}$$

because Γ is an mcst for w. Consider two cases:

Case 1: $w(\{k, \tau(r)\}) \leq w(\{\tau(r), \tau(r+1)\})$. Take $\Gamma' = (\Gamma \setminus \{\{\tau(r), \tau(r+1)\}\}) \cup \{\{k, \tau(r+1)\}\}$. By inequality (10) and the isosceles triangle property $w(\{k, \tau(r+1)\}) = w(\{\tau(r), \tau(r+1)\})$ and then Γ' is an mcst in w and Γ' and τ' fit.

Case 2: $w(\{\tau(r), \tau(r+1)\}) < w(\{k, \tau(r)\})$. Take $\Gamma' = (\Gamma \setminus \{\{k, \tau(r)\}\}) \cup \{\{k, \tau(r+1)\}\}$. By inequality (11) and the isosceles triangle property $w(\{k, \tau(r)\}) = w(\{k, \tau(r+1)\})$ and then Γ' is an mcst in w and Γ' and τ' fit. □

Theorem 1. *Let $w \in \bar{\mathcal{W}}^{N'}$. Then*

i) for each $\tau \in \Sigma_N$ there exists an mcst Γ such that Γ and τ fit.

ii) Let c_w be the mcst game corresponding to w. Then $m^{\tau}(c_w) \in \mathcal{C}(c_w)$ for all $\tau \in \Sigma_N$ and c_w is a concave game.

Proof. i) Let $\hat{\Gamma}$ be an mcst for w. Then there is at least one $\hat{\tau} \in \Sigma_N$ such that $\hat{\Gamma}$ and $\hat{\tau}$ fit. Further each τ can be obtained from $\hat{\tau}$ by a suitable sequence of neighbor switches and so, by applying Lemma 1 repeatedly, we complete the proof of assertion i).

ii) Let Γ be an mcst in N' for w and let $\tau \in \Sigma_N$ such that Γ and τ fit. By Remark 2, it follows that $m^\tau(c_w)$ coincides with the Bird allocation corresponding to Γ. Hence, again by Remark 2, $m^\tau(c_w) \in \mathcal{C}(c_w)$. Finally, by the Ichiishi theorem (Ichiishi (1981)) telling that a game is concave iff all marginal vectors are in the core of the game, it follows that c_w is a concave game. □

Let $w \in \mathcal{W}^{N'}$. We call the core of the mcst game $c_{\bar{w}}$ the Bird core of the mcst game c_w and denote it by $\mathcal{BC}(w)$. By Theorem 1 it directly follows that the Bird core $\mathcal{BC}(w)$ of the mcst game c_w is the convex hull of all the Bird allocations corresponding to the minimum cost spanning trees for $\bar{w}$. Note also that $\mathcal{BC}(w) \subseteq \mathcal{C}(c_w)$, since $c_{\bar{w}}(S) \leq c_w(S)$ for each $S \in 2^N \setminus \{\emptyset\}$ and $c_{\bar{w}}(N) = c_w(N)$ (cf. [11]).

Example 3. Consider the mcst situation w of Figure 1 and the corresponding reduced mcst situation $\bar{w}$ of Figure 2. Then

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,3\}$	$\{1,2,3\}$
c_w	8	12	12	13	24	18	23
$c_{\bar{w}}$	8	8	10	13	18	18	23

There are six minimum cost spanning trees for $\bar{w}$. Three of them lead to the Bird allocation $(8, 5, 10)$ and the other three to the Bird allocation $(5, 8, 10)$. Further, $m^\tau(c_{\bar{w}}) = (8, 5, 10)$ for $\tau \in \{(1,2,3), (1,3,2), (3,1,2)\}$ and $m^\tau(c_{\bar{w}}) = (5, 8, 10)$ for $\tau \in \{(2,1,3), (2,3,1), (3,2,1)\}$. The Bird core $\mathcal{BC}(w)$ is the convex hull of the marginal vectors of the game $c_{\bar{w}}$, that is $\mathcal{BC}(w) = \text{conv}\{(8, 5, 10), (5, 8, 10)\} \subset \mathcal{C}(c_w)$.

4 Monotonicity Properties

In [23] a class of solutions for mcst situations which are cost monotonic is introduced: the class of Obligation rules. Roughly speaking, we define a cost monotonic solution for mcst situations as a solution such that, if the costs of some edges increase, then no agent will pay less. More precisely:

Definition 2. *A solution $F : \mathcal{W}^{N'} \to \mathbb{R}^N$ is a* cost monotonic *solution if for all mcst situations $w, w' \in \mathcal{W}^{N'}$ such that $w(e) \leq w'(e)$ for each $e \in E_{N'}$, it holds that $F(w) \leq F(w')$.*

In this section we introduce a related concept of cost monotonicity for multisolutions on mcst situations. We call a correspondence $G : \mathcal{W}^{N'} \twoheadrightarrow \mathbb{R}^N$ assigning to every mcst situation w a set of cost allocations in $\mathbb{R}^N$ a *multisolution.*

Definition 3. *A multisolution* $M : \mathcal{W}^{N'} \twoheadrightarrow \mathbb{R}^N$ *is a* cost monotonic *multisolution if for all mcst situations* $w, w' \in \mathcal{W}^{N'}$ *such that* $w(e) \leq w'(e)$ *for each* $e \in E_{N'}$, *it holds that*

$$M(w) \subseteq compr^-(M(w')) \quad \text{and} \quad M(w') \subseteq compr^+(M(w)),$$

where $compr^-(B) = \{x \in \mathbb{R}^N | \exists b \in B \text{ s.t. } x_i \leq b_i \ \forall i \in N\}$ *and* $compr^+(B) = \{x \in \mathbb{R}^N | \exists b \in B \text{ s.t. } b_i \leq x_i \ \forall i \in N\}$, *for each* $B \subset \mathbb{R}^N$.

Before discussing properties of the Bird core as multisolution for mcst situations, we introduce the following propositions dealing with mcst situations originated by NA-semimetrics.

Proposition 3. *Let* $w \in \bar{\mathcal{W}}^{N'}$ *and let* Γ *be an mcst for* w *and* $\tau \in \Sigma_N$ *be such that* Γ *and* τ *fit. Then*

$$m^\tau_{\tau(j)}(c_w) = \min_{k \in (\tau(j))'} w(k, \tau(j)),$$

for each $j \in \{2, \ldots, |N|\}$.

Proof. Let $j \in \{2, \ldots, |N|\}$. Note that by Remark 1

$$m^\tau_{\tau(j)}(c_w) = c_w([\tau(j)]) - c_w((\tau(j))) = \sum_{e \in E^\Gamma_{[\tau(j)]'}} w(e) - \sum_{e \in E^\Gamma_{(\tau(j))'}} w(e). \quad (12)$$

Since Γ and τ fit, we have $E^\Gamma_{[\tau(j)]'} \setminus E^\Gamma_{(\tau(j))'} = \{\{\tau(j), s\}\}$, for some $s \in (\tau(j))'$. Because $E^\Gamma_{[\tau(j)]'}$ is an mcst for $w_{|[\tau(j)]'}$, we have $s \in \arg\min_{k \in (\tau(j))'} w(\{k, \tau(j)\})$. So

$$\sum_{e \in E^\Gamma_{[\tau(j)]'}} w(e) - \sum_{e \in E^\Gamma_{(\tau(j))'}} w(e) = \min_{k \in (\tau(j))'} w(k, \tau(j)). \quad (13)$$

From (12) and (13) follows the proposition. □

Proposition 4. *Let* $w, w' \in \bar{\mathcal{W}}^{N'}$ *be NA-semimetric mcst situations such that* $w(e) \leq w'(e)$ *for each* $e \in E_{N'}$. *Then it holds that*

$$m^\tau(c_w) \leq m^\tau(c_{w'}) \text{ for each } \tau \in \Sigma_N.$$

Proof. Let $\tau \in \Sigma_N$. By Theorem 1 there exist two mcst's Γ and Γ' for w and w', respectively, such that they both fit with τ. First note that

$$m^\tau_{\tau(1)}(c_w) = w(0, \tau(1)) \leq w'(0, \tau(1)) = m^\tau_{\tau(1)}(c_{w'}).$$

Further

$$\begin{aligned} m^\tau_{\tau(j)}(c_w) &= \min_{k \in (\tau(j))'} w(k, \tau(j)) \\ &\leq \min_{k \in (\tau(j))'} w'(k, \tau(j)) \\ &= m^\tau_{\tau(j)}(c_{w'}), \end{aligned}$$

for each $j \in \{2, \ldots, |N|\}$, where the first and the second equality follow by Proposition 3 and the inequality follows from $w(e) \leq w'(e)$ for each $e \in E_{N'}$. □

Theorem 2. *The correspondence $\mathcal{BC}$ is a cost monotonic multisolution.*

Proof. Let $w, w' \in \mathcal{W}^{N'}$ be such that $w(e) \leq w'(e)$ for each $e \in E_{N'}$. By Theorem 1 and properties of concave games, $\mathcal{BC}(w)$ is a convex set whose extreme points are the marginal vectors of the game $c_{\bar{w}}$, i.e. each element of $\mathcal{BC}(w)$ is a convex combination of marginal vectors of the game $c_{\bar{w}}$. Let $x \in \mathcal{BC}(w)$. There exist numbers α^τ, with $\tau \in \Sigma_N$, $0 \leq \alpha^\tau \leq 1$, $\sum_{\tau \in \Sigma_N} \alpha^\tau = 1$ and

$$x = \sum_{\tau \in \Sigma_N} \alpha^\tau \, m^\tau(c_{\bar{w}}). \tag{14}$$

Hence

$$\begin{aligned} x &= \textstyle\sum_{\tau \in \Sigma_N} \alpha^\tau \, m^\tau(c_{\bar{w}}) \\ &\leq \textstyle\sum_{\tau \in \Sigma_N} \alpha^\tau \, m^\tau(c_{\bar{w}'}) \\ &= x' \in \mathcal{BC}(w'), \end{aligned} \tag{15}$$

where the inequality follows by Proposition 4 and the fact that $\bar{w}(e) \leq \bar{w}'(e)$ for each $e \in E_{N'}$ and the second equality by Theorem 1, implying that $\mathcal{BC}(w) \subseteq compr^-(\mathcal{BC}(w'))$. Using a similar argument the other way around in relations (15), it follows that $\mathcal{BC}(w') \subseteq compr^+(\mathcal{BC}(w))$, which concludes the proof. □

To connect the cost monotonicity of the Bird core with cost monotonicity of Obligation rules, we need Proposition 5.

Proposition 5. *Let $F : \mathcal{W}^{N'} \to \mathbb{R}^N$ be a cost monotonic and efficient solution. Then*

i) $F(\bar{w}) = F(w)$ for every $w \in \mathcal{W}^{N'}$;
ii) If F is also stable (i.e. $F(w') \in \mathcal{C}(c_{w'})$ for every $w' \in \mathcal{W}^{N'}$), then $F(w) \in \mathcal{BC}(w)$ for every $w \in \mathcal{W}^{N'}$.

Proof. Let $w \in \mathcal{W}^{N'}$. First note that by Definition 1,

$$\bar{w}(e) \leq w(e) \text{ for each } e \in E_{N'}. \tag{16}$$

Let Γ be an mcst for w. Consider first i). By inequality (16) and cost monotonicity of F, $F(\bar{w}) \leq F(w)$. On the other hand Γ is an mcst for $\bar{w}$ too and by efficiency of F

$$\sum_{i \in N} F_i(\bar{w}) = \sum_{i \in N} F_i(w) = w(\Gamma).$$

So, $F(\bar{w}) = F(w)$. Consider now ii). By inequality (16),

$$c_{\bar{w}}(S) \leq c_w(S) \text{ for all } S \subseteq N,$$

and by Definition 1

$$c_{\bar{w}}(N) = c_w(N) = w(\Gamma).$$

Then, by stability of F, $F(\bar{w}) \in \mathcal{C}(c_{\bar{w}}) = \mathcal{BC}(w) \subseteq \mathcal{C}(c_w)$ and by result (i) $F(w) \in \mathcal{BC}(w)$ too. □

Remark 4. Proposition 5 can be extended to multisolutions which are cost monotonic and efficient (Property 1 in next section) multisolutions. From this follows that $\mathcal{BC}$ is the "largest" cost monotonic stable multisolution.

Remark 5. As previously said, in [23] we have introduced the class of Obligation rules and proved that they are both cost monotonic and stable solutions for mcst situations. So, by Proposition 5 it follows that for each $w \in \mathcal{W}^{N'}$, the set $\mathcal{F}(w) = \{\phi(w) \mid \phi \text{ is an Obligation rule}\}$ is a subset of the Bird core $\mathcal{BC}(w)$ and $\mathcal{F}(w) = \mathcal{F}(\bar{w})$.

5 An Axiomatic Characterization of the Bird Core

In order to introduce an axiomatic characterization of the Bird core, we need to prove the following fact for NA-semimetric mcst situations.

Lemma 2. *Let $w, w' \in \mathcal{W}^{N'}$ and let $\sigma \in \Sigma_{E_{N'}}$ be such that $w, w' \in K^{\sigma}$. Let $\alpha, \alpha' \geq 0$. Then $\alpha\bar{w}, \alpha'\bar{w}', \overline{\alpha w + \alpha' w'} \in K^{\hat{\sigma}}$ for some $\hat{\sigma} \in \Sigma_{E_{N'}}$.*

Proof. By relation (4), for each edge $e \in E_{N'}$, there is an edge $\bar{e} \in E_{N'}$ such that $\bar{w}(e) = w(\bar{e})$: given that $e = \{i, j\}$, $\bar{e}$ is such that $w(\bar{e}) = \min_{P \in \mathcal{P}_{ij}^{N'}} t(P)$. Note that for each w_1 in the same cone K^{σ} as w we have $\bar{w}_1(e) = w(\bar{e})$. This implies that for all pairs of edges $e_1, e_2 \in E_{N'}$

$$\bar{w}(e_1) \leq \bar{w}(e_2) \Leftrightarrow w(\bar{e}_1) \leq w(\bar{e}_2) \Leftrightarrow \bar{w}_1(e_1) \leq \bar{w}_1(e_2).$$

So, for each $\bar{\sigma} \in \Sigma_{E_{N'}}$ we have:

$$\bar{w} \in K^{\bar{\sigma}} \Leftrightarrow \bar{w}' \in K^{\bar{\sigma}}.$$

Using this fact, respectively, for αw, $\alpha' w'$ and $\alpha w + \alpha' w' \in K^{\sigma}$ in the role of w_1, we obtain

$$\bar{w} \in K^{\bar{\sigma}} \Leftrightarrow \alpha\bar{w}, \alpha'\bar{w}', \overline{\alpha w + \alpha' w'} \in K^{\bar{\sigma}},$$

for each $\bar{\sigma} \in \Sigma_{E_{N'}}$. □

Proposition 6. *Let $w, w' \in \mathcal{W}^{N'}$ and let $\sigma \in \Sigma_{E_{N'}}$ be such that $w, w' \in K^{\sigma}$. Let $\alpha, \alpha' \geq 0$. Then*

i) $\overline{\alpha w + \alpha' w'} = \alpha\bar{w} + \alpha'\bar{w}'$;
ii) $c_{\overline{\alpha w + \alpha' w'}} = \alpha c_{\bar{w}} + \alpha' c_{\bar{w}'}$.

[*The NA-semimetric mcst situations $\bar{w}, \bar{w}', \overline{\alpha w + \alpha' w'}$ are obtained via reduction of the weight functions $w, w', \alpha w + \alpha' w'$, respectively.*]

Proof. i) Note that

$$\begin{aligned}\overline{\alpha w + \alpha' w'}(\{i,j\}) &= \min_{P \in \mathcal{P}_{ij}^{N'}} \max_{e \in E(P)} \left(\alpha w(e) + \alpha' w'(e)\right) \\ &= \alpha \min_{P \in \mathcal{P}_{ij}^{N'}} \max_{e \in E(P)} w(e) \\ &+ \alpha' \min_{P \in \mathcal{P}_{ij}^{N'}} \max_{e \in E(P)} w'(e) \\ &= \alpha \bar{w}(\{i,j\}) + \alpha' \bar{w'}(\{i,j\}),\end{aligned}$$

where the second equality follows from the fact that w, w' and $\alpha w + \alpha' w'$ all belong to K^σ;

ii) Note that, by Lemma 2, $\alpha\bar{w}, \alpha'\bar{w'}, \overline{\alpha w + \alpha' w'} \in K^{\bar{\sigma}}$ for some $\bar{\sigma} \in \Sigma_{E_{N'}}$. For each $S \in 2^N \setminus \{\emptyset\}$, there is, according to Remark 3, a common mcst Γ_S for $\alpha\bar{w}, \alpha'\bar{w'}$ and $\overline{\alpha w + \alpha' w'}$. Hence

$$\begin{aligned}\alpha c_{\bar{w}}(S) + \alpha' c_{\bar{w'}}(S) &= \textstyle\sum_{e \in \Gamma_S} \alpha\bar{w}(e) + \sum_{e \in \Gamma_S} \alpha'\bar{w'}(e) \\ &= \textstyle\sum_{e \in \Gamma_S} \left(\alpha\bar{w'}(e) + \alpha'\bar{w}(e)\right) \\ &= \textstyle\sum_{e \in \Gamma_S} \left(\overline{\alpha w + \alpha' w'}(e)\right) \\ &= c_{\overline{\alpha w + \alpha' w'}}(S),\end{aligned}$$

where the third equality follows by (i). □

Some interesting properties for multisolutions on mcst situations are the following.

Property 1. The multisolution G is *efficient* (EFF) if for each $w \in \mathcal{W}^{N'}$ and for each $x \in G(w)$

$$\sum_{i \in N} x_i = w(\Gamma),$$

where Γ is a minimum cost spanning network for w on N'.

Property 2. The multisolution G has the *positive* (POS) property if for each $w \in \mathcal{W}^{N'}$ and for each $x \in G(w)$

$$x_i \geq 0$$

for each $i \in N$.

Property 3. The multisolution G has the *Upper Bounded Contribution* (UBC) property if for each $w \in \mathcal{W}^{N'}$ and every (w, N')-component $C \neq \{0\}$

$$\sum_{i \in C \setminus \{0\}} x_i \leq \min_{i \in C \setminus \{0\}} w(\{i,0\})$$

for each $x \in G(w)$.

Property 4. The multisolution G has the *Cone-wise Positive Linearity* (CPL) property if for each $\sigma \in \Sigma_{E_{N'}}$, for each pair of mcst situations $w, \widehat{w} \in K^\sigma$ and for each pair $\alpha, \widehat{\alpha} \geq 0$, we have

$$G(\alpha w + \widehat{\alpha}\widehat{w}) = \alpha G(w) + \widehat{\alpha} G(\widehat{w}).$$

[Here we denote by $\alpha G(w) + \widehat{\alpha} G(\widehat{w})$ the set $\{\alpha x + \widehat{\alpha}\widehat{x} | x \in G(w), \widehat{x} \in G(\widehat{w})\}$.]

Proposition 7. *The Bird core $\mathcal{BC}$ satisfies the properties EFF, POS, UBC and CPL.*

Proof. Let $w \in \mathcal{W}^{N'}$ and let $\sigma \in \Sigma_{E_{N'}}$ be such that $w \in K^\sigma$. Since $\mathcal{BC}(w) = \mathcal{C}(c_{\bar{w}})$, the following considerations hold:

i) For each allocation $x \in \mathcal{BC}(w)$, $\sum_{i \in N} x_i = w(\Gamma)$ for some mcst Γ by the efficiency property of the core of the game $c_{\bar{w}}$. So $\mathcal{BC}$ has the EFF property.

ii) For each allocation $x \in \mathcal{BC}(w)$, $x_i \geq 0$ for each $i \in N$ since the Bird core is the convex hull of all Bird allocations in the mcst $\bar{w}$, which are vectors in $\mathbb{R}^N_+$. So $\mathcal{BC}$ has the POS property.

iii) For each (w, N')-component $C \neq \{0\}$ and each $x \in \mathcal{BC}(w)$

$$\sum_{i \in C \setminus \{0\}} x_i \leq c_{\bar{w}}(C \setminus \{0\}) = \min_{i \in C \setminus \{0\}} w(\{i, 0\})$$

by coalitional rationality of the core of the game $c_{\bar{w}}$. So $\mathcal{BC}$ has the UBC property.

iv) Let $\sigma \in \Sigma_{E_{N'}}$, let $w, w' \in \mathcal{W}^{N'}$ be such that $w, w' \in K^\sigma$ and let $\alpha, \alpha' \geq 0$. Since the core is additive on the class of concave games (see [9]), we have

$$\mathcal{BC}(\alpha w + \alpha' w') = \mathcal{C}(c_{\overline{\alpha w + \alpha' w'}}) = \alpha \mathcal{C}(c_{\bar{w}}) + \alpha' \mathcal{C}(c_{\bar{w}'}) = \alpha \mathcal{BC}(w) + \alpha' \mathcal{BC}(w').$$

Hence $\mathcal{BC}$ has the CPL property. □

Inspired by the axiomatic characterization of the P-value ([4]) we provide the following theorem.

Theorem 3. *The Bird core $\mathcal{BC}$ is the largest multisolution which satisfies EFF, POS, UBC and CPL, i.e. for each multisolution F which satisfies EFF, POS, UBC and CPL, we have $F(w) \subseteq \mathcal{BC}(w)$, for each $w \in \mathcal{W}^{N'}$.*

Proof. We already know by Proposition 7 that the Bird core $\mathcal{BC}$ satisfies the four properties EFF, POS, UBC and CPL. Let $\Psi : \mathcal{W}^{N'} \twoheadrightarrow \mathbb{R}^N$ be a multisolution satisfying EFF, POS, UBC and CPL. Let $w \in \mathcal{W}^{N'}$ and $\sigma \in \Sigma_{E_{N'}}$ be such that $w \in K^\sigma$. We have to prove that $\Psi(w) \subseteq \mathcal{BC}(w)$. First, note that by the CPL property of Ψ

$$\Big(w(\sigma(1))\Psi(e^{\sigma,1}) + \sum_{k=2}^{|E_{N'}|} \big(w(\sigma(k)) - w(\sigma(k-1))\big)\Psi(e^{\sigma,k})\Big) = \Psi(w). \qquad (17)$$

Let $x \in \psi(w)$. According to (17) there exists $x^{e^{\sigma,k}} \in \Psi(e^{\sigma,k})$ for each $k \in \{1, \ldots, |E_{N'}|\}$ such that

$$x = w(\sigma(1))x^{e^{\sigma,1}} + \sum_{k=2}^{|E_{N'}|} \big(w(\sigma(k)) - w(\sigma(k-1))\big)x^{e^{\sigma,k}}.$$

By the UBC property, for each $k \in \{1, \ldots, |E_{N'}|\}$ and for each $(e^{\sigma,k}, N')$-component $C \neq \{0\}$ we have

$$\sum_{i \in C \setminus \{0\}} x_i^{e^{\sigma,k}} \leq \min_{i \in C \setminus \{0\}} e^{\sigma,k}(\{i,0\}) = \begin{cases} 0 \text{ if } 0 \in C \\ 1 \text{ if } 0 \notin C \end{cases} \tag{18}$$

implying that

$$\sum_{i \in N} x_i^{e^{\sigma,k}} = \sum_{C \in \mathcal{C}(e^{\sigma,k})} \sum_{j \in C \setminus \{0\}} x_j^{e^{\sigma,k}} \leq |\mathcal{C}(e^{\sigma,k})| - 1 = e^{\sigma,k}(\Gamma),$$

where Γ is a minimum spanning network on N' for the simple mcst situation $e^{\sigma,k}$. By the EFF property, we have $\sum_{i \in N} x_i^{e^{\sigma,k}} = e^{\sigma,k}(\Gamma)$, and then inequalities in relation (18) are equalities, that is

$$\sum_{i \in C \setminus \{0\}} x_i^{e^{\sigma,k}} = \begin{cases} 0 \text{ if } 0 \in C \\ 1 \text{ if } 0 \notin C. \end{cases} \tag{19}$$

Now, consider the game $c_{\overline{e^{\sigma,k}}}$ corresponding to the simple mcst situation $\overline{e^{\sigma,k}}$. Note that for each $S \in 2^N \setminus \{\emptyset\}$,

$$c_{\overline{e^{\sigma,k}}}(S) = |\{C \ : \ C \text{ is a } (e^{\sigma,k}, N') - \text{component}, C \cap S \neq \emptyset, 0 \notin C\}|,$$

which is the number of $(e^{\sigma,k}, N')$-components not connected to 0 in $e^{\sigma,k}$ with at least one node in the player set S. By (19) and the POS property, it follows that $\sum_{i \in S} x_i^{e^{\sigma,k}} \leq c_{\overline{e^{\sigma,k}}}(S)$ and together with the EFF property we have $x^{e^{\sigma,k}} \in \mathcal{C}(c_{\overline{e^{\sigma,k}}}) = \mathcal{BC}(e^{\sigma,k})$. Moreover, from Proposition 6 it follows

$$x = \Big(w(\sigma(1))x^{e^{\sigma,1}} + \sum_{k=2}^{|E_{N'}|} \big(w(\sigma(k)) - w(\sigma(k-1))\big)x^{e^{\sigma,k}}\Big) \in \mathcal{C}(c_{\bar{w}}) = \mathcal{BC}(w).$$

Keeping into account relation (17), we have $\Psi(w) \subseteq \mathcal{BC}(w)$. □

6 Final Remarks

This paper deals mainly with the Bird core of an mcst situation and its monotonicity and additivity properties.

Given an mcst $w \in \mathcal{W}^{N'}$ and an mcst Γ for N' in w, the Bird core has been introduced (cf. Bird, 1976) as the core of the mcst game (N, c_{w^Γ}) corresponding to the mcst situation w^Γ defined as in relation (5).

From a combinatorial perspective, Proposition 2 allows for a relevant reduction in the number of operations needed to obtain the minimal mcst situation corresponding to an mcst w. In fact, by means of relation (4) it is

not necessary anymore to solve the mcst problem in w finding an optimal spanning tree Γ and then computing w^{Γ} as defined by relation (5).

The attention to monotonicity properties of solutions for cost and reward sharing situations is growing in the literature.

In [21] attention is paid to population monotonic allocation schemes (pmas), in [7] and [24] to bi-monotonic allocation schemes (bi-mas) and in [5] to type monotonic allocation schemes. For mcst-situations, the existence of population monotonic allocation schemes has been established in [19]. For special directed mcst-situations also pmas-es exist as is shown in [17].

As we already said in the introduction, the problem of finding cost monotonic allocation rules has been tackled in [10], paying attention only to the agents who are directly involved in the cost increasing. In [23] so called Obligation rules for mcst-situations turn out to be cost monotonic (with respect to all the agents) and induce also pmas-es. A special Obligation rule is the P-value discussed in [4] (see also [12], [11], [2], [18]).

In the axiomatic characterization of Section 5, we use very intuitive axioms (UBC, EFF, POS and CPL) to characterize the Bird core. Let $w \in \mathcal{W}^{N'}$. From the game theoretical point of view, the UBC property together with the EFF property selects a subset of the imputation set of the mcst game c_w, i.e. the set of imputations which also satisfy the intermediate stability conditions for coalitions of players that are (w, N')-connected. Note that for such coalitions checking for the intermediate stability of an allocation is very easy (just look at the minimum distance from the source). The POS property guarantees that no players should be subsidized from others according to some allocations: all the players must pay at least zero of the total cost. One can easily check that EFF, POS and UBC properties are satisfied by many allocation rules for mcst situations, like the Bird rule (Bird 1976), Obligation rules ([23]), Construct & Charge rules ([18]) but not from classical game theoretical solutions, like the Shapley value, for example. For a deeper game theoretical view of the CPL property, we refer to [19], where CPL formed the base for a decomposition theorem showing that every mcst game can be written as nonnegative combination of mcst games corresponding to simple mcst situations. The CPL property for solutions has been also used to axiomatically characterize the P-value in [4].

For further considerations on the additivity properties of solutions see also [6], [22].

References

1. H. Aarts. Minimum cost spanning tree games and set games. PhD Dissertation, Univ. of Twente, The Netherlands, 1994.
2. G. Bergantiños and J.J. Vidal-Puga. Defining rules in cost spanning tree problems through the canonical form. EconPapers (RePEc:wpa:wuwpga 0402004), 2004.
3. C.G. Bird. On cost allocation for a spanning tree: a game theoretic approach. Networks 6:335-350, 1976.

4. R. Branzei, S. Moretti, H. Norde, and S. Tijs. The P-value for cost sharing in minimum cost spanning tree situations. Theory and Decision 56:47-61, 2004.
5. R. Branzei, T. Solymosi, and S. Tijs. Type monotonic allocation schemes for multi-glove games. CentER DP 2002-117, Tilburg Univ., The Netherlands, 2002.
6. R. Branzei and S. Tijs. Additivity regions for solutions in cooperative game theory. Libertas Mathematica 21:155-167, 2001.
7. R. Branzei, S. Tijs, and J. Timmer. Information collecting situations and bi-monotonic allocation schemes. Math. Meth. Oper. Res., 54:303-313, 2001.
8. A. Claus and D.J. Kleitman. Cost allocation for a spanning tree. Networks 3:289-304, 1973.
9. I. Dragan, J. Potters, and S. Tijs. Superadditivity for solutions of coalitional games. Libertas Mathematica 9:101-110, 1989.
10. B. Dutta and A. Kar. Cost monotonicity, consistency and minimum cost spanning tree games. Games and Economic Behavior, 48:223-248, 2004.
11. V. Feltkamp. Cooperation in controlled network structures. PhD Dissertation, Tilburg Univ., The Netherlands, 1995.
12. V. Feltkamp, S. Tijs, and S. Muto. On the irreducible core and the equal remaining obligations rule of minimum cost spanning extension problems. CentER DP 1994 nr.106, Tilburg Univ., The Netherlands, 1994.
13. D. Granot and A. Claus. Game theory application to cost allocation for a spanning tree. Working Paper 402, Fac. of Commerce and Business Administration, Univ. of British Columbia, 1976.
14. D. Granot and G. Huberman. On minimum cost spanning tree games. Mathematical Programming 21:1-18, 1981.
15. T. Ichiishi. Super-modularity: applications to convex games and the greedy algorithm for LP. Journal of Economic Theory 25:283-286, 1981.
16. J.B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc., 7:48-50, 1956.
17. S. Moretti, H. Norde, K.H. Pham Do, and S. Tijs. Connection problems in mountains and monotonic allocation schemes. Top 10:83-99, 2002.
18. S. Moretti, S. Tijs, R. Branzei, H. Norde. Cost monotonic 'construct and charge' rules for connection situations. Working Paper, 2005.
19. H. Norde, S. Moretti, and S. Tijs. Minimum cost spanning tree games and population monotonic allocation schemes. European J. Oper. Res. 154:84-97, 2004.
20. R.C. Prim. Shortest connection networks and some generalizations. Bell Systems Technical Journal 36:1389-1401, 1957.
21. Y. Sprumont. Population monotonic allocation schemes for cooperative games with transferable utility. Games and Economic Behavior 2:378-394, 1990.
22. S. Tijs and R. Branzei. Additive stable solutions on perfect cones of cooperative games. Int. J. Game Theory 31:469-474, 2002.
23. S. Tijs, R. Branzei, S. Moretti and H. Norde. Obligation rules for minimum cost spanning tree situations and their monotonicity properties. CentER DP 2004-53, Tilburg Univ., The Netherlands, 2004 (to appear in European J. Oper. Res.).
24. M. Voorneveld, S. Tijs, and S. Grahn. Monotonic allocation schemes in clan games. Math. Meth. Oper. Res., 56:439-449, 2002.

A Parametric Family of Mixed Coalitional Values *

Francesc Carreras[1] and María Albina Puente[2]

[1] Department of Applied Mathematics II and Industrial Engineering School of Terrassa, Polytechnic University of Catalonia, Spain. francesc.carreras@upc.edu

[2] Department of Applied Mathematics III and Polytechnic School of Manresa, Polytechnic University of Catalonia, Spain. m.albina.puente@upc.edu

Summary. We introduce here a family of mixed coalitional values. They extend the binomial semivalues to games endowed with a coalition structure, satisfy the property of symmetry in the quotient game and the quotient game property, generalize the symmetric coalitional Banzhaf value introduced by Alonso and Fiestras and link and merge the Shapley value and the binomial semivalues. A computational procedure in terms of the multilinear extension of the original game is also provided and an application to political science is sketched.

1 Introduction

The parallel axiomatic characterization stated by Feltkamp [8] shows that the only difference between the Shapley value [19] and the Banzhaf value (cf.[13]), as allocation rules for all cooperative games, is that the first satisfies efficiency while the second satisfies the total power property.

In the framework of cooperative games with a coalition structure, other essential differences also arise between the Owen value [14] and the modified Banzhaf value or Owen–Banzhaf value [16]. The Owen–Banzhaf value fails to satisfy the property of symmetry in the quotient game and the quotient game property, which are fulfilled by the Owen value.

Alonso and Fiestras [1] suggested a modification of the Owen–Banzhaf value that satisfies these two properties and can therefore be compared with the Owen value in terms analogous to Feltkamp's. Our aim here is to introduce the notion of coalitional binomial semivalue as a wide generalization of the Alonso–Fiestras value (essentially: $p \in [0,1]$ instead of $p = 1/2$) in order to get a symmetric coalitional binomial semivalue that still satisfies the property of symmetry in the quotient game and the quotient game property, so that it

* Research partially supported by Grant BFM 2003–01314 of the Science and Technology Spanish Ministry and the European Regional Development Fund.

differs from the Owen value just in satisfying a total power property instead of efficiency.

The organization of the paper is as follows. In Section 2, a minimum of preliminaries is provided. Section 3 is devoted to define and study the symmetric coalitional binomial semivalue, and it includes an axiomatic characterization that parallels Owen's [14] for the Owen value. In Section 4 we present a computation procedure for the symmetric coalitional binomial semivalue. Finally, Section 5 contains a remark on simple games and a detailed example.

2 Preliminaries

2.1 Games and Semivalues

Let N be a finite set of players and 2^N be the set of its coalitions (subsets of N). A cooperative game on N is a function $v : 2^N \to \mathbb{R}$, that assigns a real number $v(S)$ to each coalition $S \subseteq N$ with $v(\emptyset) = 0$. A game v is monotonic if $v(S) \leq v(T)$ whenever $S \subseteq T \subseteq N$. A player $i \in N$ is a dummy in v if $v(S \cup \{i\}) = v(S) + v(\{i\})$ for all $S \subseteq N \backslash \{i\}$, and null if, moreover, $v(\{i\}) = 0$. Two players $i, j \in N$ are symmetric in v if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \backslash \{i, j\}$.

Endowed with the natural operations for real–valued functions, the set of all cooperative games on N is a vector space $\mathcal{G}_N$. For every nonempty coalition $T \subseteq N$, the unanimity game u_T is defined by $u_T(S) = 1$ if $T \subseteq S$ and $u_T(S) = 0$ otherwise. Finally, every permutation θ of N induces a linear automorphism of $\mathcal{G}_N$ given by $(\theta v)(S) = v(\theta^{-1} S)$ for all $S \subseteq N$ and all v.

By a value on $\mathcal{G}_N$ we will mean a map $f : \mathcal{G}_N \to \mathbb{R}^N$, that assigns to every game v a vector $f[v]$ with components $f_i[v]$ for all $i \in N$.

Following Weber's [22] axiomatic description, $\psi : \mathcal{G}_N \to \mathbb{R}^N$ is a semivalue iff it satisfies the following properties:

(i) *linearity*: $\psi[v + v'] = \psi[v] + \psi[v']$ (*additivity*) and $\psi[\lambda v] = \lambda \psi[v]$ for all $v, v' \in \mathcal{G}_N$ and $\lambda \in \mathbb{R}$;

(ii) *anonymity*: $\psi_{\theta i}[\theta v] = \psi_i[v]$ for all θ on N, $i \in N$, and $v \in \mathcal{G}_N$;

(iii) *positivity*: if v is monotonic, then $\psi[v] \geq 0$;

(iv) *dummy player property*: if $i \in N$ is a dummy in game v, then $\psi_i[v] = v(\{i\})$.

There is an interesting characterization of semivalues, by means of weighting coefficients, due to Dubey, Neyman and Weber [7]. Set $n = |N|$. Then: (a) for every weighting vector $\{p_k\}_{k=0}^{n-1}$ such that $\sum_{k=0}^{n-1} p_k \binom{n-1}{k} = 1$ and $p_k \geq 0$ for all k, the expression

$$\psi_i[v] = \sum_{S \subseteq N \backslash \{i\}} p_s [v(S \cup \{i\}) - v(S)] \quad \text{for all } i \in N \text{ and all } v \in \mathcal{G}_N,$$

where $s = |S|$, defines a semivalue ψ; (b) conversely, every semivalue can be obtained in this way; (c) the correspondence given by $\{p_k\}_{k=0}^{n-1} \mapsto \psi$ is bijective.

Well known examples of semivalues are the Shapley value φ [19], for which $p_k = 1/n\binom{n-1}{k}$, and the Banzhaf value β (cf.[13]), for which $p_k = 2^{1-n}$. The Shapley value φ is the only efficient semivalue, in the sense that the equality $\sum_{i \in N} \varphi_i[v] = v(N)$ holds for every $v \in \mathcal{G}_N$.

Notice that these two classical values are defined for each N. The same happens with the binomial semivalues, introduced by Puente [18] as follows. Let $p \in [0,1]$ and $p_k = p^k(1-p)^{n-k-1}$ for $k = 0, 1, \ldots, n-1$. Then $\{p_k\}_{k=0}^{n-1}$ is a weighting vector and defines a semivalue that will be denoted as ψ^p and called the p–binomial semivalue. Using the convention that $0^0 = 1$, the definition makes sense also for $p = 0$ and $p = 1$, where we respectively get the dictatorial index ψ^0 and the marginal index ψ^1, introduced by Owen [15] and such that $\psi_i^0[v] = v(\{i\})$ and $\psi_i^1[v] = v(N) - v(N \backslash \{i\})$ for all $i \in N$ and all $v \in \mathcal{G}_N$. Of course, $p = 1/2$ gives $\psi^{1/2} = \beta$ —the Banzhaf value.

In fact, semivalues are defined on cardinalities rather than on specific player sets: this means that a weighting vector $\{p_k\}_{k=0}^{n-1}$ defines a semivalue ψ on all N such that $n = |N|$. When necessary, we shall write $\psi^{(n)}$ for a semivalue on cardinality n and $p_k^{(n)}$ for its weighting coefficients. A semivalue $\psi^{(n)}$ induces semivalues $\psi^{(t)}$ for all cardinalities $t < n$, recurrently defined by the Pascal triangle (inverse) formula given by Dragan [6]:

$$p_k^{(t)} = p_k^{(t+1)} + p_{k+1}^{(t+1)} \quad \text{for} \quad 0 \leq k < t.$$

A series $\psi = \{\psi^{(n)}\}_{n=1}^{\infty}$ of semivalues, one for each cardinality, is a multisemivalue if it satisfies Dragan's recurrence formula. Thus, the Shapley and Banzhaf values and all binomial semivalues are multisemivalues.

2.2 Games with Coalition Structure

Let us consider a finite set, say, $N = \{1, 2, \ldots, n\}$. We will denote by $P(N)$ the set of all partitions of N. Each $P \in P(N)$ is called a coalition structure or system of unions on N. The so–called trivial coalition structures are $P^n = \{\{1\}, \{2\}, \ldots, \{n\}\}$ and $P^N = \{N\}$. A cooperative game with a coalition structure is a pair $[v; P]$, where $v \in \mathcal{G}_N$ and $P \in P(N)$ for a given N. We denote by $\mathcal{G}_N^{cs}$ the set of all cooperative games with a coalition structure and player set N.

If $[v; P] \in \mathcal{G}_N^{cs}$ and $P = \{P_1, P_2, \ldots, P_m\}$, the quotient game v^P is the cooperative game played by the unions, or, rather, by the set $M = \{1, 2, \ldots, m\}$ of their representatives, as follows:

$$v^P(R) = v\Big(\bigcup_{r \in R} P_r\Big) \quad \text{for all } R \subseteq M.$$

Unions P_r, P_s are said to be symmetric in $[v;P]$ if r,s are symmetric players in v^P.

By a coalitional value on $\mathcal{G}_N^{cs}$ we will mean a map $g : \mathcal{G}_N^{cs} \to \mathbb{R}^N$, which assigns to every pair $[v;P]$ a vector $g[v;P]$ with components $g_i[v;P]$ for each $i \in N$.

The Owen value [14] is the coalitional value Φ defined by

$$\Phi_i[v;P] = \sum_{R \subseteq M \setminus \{k\}} \sum_{T \subseteq P_k \setminus \{i\}} \frac{1}{mp_k} \frac{1}{\binom{m-1}{r}} \frac{1}{\binom{p_k-1}{t}} [v(Q \cup T \cup \{i\}) - v(Q \cup T)]$$

for all $i \in N$ and $[v;P] \in \mathcal{G}_N^{cs}$, where $P_k \in P$ is the union such that $i \in P_k$ and $Q = \bigcup_{r \in R} P_r$. It was axiomatically characterized by Owen [14] as the only coalitional value that satisfies the following properties: the natural extensions to this framework of

- *efficiency*
- *additivity*
- *the dummy player property*

and also

- *symmetry within unions*: if $i,j \in P_k$ are symmetric in v then

$$\Phi_i[v;P] = \Phi_j[v;P]$$

- *symmetry in the quotient game*: if $P_r, P_s \in P$ are symmetric in $[v;P]$ then

$$\sum_{i \in P_r} \Phi_i[v;P] = \sum_{j \in P_s} \Phi_j[v;P].$$

The Owen value is a coalitional value of the Shapley value φ in the sense that $\Phi[v;P^n] = \varphi[v]$ for all $v \in \mathcal{G}_N$. Besides, $\Phi[v;P^N] = \varphi[v]$. Finally, as Φ is defined for any N, the following property makes sense and is also satisfied:

- *quotient game property*: for all $[v;P] \in \mathcal{G}_N^{cs}$,

$$\sum_{i \in P_k} \Phi_i[v;P] = \Phi_k[v^P;P^m] \quad \text{for all } P_k \in P.$$

The Owen value can be viewed as a two–step allocation rule. First, each union P_k receives its payoff in the quotient game according to the Shapley value; then, each P_k splits this amount among its players by applying the Shapley value to a game played in P_k as follows: the worth of each subcoalition T of P_k is the Shapley value that T would get in a "pseudoquotient game" played by T and the remaining unions on the assumption that $P_k \setminus T$ leaves the game, i.e. the quotient game after replacing P_k with T. This is the way to bargain within the union: each subcoalition T claims the payoff it would obtain when dealing with the other unions in absence of its partners in P_k.

The Owen–Banzhaf value B [16] follows a similar scheme. The resulting formula parallels that of the Owen value given above with the sole change of coefficient $1/mp_k\binom{m-1}{r}\binom{p_k-1}{t}$ by $2^{1-m}2^{1-p_k}$. This value, which is a coalitional value of the Banzhaf value β, does not satisfy efficiency, but neither symmetry in the quotient game nor the quotient game property. The bargaining interpretation is the same as in the case of the Owen value by replacing everywhere the Shapley value with the Banzhaf value.

Alonso and Fiestras [1] introduced a modification of the Owen–Banzhaf value. In this case, the coefficient of each marginal contribution is replaced with $2^{1-m}/p_k\binom{p_k-1}{t}$. This symmetric coalitional Banzhaf value Π satisfies the same properties as the Owen value, with the sole exception of efficiency — replaced by a total power property—, as well as the quotient game property, and it is a coalitional value of the Banzhaf value.

Example (Alonso and Fiestras [1]). Let us take $n = 5$ and consider the unanimity game u_N and the coalition structure $P = \{P_1, P_2\}$ where $P_1 = \{1, 2, 3\}$ and $P_2 = \{4, 5\}$. Notice that the quotient game is $u_N^P = u_M$, where $M = \{1, 2\}$. It is not difficult to obtain the following values:

$$\begin{aligned}\beta[u_N] &= (1/16, 1/16, 1/16, 1/16, 1/16),\\ \beta[u_N^P] &= (1/2, 1/2),\\ B[u_N; P] &= (1/8, 1/8, 1/8, 1/4, 1/4).\end{aligned}$$

As P_1 and P_2 are symmetric in $[u_N; P]$, it follows that the Owen–Banzhaf value B fails to satisfy the property of symmetry in the quotient game. Neither the quotient game property is fulfilled by B in this instance. Instead

$$\Pi[u_N; P] = (1/6, 1/6, 1/6, 1/4, 1/4)$$

so that both properties are satisfied by the Alonso–Fiestras value Π (here and elsewhere).

3 The Symmetric Coalitional Binomial Semivalue

In this section we define and study a "coalitional version" of each p–binomial semivalue for games with coalition structure. This includes, besides the explicit formula, an axiomatic characterization and an interpretation in terms of a two–step bargaining process, among unions, first, and among players within each union later. We recall that ψ^p denotes, for each $p \in [0, 1]$, the p–binomial semivalue acting on a fixed $\mathcal{G}_N$, and also the following notion (cf. Puente [18]).

Definition 1. *Let $p \in [0, 1]$. A value f on $\mathcal{G}_N$ satisfies the p–binomial total power property if*

$$\sum_{i \in N} f_i[v] = \sum_{i \in N} \sum_{S \subseteq N \setminus \{i\}} p^s (1-p)^{n-s-1} [v(S \cup \{i\}) - v(S)] \quad \textit{for all } v \in \mathcal{G}_N.$$

The Owen (resp., Owen–Banzhaf) value is a natural extension of the Shapley (resp., Banzhaf) value to games with a coalition structure. We generalize this idea.

Definition 2. *Given a value f on $\mathcal{G}_N$, a coalitional value of f is a coalitional value g on $\mathcal{G}_N^{cs}$ such that $g[v;P^n] = f[v]$ for all $v \in \mathcal{G}_N$.*

Let g be a coalitional value of the p–binomial semivalue ψ^p defined for all N, and assume that g satisfies the quotient game property. Then, for a given N and any $[v;P] \in \mathcal{G}_N^{cs}$,

$$\sum_{i\in N} g_i[v;P] = \sum_{k\in M}\sum_{i\in P_k} g_i[v;P] = \sum_{k\in M} g_k[v^P;P^m] = \sum_{k\in M} \psi_k^p[v^P] =$$
$$= \sum_{k\in M}\sum_{R\subseteq M\setminus\{k\}} p^r(1-p)^{m-r-1}[v^P(R\cup\{k\}) - v^P(R)].$$

This motivates the next definition, that is an adaptation of the p–binomial total power property to games with a coalition structure.

Definition 3. *Let $p \in [0,1]$. A coalitional value g on $\mathcal{G}_N^{cs}$ satisfies the coalitional p–binomial total power property if, for all $[v;P] \in \mathcal{G}_N^{cs}$,*

$$\sum_{i\in N} g_i[v;P] = \sum_{k\in M}\sum_{R\subseteq M\setminus\{k\}} p^r(1-p)^{m-r-1}[v^P(R\cup\{k\}) - v^P(R)].$$

The next statement defines and axiomatically characterizes, for each $p \in [0,1]$, the symmetric coalitional p–binomial semivalue, which will be denoted as Ω^p. We need a previous lemma.

Lemma 1. *Let $\emptyset \neq S \subseteq N$, $s = |S|$ and $i \in N$. Then $\psi_i^p[u_S] = p^{s-1}$ if $i \in S$, and $\psi_i^p[u_S] = 0$ otherwise.*

Proof. Let $i \in S$. By the definition of the weighting coefficients of ψ^p we have

$$\psi_i^p[u_S] = \tbinom{n-s}{0}p^{s-1}(1-p)^{n-s} + \tbinom{n-s}{1}p^s(1-p)^{n-s-1} + \cdots + \tbinom{n-s}{n-s}p^{n-1} =$$

$$= p^{s-1}[p+(1-p)]^{n-s} = p^{s-1}.$$

If $i \notin S$, the dummy player property yields $\psi_i^p[u_S] = 0$. □

Theorem 1. *Let $p \in [0,1]$. For any N there is a unique coalitional value on $\mathcal{G}_N^{cs}$ that satisfies additivity, the dummy player property, symmetry within unions, symmetry in the quotient game, and the coalitional p–binomial total power property. Given $[v;P] \in \mathcal{G}_N^{cs}$, this value allocates to each player $i \in N$ the real number*

$$\Omega_i^p[v;P] = \sum_{R\subseteq M\setminus\{k\}}\sum_{T\subseteq P_k\setminus\{i\}} p^r(1-p)^{m-r-1}\frac{1}{p_k\binom{p_k-1}{t}}[v(Q\cup T\cup\{i\}) - v(Q\cup T)],$$

where $P_k \in P$ is the union such that $i \in P_k$ and $Q = \bigcup_{r \in R} P_r$. Moreover, Ω^p is a coalitional value of the p–binomial semivalue ψ^p and satisfies the quotient game property.

Proof. (a) (Existence) It suffices to show that the coalitional value Ω^p given by the above formula satisfies the five properties enumerated in the statement.

1. Additivity. It merely follows from the expression of $\Omega_i^p[v; P]$.

2. Dummy player property. Let $i \in N$ be a dummy player in game v and P be any coalition structure. Assume $i \in P_k$. Then $v(Q \cup T \cup \{i\}) - v(Q \cup T) = v(\{i\})$ for all R and T. As, moreover,

$$\sum_{R \subseteq M \setminus \{k\}} p^r (1-p)^{m-r-1} = 1 \quad \text{and} \quad \sum_{T \subseteq P_k \setminus \{i\}} \frac{1}{p_k \binom{p_k - 1}{t}} = 1,$$

we conclude that $\Omega_i^p[v; P] = v(\{i\})$.

3. Symmetry within unions. Let $i, j \in P_k \in P$ be symmetric players in game v. For each $R \subseteq M \setminus \{k\}$ and $T \subseteq P_k \setminus \{i, j\}$, let $\Delta(R, T, h) = v(Q \cup T \cup \{h\}) - v(Q \cup T)$ for $h = i, j$. Then, by the symmetric position of i, j in v,

$$f(R,T) = \Delta(R,T,i) - \Delta(R,T,j) = 0 \quad \text{and}$$
$$g(R,T) = \Delta(R,T \cup \{j\}, i) - \Delta(R, T \cup \{i\}, j) = 0,$$

so that $\Omega_i^p[v; P] - \Omega_j^p[v; P]$ is equal to

$$\sum_{R \subseteq M \setminus \{k\}} p^r (1-p)^{m-r-1} \sum_{T \subseteq P_k \setminus \{i,j\}} \left[\frac{f(R,T)}{p_k \binom{p_k - 1}{t}} + \frac{g(R,T)}{p_k \binom{p_k - 1}{t+1}} \right] = 0.$$

4. Coalitional p–binomial total power property. Let $[v; P] \in \mathcal{G}_N^{cs}$. Fixing $k \in M$, for every $R \subseteq M \setminus \{k\}$ we consider the game $v_R \in \mathcal{G}_{P_k}$ defined by

$$v_R(T) = v(Q \cup T) - v(Q) \quad \text{for all } T \subseteq P_k.$$

The Shapley value gives, for each $i \in P_k$,

$$\varphi_i[v_R] = \sum_{T \subseteq P_k \setminus \{i\}} \frac{1}{p_k \binom{p_k - 1}{t}} [v(Q \cup T \cup \{i\}) - v(Q \cup T)].$$

Using the efficiency of φ, we get

$$\sum_{i \in P_k} \varphi_i[v_R] = v_R(P_k) = v(Q \cup P_k) - v(Q) = v^P(R \cup \{k\}) - v^P(R).$$

Hence

$$\sum_{i \in P_k} \Omega_i^p[v; P] = \sum_{R \subseteq M \setminus \{k\}} p^r (1-p)^{m-r-1} [v^P(R \cup \{k\}) - v^P(R)] = (\psi^p)_k^{(m)}[v^P]$$

and, finally,

$$\sum_{i\in N} \Omega_i^p[v;P] = \sum_{k\in M} \sum_{R\subseteq M\setminus\{k\}} p^r(1-p)^{m-r-1}[v^P(R\cup\{k\}) - v^P(R)].$$

5. Symmetry in the quotient game. It readily follows from the relationship

$$\sum_{i\in P_k} \Omega_i^p[v;P] = (\psi^p)_k^{(m)}[v^P],$$

stated in the previous point, and the anonymity (whence symmetry) of the p–binomial semivalue ψ^p.

(b) (Uniqueness) Let g be a coalitional value on $\mathcal{G}_N^{cs}$ satisfying the five properties. Using additivity and the fact that the unanimity games form a basis of $\mathcal{G}_N$, it suffices to show that g is completely determined by its action on any pair of the form $[\lambda u_S;P]$, where $\lambda\in\mathbb{R}$, $\emptyset\neq S\subseteq N$ and $P\in P(N)$. By the dummy player property, $g_i[\lambda u_S;P]=0$ if $i\notin S$. This leaves us with players $i\in S$. Let $S'=\{k\in M : S\cap P_k\neq\emptyset\}$ and, for every $k\in S'$, $S'_k=S\cap P_k$. It is easy to see that $(\lambda u_S)^P=\lambda u_{S'}$. From the coalitional p–binomial total power property, and applying Lemma 1, we have

$$\sum_{i\in N} g_i[\lambda u_S;P] = \sum_{k\in M} (\psi^p)_k^{(m)}[\lambda u_{S'}] = \sum_{k\in S'} (\psi^p)_k^{(m)}[\lambda u_{S'}] = \lambda s' p^{s'-1}.$$

Now, from symmetry in the quotient game, if $k\in S'$ then

$$\sum_{i\in S'_k} g_i[\lambda u_S;P] = \sum_{i\in P_k} g_i[\lambda u_S;P] = \lambda p^{s'-1}$$

and, finally, using symmetry within unions,

$$g_i[\lambda u_S;P] = \frac{\lambda p^{s'-1}}{s'_k} \quad \text{for any } i\in S'_k.$$

As $S=\bigcup_{k\in S'} S'_k$, this concludes the proof that g is univocally determined.

(c) Ω^p is a coalitional value of the p–binomial semivalue ψ^p. Indeed, for $P=P^n$ the explicit formula of Ω^p reduces to

$$\Omega_i^p[v;P^n] = \sum_{R\subseteq N\setminus\{i\}} p^r(1-p)^{m-r-1}[v(R\cup\{i\}) - v(R)] = \psi_i^p[v].$$

Finally, the quotient game property: as we have seen when showing the symmetry in the quotient game in part (a) of this proof, and using the preceding property for $\mathcal{G}_M^{cs}$,

$$\sum_{i\in P_k} \Omega_i^p[v;P] = (\psi^p)_k^{(m)}[v^P] = \Omega_k^p[v^P;P^m]. \qquad \square$$

Remark 1. (a) The symmetric coalitional p–binomial semivalue is a natural (and wide) generalization of Alonso and Fiestras' symmetric coalitional Banzhaf value, since $\Omega^{1/2} = \Pi$.

(b) Ω^p relates not only to the p–binomial semivalue ψ^p (of which it is a coalitional value) but also to the Shapley value φ, as $\Omega^p[v; P^N] = \varphi[v]$ for any $v \in \mathcal{G}_N$.

(c) From Theorem 1 it follows that the only axiomatic difference between the Owen value Φ and the symmetric coalitional p–binomial semivalue Ω^p is that the former satisfies efficiency whereas the latter satisfies the coalitional p–binomial total power property, in a way that parallels the distinction between the Shapley value φ and the p–binomial semivalue ψ^p.

(d) The symmetric coalitional p–binomial semivalue represents a two–step bargaining procedure where, first, the unions are allocated in the quotient game the payoff given by the p–binomial semivalue ψ^p and, then, this payoff is efficiently shared within each union according to the Shapley value φ.

4 A Computation Procedure

The multilinear extension (cf.[12]) of a game $v \in \mathcal{G}_N$ is the real–valued function defined on $\mathbb{R}^N$ by

$$f(x_1, x_2, \ldots, x_n) = \sum_{S \subseteq N} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) v(S).$$

As is well known, both the Shapley and Banzhaf values of any game v can be easily obtained from its multilinear extension (cf.[12, 13]). The latter procedure extends well to any p–binomial semivalue (cf.[18]).

In the context of games with a coalition structure, the multilinear extension technique has been also applied to computing the Owen value Φ (cf.[17]), as well as the Owen–Banzhaf value B (cf.[4]) and the symmetric coalitional Banzhaf value Π (cf.[2]). In this section we present a method to compute the symmetric coalitional p–binomial semivalue Ω^p by means of the multilinear extension of the game.

Theorem 2. *Let $p \in [0,1]$ and $[v; P] \in \mathcal{G}_N^{cs}$ be a cooperative game with a coalition structure. Then the following steps lead to the symmetric coalitional p–binomial semivalue of any player $i \in P_k$ in $[v; P]$.*

1. *Obtain the multilinear extension $f(x_1, x_2, \ldots, x_n)$ of game v.*
2. *For every $r \neq k$ and all $h \in P_r$, replace the variable x_h with y_r. This yields a new function of x_j for $j \in P_k$ and y_r for $r \in M \backslash \{k\}$.*
3. *In this function, reduce to 1 all higher exponents, i.e. replace with y_r each y_r^q such that $q > 1$. This gives a new multilinear function that we denote as $g_k((x_j)_{j \in P_k}, (y_r)_{r \in M \backslash \{k\}})$.*

4. *In the function obtained in step 3, substitute each* y_r *by* p. *This provides a new function* $\alpha_k((x_j)_{j\in P_k})$ *defined by*

$$\alpha_k((x_j)_{j\in P_k}) = g_k((x_j)_{j\in P_k}, (p)_{r\in M\setminus\{k\}}).$$

5. *Finally, the symmetric coalitional* p*-binomial semivalue of player* $i \in P_k$ *in* $[v; P]$ *is given by*

$$\Omega_i^p[v; P] = \int_0^1 \frac{\partial \alpha_k}{\partial x_i}(z, z, \ldots, z)dz.$$

Proof. By the second and third steps, we get a multilinear function where all terms corresponding to coalitions S such that $S \cap P_r \neq \emptyset$ and $(N\setminus S) \cap P_r \neq \emptyset$ for some $r \in M\setminus\{k\}$ vanish. Indeed, in step 2, the terms corresponding to these coalitions include expressions of the form $cy_r^{q_1}(1-y_r)^{q_2}$, with $q_1, q_2 \in \mathbb{N}$, and in step 3 these terms turn on $c(y_r - y_r)$ thus getting zero. Hence, the only coalitions S for which the corresponding term of the (initial) multilinear extension may not vanish after steps 2 and 3 are those of the form $S = Q \cup T$, where $T \subseteq P_k$ and $Q = \cup_{r\in R} P_r$ for some $R \subseteq M\setminus\{k\}$. The function arising from step 3 is therefore

$$g_k((x_j)_{j\in P_k}, (y_r)_{r\in M\setminus\{k\}}) =$$

$$\sum_{T\subseteq P_k} \sum_{R\subseteq M\setminus\{k\}} \prod_{j\in T} x_j \prod_{j\in P_k\setminus T} (1-x_j) \prod_{r\in R} y_r \prod_{r\notin R\cup\{k\}} (1-y_r) v(Q\cup T).$$

Substituting each y_r by p (step 4) gives

$$\alpha_k((x_j)_{j\in P_k}) = \sum_{T\subseteq P_k} \sum_{R\subseteq M\setminus\{k\}} \prod_{j\in T} x_j \prod_{j\in P_k\setminus T} (1-x_j) p^r (1-p)^{m-r-1} v(Q\cup T).$$

By differentiating function $\alpha_k((x_j)_{j\in P_k})$ with respect to x_i

$$\frac{\partial \alpha_k}{\partial x_i}((x_j)_{j\in P_k}) = \sum_{T\subseteq P_k\setminus\{i\}} \sum_{R\subseteq M\setminus\{k\}}$$

$$\prod_{j\in T} x_j \prod_{j\in P_k\setminus(T\cup\{i\})} (1-x_j) p^r (1-p)^{m-r-1} [v(Q\cup T\cup\{i\}) - v(Q\cup T)].$$

Finally, by step 5,

$$\int_0^1 \frac{\partial \alpha_k}{\partial x_i}(z, z, \ldots, z)dz =$$

$$\sum_{T\subseteq P_k\setminus\{i\}} \sum_{R\subseteq M\setminus\{k\}} p^r (1-p)^{m-r-1} [v(Q\cup T\cup\{i\}) - v(Q\cup T)] \int_0^1 z^t (1-z)^{p_k-t-1} dz =$$

$$\sum_{T\subseteq P_k\setminus\{i\}} \sum_{R\subseteq M\setminus\{k\}} p^r (1-p)^{m-r-1} \frac{t!(p_k-t-1)!}{p_k!} [v(Q\cup T\cup\{i\}) - v(Q\cup T)]$$

$$= \Omega_i^p[v; P],$$

completing the proof. □

5 A Remark and an Example

Simple games form an especially interesting class of cooperative games. Not only as a test bed for many cooperative concepts, but also for the variety of their interpretations, often far from game theory. In particular, they have been intensively applied to describe and analyze collective decision–making mechanisms —weighted majority games have played a crucial role here—, and the notion of voting power has been closely attached to them.

A cooperative game v on N is simple if it is monotonic and $v(S) = 0$ or 1 for every $S \subseteq N$. A coalition $S \subseteq N$ is winning in v if $v(S) = 1$ (otherwise it is called losing), and $W(v)$ denotes the set of winning coalitions in v. Due to monotonicity, the set $W^m(v)$ of all minimal winning coalitions determines $W(v)$ and hence the game. A simple game v is a weighted majority game if there are nonnegative weights $w_1, w_2, \ldots, w_n$ allocated to the players and a positive quota q such that

$$v(S) = 1 \quad \text{iff} \quad \sum_{i \in S} w_i \geq q.$$

We then write $v = [q; w_1, w_2, \ldots, w_n]$.

Let $\mathcal{SG}_N$ denote the set of all simple games on a given player set N. A power index on $\mathcal{SG}_N$ is a function $f : \mathcal{SG}_N \to \mathbb{R}^N$. All properties stated for values in this paper —with the sole exception of additivity and linearity— make sense for power indices. As $\mathcal{SG}_N$ is a lattice under the standard composition laws given by $(v \vee v')(S) = \max\{v(S), v'(S)\}$ and $(v \wedge v')(S) = \min\{v(S), v'(S)\}$, we will say that a power index f satisfies the transfer property if

$$f[v \vee v'] = f[v] + f[v'] - f[v \wedge v'] \quad \text{for all } v, v' \in \mathcal{SG}_N.$$

Very recently, Carreras, Freixas and Puente [5] gave an axiomatic characterization and a combinatorial description in terms of weighting coefficients for (the restrictions of) semivalues as power indices, which parallel the corresponding ones for semivalues on general cooperative games.

Let $\mathcal{SG}_N^{cs}$ be the set of all simple games with a coalition structure on N. A coalitional power index on $\mathcal{SG}_N^{cs}$ is a function $g : \mathcal{SG}_N^{cs} \to \mathbb{R}^N$. All properties stated for coalitional values in this paper —excluding again additivity and linearity—, as well as the natural extension of the transfer property, make sense for coalitional power indices. Vázquez, van den Nouweland and García–Jurado [21] carried out an axiomatic characterization of the (restricted) Owen value as a coalitional power index by means of efficiency, the transfer property, the dummy player property, symmetry within unions and symmetry in the quotient game.

In a similar way, we have found a "parallel" axiomatic characterization of the symmetric coalitional binomial semivalues as power indices (that is, restricted to $\mathcal{SG}_N^{cs}$) that we state without proof because it is very similar to that of Theorem 1.

Corollary 1. *Let $p \in [0,1]$. For any N there is a unique coalitional power index on $\mathcal{SG}_N^{cs}$ that satisfies the coalitional p–binomial total power property, the transfer property, the dummy player property, symmetry within unions and symmetry in the quotient game. It is the restriction of the symmetric coalitional p–binomial semivalue Ω^P to $\mathcal{SG}_N^{cs}$.*

Besides, this index satisfies the quotient game property and reduces to the (restricted) p–binomial semivalue ψ^p when $P = P^n$ and to the Shapley–Shubik power index φ when $P = P^N$.

In the following example, we shall apply some values and coalitional values (mainly ψ^p and Ω^p) to the analysis of an interesting political structure: the current Catalonia Parliament. All values have been computed using the multilinear extension technique.

In the papers by Straffin [20], Laruelle [9] and Laruelle and Valenciano [10], the Banzhaf value β is suggested as a power measure more suitable than the Shapley value. The natural generalization to semivalues has been argued by Laruelle and Valenciano [11], Carreras and Freixas [3], and Carreras, Freixas and Puente [5]. By considering here binomial semivalues, we look at the Banzhaf value in perspective, as will be shown by the results of our analysis.

Therefore, our study of alliances will be based on the bargaining process corresponding to the symmetric coalitional binomial semivalues Ω^p: first, a power notion is shared among unions in the quotient game by means of the Banzhaf value or a binomial semivalue; then, the power so got by each union is shared among its members by using the Shapley value. This will reflect that both bargaining steps are of different nature. Notice that, once an alliance is formed —and, especially, if it supports a coalition government—, cabinet ministries, parliamentary and institutional positions, budgets, and other political responsibilities have to be distributed efficiently among the parties of the coalition, hence in a way as closely as possible to the one suggested by the Shapley value. At this point, the quotient game property and symmetry in the quotient game become very relevant properties. In fact, they are connected because if a coalitional value satisfies the quotient game property (as is the case for all Ω^p) and it is a coalitional value of the Banzhaf value (or a p–binomial semivalue) then symmetry in the quotient game follows from the anonymity of β (or of ψ^p).

Which is the reason for letting p range from 0 to 1? Notice that a reasonable regularity assumption on players' behavior is that the probability to form coalitions follows a monotonic (increasing or decreasing) behavior. Then, it is not difficult to see that the only semivalues such that $p_{k+1} = \lambda p_k$ for all k (maybe the simplest form of monotonicity) are precisely the p–binomial semivalues, in which case $\lambda = \frac{p}{1-p}$ for any $p \in [0,1]$. For example, $p = 0.1$ means that the players are very reticent to form coalitions, whereas $p = 0.8$ means that great coalitions are likelier. The neutral case $p = 0.5$ corresponds

to the Banzhaf value. Table 1 shows, for $n = 5$, the weighting coefficients of ψ^p for several values of p.

Table 1. Weighting coefficients of some p–binomial semivalues ψ^p for $n = 5$

	$p = 0.1$	$p = 0.4$	$p = 0.5$ (Banzhaf)	$p = 0.8$
$p_0 = (1-p)^4$	0.6561	0.1296	0.0625	0.0016
$p_1 = p(1-p)^3$	0.0729	0.0864	0.0625	0.0064
$p_2 = p^2(1-p)^2$	0.0081	0.0576	0.0625	0.0256
$p_3 = p^3(1-p)$	0.0009	0.0384	0.0625	0.1024
$p_4 = p^4$	0.0001	0.0256	0.0625	0.4096

As we will see, almost all allocations $\psi_i^p[v]$ and coalitional allocations $\Omega_i^p[v; P]$ will show factors $p(1-p)$. Furthermore, the maximum or the minimum of all these allocations for each player will be attained in case $p = 0.5$, that respectively correspond to the Banzhaf value $\beta = \psi^{1/2}$ or to the Alonso–Fiestras coalitional value $\Pi = \Omega^{1/2}$. These properties would not have been discovered if only the case $p = 0.5$ were considered.

Example (*The Catalonia Parliament, Legislature 2003–2007*). Five parties elected members to the Catalonia Parliament (135 seats) in the elections held on 16 November 2003, giving rise to a seat distribution that can be represented by the weighted majority game $v = [68; 46, 42, 23, 15, 9]$.

Let us briefly describe ideologically the agents in this game:

1: CiU (Convergència i Unió), Catalan nationalist middle–of–the–road coalition of two federated parties.
2: PSC (Partit dels Socialistes de Catalunya), moderate left–wing socialist party, federated to the Partido Socialista Obrero Español.
3: ERC (Esquerra Republicana de Catalunya), radical Catalan nationalist left–wing party.
4: PPC (Partit Popular de Catalunya), conservative party, Catalan delegation of the Partido Popular.
5: ICV (Iniciativa per Catalunya–Verds), coalition of Catalan eurocommunist parties, federated to Izquierda Unida, and ecologist groups ("Verds").

Notice that

$$W^m(v) = \{\{1,2\},\{1,3\},\{1,4,5\},\{2,3,4\},\{2,3,5\}\},$$

so that players 2 and 3 on one hand, and 4 and 5 on the other, are symmetric in v.

We show in Table 2 the evaluation of v given by several binomial semivalues ψ^p. The total power is $\tau^p[v] = \sum_{i \in N} \psi_i^p[v]$.

Table 2. Initial power distribution in the Catalonia Parliament 2003–2007

	$\psi_i^p[v]$	$p=0.1$	$p=0.4$	$p=0.5$	$p=0.8$
1. CiU	$p(1-p)(2+2p-2p^2)$	0.1962	0.5952	0.6250	0.3712
2. PSC	$p(1-p)(1+2p-2p^2)$	0.1062	0.3552	0.3750	0.2112
3. ERC	$p(1-p)(1+2p-2p^2)$	0.1062	0.3552	0.3750	0.2112
4. PPC	$p(1-p)(2p-2p^2)$	0.0162	0.1152	0.1250	0.0512
5. ICV	$p(1-p)(2p-2p^2)$	0.0162	0.1152	0.1250	0.0512
$\tau^p[v]$	$p(1-p)(4+10p-10p^2)$	0.4410	1.5360	1.6250	0.8960

It is easy to see that the allocations found for p and $1-p$ would coincide because the game is *decisive* (proper and strong). Notice that the proportions between the allocations to the players decrease as p approaches 0.5 from any of the extreme possibilities (0 or 1). Also notice that the maximum allocation (power) for any player and the maximum total power are got for $p=0.5$ (Banzhaf value).

Next we are interested in the study and comparison of the politically most plausible coalition structures. In each case, we have computed the coalitional value Ω^p for all $p \in [0,1]$ and also Π (for $p=1/2$) and the coalitional p–binomial total power $T^p[v;P] = \sum_{i \in N} \Omega_i^p[v;P]$. The results are as follows:

- The left–wing majority alliance PSC+ERC+ICV. The corresponding coalition structure is $P=\{\{2,3,5\},\{1\},\{4\}\}$, and the coalitional values are:

$$\Pi[v;P] = (0, 5/12, 5/12, 0, 2/12),$$
$$\Omega^p[v;P] = \left(0, \frac{1+p-p^2}{3}, \frac{1+p-p^2}{3}, 0, \frac{1-2p+2p^2}{3}\right),$$
$$T^p[v;P] = 1.$$

  Notice that $\Omega_i^p[v;P] > \psi_i^p[v]$ for all $p \in [0,1]$ and $i=2,3,5$, and also that $p=0.5$ gives the maximum of $\Omega^p[v;P]$ for PSC and ERC but, at the same time, the minimum of $\Omega^p[v;P]$ for ICV.
  Incidentally, in this case $B[v;P] = (0, 3/8, 3/8, 0, 1/8)$, so that B fails to satisfy the quotient game property and the sharing of the dictatorial power is by no means convincing because of its inefficiency.
- The catalanist majority alliance CiU+ERC. The corresponding coalition structure is $P=\{\{1,3\},\{2\},\{4\},\{5\}\}$, and the coalitional values are:

$$\Pi[v;P] = (5/8, 0, 3/8, 0, 0),$$
$$\Omega^p[v;P] = \left(\frac{1+p-p^2}{2}, 0, \frac{1-p+p^2}{2}, 0, 0\right),$$
$$T^p[v;P] = 1.$$

In this case $\Pi_i[v;P] = \beta_i[v]$ but $\Omega_i^p[v;P] > \psi_i^p[v]$ for all $p \in [0,1]$ and $i = 1,3$ (unless $p = 0.5$, where the equality holds). Here $p = 0.5$ gives the maximum of $\Omega^p[v;P]$ for CiU and the minimum for ERC.

A most convenient way to analyze these evaluations of the coalitional behavior will consist of considering different values of p, and we will take 0.1, 0.4, 0.5 (this gives Π) and 0.8. Tables 3–6 show all these particular allocations but we prefer the following order: $p = 0.5$, $p = 0.4$, $p = 0.8$ and $p = 0.1$.

Table 3. Evaluation according to ψ^p and Ω^p for $p = 0.5$

scenario	value	CiU	PSC	ERC	PPC	ICV
initial (no alliance)	β	0.6250	0.3750	0.3750	0.1250	0.1250
PSC+ERC+ICV	B	0	0.3750	0.3750	0	0.1250
PSC+ERC+ICV	Π	0	0.4167	0.4167	0	0.1667
CiU+ERC	$B = \Pi$	0.6250	0	0.3750	0	0

The important point arises when considering the majority formation. According to the Owen–Banzhaf value B, forming a winning coalition does not change the power of its members with regard to the initial distribution, although it serves to reduce the outside parties to a null position. Instead, from the viewpoint of the symmetric coalitional Banzhaf value Π, coalition PSC+ERC+ICV clearly increases the power of each one of its members, and hence it suggests to ERC the convenience to choose this coalition (which also satisfies its partners, PSC and ICV) instead of CiU+ERC.

Therefore, we have to point out here that after a short period of negotiations, precisely concerning these two options, alliance PSC+ERC+ICV was actually formed and got the regional government of Catalonia, ending 23 years of CiU governments headed by Jordi Pujol (under absolute majority of this party or with the parliamentary support of PPC). The actual sharing of positions gave the presidency of the government to Pasqual Maragall (PSC) but the presidency of the Parliament and the "Conseller en cap" position (a sort of vice–presidency of the government) to Ernest Benach and Josep Lluis Carod Rovira (both ERC), respectively. The remaining cabinet positions ("conselleries") were distributed in the proportion 8:5:2 among the three parties.

Table 4. Evaluation according to ψ^p and Ω^p for $p = 0.4$

scenario	CiU	PSC	ERC	PPC	ICV
initial (no alliance)	0.5952	0.3552	0.3552	0.1152	0.1152
PSC+ERC+ICV	0	0.4133	0.4133	0	0.1733
CiU+ERC	0.6200	0	0.3800	0	0

We recall that the allocations on this (decisive) game for a given p are the same as for $1 - p$, so that our comments on Table 4 ($p = 0.4$) are the same as they would be for $p = 0.6$, and the analogue holds for Tables 5 and 6.

By comparing the results given in Table 3 with those of Table 4, where it is assumed that players are not indifferent to join a coalition of any size but, rather, they prefer not too big coalitions (as $p = 0.4$), here we find that, not only in the case of PSC+ERC+ICV but also in the case of CiU+ERC, every party entering such a coalition clearly increases its power. However, from ERC's viewpoint, coalition PSC+ERC+ICV gives again the best fraction of coalitional power.

Table 5. Evaluation according to ψ^p and Ω^p for $p = 0.8$

scenario	CiU	PSC	ERC	PPC	ICV
initial (no alliance)	0.3712	0.2112	0.2112	0.0512	0.0512
PSC+ERC+ICV	0	0.3867	0.3867	0	0.2267
CiU+ERC	0.5800	0	0.4200	0	0

It is worthy of mention that almost all (initial or coalitional) power allocations given in Table 5 are lower than in the previous cases. The only exceptions are for ICV in PSC+ERC+ICV and ERC in CiU+ERC. The new feature here is that, in these circumstances ($p = 0.8$), ERC would clearly prefer CiU+ERC instead of PSC+ERC+ICV.

Table 6. Evaluation according to ψ^p and Ω^p for $p = 0.1$

scenario	CiU	PSC	ERC	PPC	ICV
initial (no alliance)	0.1962	0.1062	0.1062	0.0162	0.0162
PSC+ERC+ICV	0	0.3633	0.3633	0	0.2733
CiU+ERC	0.5450	0	0.4550	0	0

Finally, Table 6 exhibits the same trends as Table 5 but they are even strengthened. Again, ERC would prefer CiU+ERC, and notice that the increase of its power in agreeing to form this coalition would be greater than in the previous case.

It is not difficult to see that ERC would prefer option PSC+ERC+ICV instead of CiU+ERC if, and only if, $p \in (\frac{5-\sqrt{5}}{10}, \frac{5+\sqrt{5}}{10})$, would remain indifferent if $p = \frac{5\pm\sqrt{5}}{10}$ and would prefer CiU+ERC if $p \notin [\frac{5-\sqrt{5}}{10}, \frac{5+\sqrt{5}}{10}]$.

As a conclusion of our analysis, we find that the evaluation of games and games with a coalition structure by means of binomial semivalues and symmetric coalitional binomial semivalues provides a new approach to the study of the coalitional bargaining. Some general properties sketched only on the basis of this instance should deserve further attention.

References

1. J.M. Alonso and M.G. Fiestras. Modification of the Banzhaf value for games with a coalition structure. Annals Oper. Res., 109: 213–227, 2002.

2. J.M. Alonso, F. Carreras, and M.G. Fiestras. The multilinear extension and the symmetric coalition Banzhaf value. Working Paper MA2–IR–02–00031, Dep. of Applied Mathematics II, Polytechnic Univ. of Catalonia, Spain, 2002. Forthcoming in Theory and Decision.
3. F. Carreras and J. Freixas. Semivalue versatility and applications. Annals Oper. Res., 109: 343–358, 2002.
4. F. Carreras and A. Magaña. The multilinear extension and the modified Banzhaf–Coleman index. Mathematical Social Sciences, 28: 215–222, 1994.
5. F. Carreras, J. Freixas, and M.A. Puente. Semivalues as power indices. European J. Oper. Res., 149: 676–687, 2003.
6. I. Dragan. Some recursive definitions of the Shapley value and other linear values of cooperative TU games. Working paper 328, Univ. of Texas at Arlington, 1997.
7. P. Dubey, A. Neyman, and R.J. Weber. Value theory without efficiency. Math. Oper. Res., 6: 122–128, 1981.
8. V. Feltkamp. Alternative axiomatic characterizations of the Shapley and Banzhaf values. Int. J. Game Theory 24: 179–186, 1995.
9. A. Laruelle. On the choice of a power index. IVIE Discussion Paper WP–AD99–10, Inst. Valenciano de Investigaciones Económicas, Valencia, Spain, 1999.
10. A. Laruelle and F. Valenciano. Shapley–Shubik and Banzhaf indices revisited. Math. Oper. Res., 26: 89–104, 2001.
11. A. Laruelle and F. Valenciano. Semivalues and voting power. Discussion Paper 13, Dep. of Applied Economics IV, Basque Country Univ., Spain, 2001.
12. G. Owen. Multilinear extensions of games. Management Science 18: 64–79, 1972.
13. G. Owen. Multilinear extensions and the Banzhaf value. Naval Res. Logistics Quarterly 22: 741–750, 1975.
14. G. Owen. Values of games with a priori unions. In: Mathematical Economics and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76–88, 1977.
15. G. Owen. Characterization of the Banzhaf–Coleman index. SIAM J. Appl. Math., 35: 315–327, 1978.
16. G. Owen. Modification of the Banzhaf-Coleman index for games with a priori unions. In: Power, Voting and Voting Power (M.J. Holler, ed.), 232–238, 1982.
17. G. Owen and E. Winter. Multilinear extensions and the coalitional value. Games and Economic Behavior 4: 582–587, 1992.
18. M.A. Puente. Aportaciones a la representabilidad de juegos simples y al cálculo de soluciones de esta clase de juegos (in Spanish). Ph.D. Thesis. Polytechnic University of Catalonia, Spain, 2000.
19. L.S. Shapley. A value for n-person games. In: Contributions to the Theory of Games II (H.W. Kuhn and A.W. Tucker, eds.), Princeton Univ. Press, 307–317, 1953.
20. P.D. Straffin. The Shapley–Shubik and Banzhaf power indices. In: The Shapley Value: Essays in Honor of Lloyd S. Shapley (A.E. Roth, ed.), Cambridge Univ. Press, 71–81, 1988.
21. M. Vázquez, A. van den Nouweland, and I. García–Jurado. Owen's coalitional value and aircraft landing fees. Mathematical Social Sciences 34: 273–286, 1997.
22. R.J. Weber. Probabilistic values for games. In: The Shapley Value: Essays in Honor of Lloyd S. Shapley (A.E. Roth, ed.), Cambridge Univ. Press, 101–119, 1988.

Part IV

Industrial Applications and Numerical Testing

Complementarity Problems in Restructured Natural Gas Markets

Steven Gabriel[1] and Yves Smeers[2]

[1] Assistant Professor, Department of Civil and Environmental Engineering, Applied Mathematics and Scientific Computation Program, University of Maryland, College Park, Maryland 20742 U.S.A. sgabriel@umd.edu

[2] Tractebel Professor of Energy Economics, Department of Mathematical Engineering and Center for Operations Research and Econometrics, Université catholique de Louvain, Louvain-la-Neuve, Belgium. smeers@core.ucl.ac.be

Summary. The restructuring of the gas industry did not so far generate the same modeling activity as in electricity. While the literature of activity in electricity market models is now abundant, it is still rather scant on the gas side. This paper surveys some of the existing models and attempts to take advantage of the wealth of knowledge available in electricity in order to develop relevant models of restructured gas markets. The presentation is in three parts. The first one gives a blueprint of the market architectures inherited from the European and North American gas legislation. It also introduces a prototype optimization model and its interpretation in terms of perfect competition between agents operating on the restructured market. The second part extends the model to the case where marketers have market power. The third part considers more complex issues related to regulation of access to the network and existence of market power with different types of agents. Equilibrium models are commonly formulated as complementarity problems and the same mathematical programming framework is adopted here. Many models are single stage; there are generally easy to formulate and well known computationally. But many phenomena require two stage models that are much more intricate and on which much less is known. The paper is thus also aimed at pinpointing possible avenues for mathematical programming research.

1 Introduction

Natural gas markets in Europe and North America have recently witnessed significant changes brought about by government regulation and other market forces. An example of a regulatory measure is the U.S. Federal Energy Regulatory Commission (FERC) order 636, requiring open access service to qualified shippers (www.ferc.gov). In essence, this order transformed gas pipelines from buyers, transporters, and sellers of gas to open access transporters paving the way for new entities such as marketers to become more significant players that

might exert market power. In the European Union, similar legal measures for dividing the gas sellers and network operators have also been considered [13] as part of the restructuring and deregulation of the natural gas markets.

The EU currently imports 45 % of its natural gas [32] and this share is expected to rise given limited resources in the EU [5]). Four countries, Russia, Norway, Algeria, and the Netherlands accounted for some 87.7% of all EU gas imports in 2001 ([8] from Energie Bulletin 4145 p.5). Given the declining resources of the United Kingdom, the Netherlands will be the only major internal supplier in the coming years [5]. The potential for market power among the few producers is apparent and natural gas supply security has been addressed in the so-called "Green Paper" [12] and the European Commission (EC) directive 2004/67/EC. The increase in natural gas demand is driven in part by environmental concerns such as the Kyoto Protocol [42] and the fact that natural gas has a lower carbon content than oil or coal [33]. Other reasons for increased importance of natural gas such as the long-term supply situation, or cost-effectiveness are important factors as well.

From a modeling perspective, the traditional system optimization approach for the restructured natural gas markets in Europe and North America will not be the best choice. First, and in contrast with electricity, the gas market, whether in North America or in Europe has never been an integrated system amenable to a full optimization problem given the potentially divergent interests of the main players. Second, given the realities of the new marketplace, such models will fail to capture the important (potential) oligopolistic behavior of market players (e.g., producers in Europe, marketers in Europe or North America).

In general, the introduction of competition in the network industries (e.g. electricity, telecommunication, natural gas) stems from the following idea: one should keep (or allow to be kept), a single company for those activities considered as a natural monopoly, i.e., not competitive by default. One should allow entry in other activities to permit competition.

One way to model this mixture of regulated and non-regulated behavior, with the latter being either perfect or imperfect competition, is to depict all the market players solving separate optimization problems. The Karush-Kuhn-Tucker (KKT) conditions [2] for these optimization problems taken together with market-clearing conditions constitute a market equilibrium problem typically expressed as a nonlinear complementarity problem (NCP) [30]. Complementarity problems or the related variational inequality problems (VI) have been studied in a variety of engineering and economic settings for a number of years [4]. However, only recently have there been NCP/VI (hereafter called "complementarity") models of the natural gas market with full market detail. Some previous examples of imperfect competition models for natural gas markets in Europe concentrating on specific market segments include the early works of [31] and [11] who considered Nash-Cournot producers and a Stackelberg production market, respectively. These works concentrating on the production side were extended for example, in [29] and [6], who con-

sidered stochastic aspects and a duopoly of producers, respectively. These models all departed from traditional system optimization approaches such as maximizing total surplus [40] but lacked sufficient market detail on all the players as might be found in large-scale, detailed system optimization market models such as: the Natural Gas Transmission and Distribution Model of the U.S. Department's National Energy Modeling System and its predecessors ([1, 15, 18, 38, 39]), the Gas Systems Analysis Model for the North American market ([19, 20]), to name just a few.

Two recent models, have combined both sufficient market detail with the complementarity approach for the new markets in Europe and North America. The first model, GASTALE ([3, 14]) based in part on the work by [25, 26]. considers Nash-Cournot producers with conjectured supply functions for the European market. In addition it also includes perfectly competitive transportation and storage sectors combined with multiple consumption sectors and seasons. Gabriel et al. [21, 22, 23] have developed a model of the North American natural gas market in which marketers compete non-cooperatively against each other as Nash-Cournot players with the transporation, production, storage, and peak gas sectors taken to be perfectly competitive. Also, multiple seasons and consumption sectors are modeled.

Given the recent restructuring in natural gas markets and their importance to the energy sector, an analysis of appropriate modeling formulations is needed. This is the main goal of this paper. In Section 2 we briefly describe the functions of the various market players and provide a simple illustrative example to clarify. We also recall some mathematical programming paradigms that are used in the rest of the paper. In Section 3 we describe as a starting point, perfectly competitive behavior for these players and analyze the resulting KKT conditions for each of the players' optimization problems. Section 4 contrasts this behavior with imperfect competition among some of the players, analyzing the key differences. Both the perfect competition model of Section 3 and the imperfect competition models of Section 4 are relatively easy mathematical programming problems. Section 5 introduces more difficult considerations. It introduces transmission problems that it treats both in an average cost and Ramsey-Boiteux context. The first model guarantees neither the existence nor the uniqueness of an equilibrium. The second model is a non-convex optimization problem. More complex situations of imperfect competition are treated in Section 6, where one envisions situations where different classes of agents operating in the gas market may have market power. This leads to two-stage equilibrium problems that may not have pure strategy solutions. Many of these models have not been treated yet in the literature. The paper is thus a survey of work to be done as well. This is the message developed in the conclusion.

2 Natural Gas Market Players

The supply chain for natural gas begins with producers that extract gas from either onshore or offshore reservoirs. The producers can be assumed to potentially exert market power (as is the case in Europe) or behave in a manner consistent with perfect competition (as in North America). The next step is to transport the gas from production sites to either storage facilities, the citygate, or directly to the consumption sectors (e.g., residential, commercial, industrial, and power generation). Pipeline companies own and operate these transportation routes and are subject to regulated rates (e.g., by FERC in the U.S. and by National Regulatory authorities (NRA) in the EU). Storage operators take advantage of seasonal arbitrage by buying and injecting gas into storage in the low demand season (non-winter) and then selling it to consumers in the high demand season (winter). Storage operators can be taken to be regulated or oligopolistic depending on the local regulations. The EU Directive 2003/55/EC has much weaker regulatory requirements on storage than on transport. It only imposes access to storage on negotiated terms but does not impose any price regulation. Owners of storage facilities are thus only subject to general competition law and possibly to any additional regulatory obligation imposed by the Member States where their facilities are located. Marketers (also known as shippers) are responsible for contracting with pipeline companies to procure the gas and sell it to end-users. The marketers are generally less subject to national regulation and can reasonably be modeled as players with market power given their important position and the new deregulated marketplace for natural gas. Specifically, in the EU, Directive 2003/55/EC does not impose any regulation to marketers which are thus only subject to general competition law and possibly to the regulation that individual Member States may find necessary. Additonally, one can also consider peak demand players who supply extra gas in times of high demand. This supply may be in the form of liquefied natural gas (LNG) or propane/air mixtures. Perfect or imperfect competition could be appropriate for these players as well.

It should be noted that these players each can be modeled as solving an optimization problem in which an abstraction of their operations is assumed. For example, production rates are constrained by the number of available rigs, pressure in the reservoirs, and so on. A full consideration of all engineering aspects for these and the other players would no doubt lead to intractable, non-convex problems, thus making the computation of a complementarity-based equilibrium very difficult at best. For these reasons, an abstraction of their operations is generally taken. Also, it is important to note that in some cases, one parent company may have control over several levels of the natural gas supply chain described above. However, regulations are in effect to try to balance out the field between independent players and ones which are part of a larger company operating on several levels of the supply chain [24]. In the

European Union, this control of concentration is left to general competition law.

2.1 An Illustrative Example

To clarify how the natural gas market can be modeled, consider the following example, simplified from Gabriel et al. [21] and depicted in Figure 1.

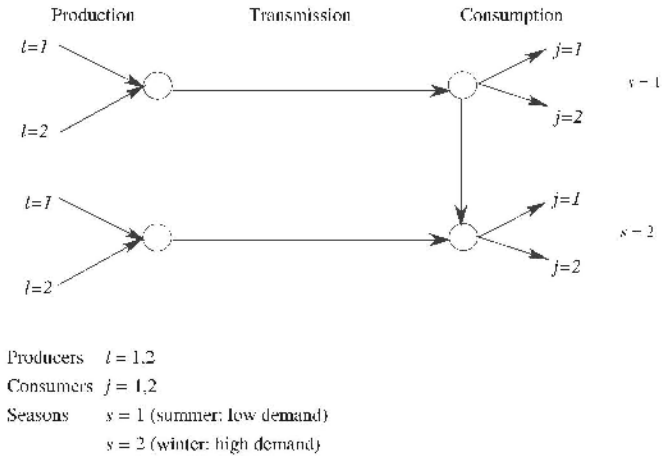


Fig. 1. A simplified example

There are two producers separated from the market by a pipeline (we neglect the distribution system). Storage facilities are located at the end of the pipeline close to the market. There are two market segments, residential and industrial. The problem refers to a single-year horizon decomposed into two seasons. Demand is low in the first season and high in the second. Gas is stored in the low season and extracted in the high season. In order to simplify notation, we assume that both seasons have the same number of days. We also neglect all losses whether from transmission or storage operations. Lastly, depending on the case at hand, it is useful to consider domestic pipelines (entirely located in a country), crossborder pipelines (e.g. crossing an European border) or long distance transportation pipelines (e.g. bringing Russian gas to European borders through Ukraine). Each of these raises new questions which are not treated here. Instead we assume a single pipeline for clarity.

2.1.1. Production

The description of the production of natural gas is reduced to a function giving the cost of extracting the quantity of gas,$(q_{\ell 1}, q_{\ell 2})$ for seasons 1 and 2,

respectively, for producer ℓ.

$$\begin{aligned} \text{Cost} &= \textstyle\sum_s EC_{\ell s}(q_{\ell s}) \\ & q_{\ell 1} \geq 0, q_{\ell 2} \geq 0 \\ & \text{(EC for Extraction Cost)} \end{aligned} \tag{1}$$

All engineering complexity associated with extracting the gas from the reservoir is thus bypassed. Stylized descriptions of this type are frequently adopted in economics where these functions are used to construct analytical models. In contrast, computable models rely on formulations that allow for more detailed engineering descriptions of the gas production process [19, 20]. Even though we use a stylized representation of the cost function, such as found in economic models, we keep in mind that one should be able to replace it by a process model of gas production at least as long as one remains within descriptions commonly amenable to optimization models (e.g. [4]).

2.1.2. Transportation

Pipeline transportation is also represented in a very simplified form that neglects all technological characteristics arising from the pressure and flow relationship or representation of compressors; see [10] for details. We simply assume that thepipeline has a maximum capacity $\overline{f}_s$ based on the flow f_s. The owner of the pipeline incurs both short and long-run transportation costs. The costs for the pipeline owner is represented as follows

$$\begin{aligned} \text{Cost} &= \textstyle\sum_s TC_s(f_s) \\ & \overline{f}_s - f_s \geq 0 \qquad f_s \geq 0 \qquad s = 1,2 \\ & \text{(TC for Transportation Cost)} \end{aligned} \tag{2}$$

2.1.3. Storage

Storage is also modeled in the simplest possible form for ease of presentation. Because we neglect losses, the amount recovered in the high demand season is equal to the amount injected in the low demand season. Injection and withdrawal operations respectively cause injection and withdrawal costs and there is also a maximum injection rate. Because there are only two periods, this maximal injection rate also limits the amount that can be stored. Additional constraints on volumetric rates could also be included but are left out from this simple example. The maximal injection rate thus also plays the role of a storage volumetric constraint related to working gas in the reservoir. The associated costs are as follows

$$\begin{aligned} \text{Cost} &= IC(i) + WC(w) \\ & i - w \geq 0 \\ & \overline{i} - i \geq 0, w \geq 0 \\ & (IC\text{: injection cost; } WC\text{: withdrawal cost}) \end{aligned} \tag{3}$$

where i, w are the injection and withdrawal amounts, respectively, with $\overline{i}$ the injection capacity.

2.1.4. Demand

In general, the demand for each sector will be a function of the price in that sector, which itself is a (decision) variable to be endogenously determined. For illustrative purposes, we assume a fixed demand in each season. We let

$$d_{js} \text{ be the demand of consumer } j \text{ in season } s. \tag{4}$$

2.1.5. From Transaction Costs to Marketer's Costs

The supply of natural gas involves procuring gas from the producers, securing transportation and storage services and selling the gas to the final consumers. These activities imply transaction costs for an integrated company such as the one described by a single optimization model. Because these activities will take on a different interpretation later, it is convenient to single them out in preparation for the rest of the paper. We therefore define the following variables that will later be bundled into a marketer activity. Specifically the marketing department of the integrated company

procures gas from producer ℓ in season s	$mq_{\ell s}$	
procures transmission services in season s	mf_s	
procures storage services (injection and withdrawal)	mi and mw	(5)
sells gas to customer j in season s	md_{js}	

with the first letter m denoting that it is a marketing variable. For the sake of simplification we shall only refer to the transactions costs due to the selling of the gas (variable md_{js}) and not consider the other marketing variables in the following. We let mn_{js} be the unit cost of selling gas to consumer j in season s.

2.1.6. A System Optimization Model

As a point of comparison, it is natural to first state the overall natural gas problem in standard production management terms or as a system optimization problem. Specifically, there are costs to produce, transport and store the gas before delivering it to the final consumers. There are also transaction costs of coordinating these activities. As indicated above, we limit our description to the sole transaction costs incurred because of the marketing of gas (sales activity). The most efficient approach in optimization terms is to minimize the sum of all these costs given as

$$
\begin{aligned}
\min \textstyle\sum_{\ell} \sum_{s} EC_{\ell s}(q_{\ell s}) + \sum_{s} TC_s(f_s) + IC(i) \\
+ WC(w) + \textstyle\sum_{j} \sum_{s} mn_{js} \cdot md_{js} \\
\text{s.t.} \quad \textstyle\sum_{\ell} q_{\ell s} - f_s \geq 0 \qquad s = 1,2 \\
f_1 - i \geq \textstyle\sum_j md_{j1} \\
f_2 + w \geq \textstyle\sum_j md_{j2} \\
i - w \geq 0 \\
\overline{f} - f \geq 0 \\
\overline{i} - i \geq 0 \\
md_{js} \geq d_{js} \qquad s = 1,2; j = 1,2 \\
q_{\ell s}, f_s, i, w \geq 0
\end{aligned}
\tag{6}
$$

Problem (6) is a very simplified representation of a natural gas market operated by a single integrated company. The primary usefulness is to serve as a basis of comparison for more complicated models to be presented below. Indeed, our goal is to progressively transform this small problem with the view of encompassing some of the concerns typically faced by market analysts, regulators, and economists. We assume throughout the paper that all cost functions are convex and differentiable. This approximation is commonly made in economic models. Differentiability can be relaxed at the cost of more complex formulations that we prefer to avoid in this paper. Adding an assumption of quadratic function would also make our complementarity problems linear complementarity problems (LCP).

This type of approach has been extensively used in the discussion of the restructuring of the electricity industry. Many arguments have been developped on the basis of electric power models, comparatively as simple as problem (6) and were eventually transformed into full size computable models for looking at policy and strategic questions. We adopt the same philosophy: starting from a simple optimization model, we progressively introduce economic questions that reflect some of the aspects of the restructuring of the natural gas sector. While there has been considerable modeling activity along these lines in the electricity sector, this has not taken place yet in the gas sector.

The approach is also interesting from an optimization point of view. Some of the models emerging from the process are standard complementarity problems which are now well understood. Other models are optimization problems subject to equilibrium constraints. These problems are much more recent even though their literature is already abundant. Also, other models are equilibrium problems subject to equilibrium constraints, a particular case of Generalized Nash Equilibrium problems. These are quite recent models that turn out to be quite difficult to analyze and computationally challenging to solve. At this stage, such models have received little attention in the literature. Last, but possibly not least, the simplified mathematical programming problems formulated here, can easily be made more challenging by adding all the technological complexities neglected in this presentation.

Classes of Mathematical Programming Problems

Before proceeding with building up a more complicated model of (6), we recall the KKT conditions, complementarity problems, and other mathematical programs that are relevant. A detailed discussion of the properties of these various mathematical programs as well as applications thereof can be found in [4]. We use throughout the notation $0 \leq a \perp b \geq 0$ which expresses the set of relations

$$a \geq 0 \qquad b \geq 0 \qquad ab = 0.$$

1. Karush-Kuhn-Tucker Conditions for a Convex Optimization Problem: Consider a standard nonlinear programming problem of the form

$$\begin{array}{ll} \min & f(x) \\ s.t. & g_i(x) \leq 0 \;\; i = 1, \ldots, m \\ & h_j(x) = 0 \;\; j = 1, \ldots, p \end{array}$$

where $f, g_i : R^n \to R$, are convex functions and $h_j : R^n \to R$ are affine functions. The KKT conditions are then sufficient for optimality ([4]). These conditions are to find a decision vector $x \in R^n$, an inequality Lagrange multiplier vector $u \in R^m$, and an equality Lagrange multiplier vector $v \in R^p$ such that

$$\begin{array}{ll} \nabla f(x) + \sum_i u_i \nabla g_i(x) + \sum_j v_j \nabla h_j(x) = 0 & \\ g_i(x) \leq 0, u_i \geqslant 0, g_i(x) u_i = 0 & \forall i \\ h_j(x) = 0, \;\; v_j \text{ unconstrained} & \forall j \end{array}$$

These KKT conditions are a special case of a nonlinear complementarity problem with both equations and inequalities, called a mixed complementarity problem (MCP) and given as follows.

2. Mixed Complementarity Problems: Find $x \in R^{n_1}, y \in R^{n_2}$ such that

$$\begin{array}{l} 0 \leq F(x,y) \perp x \geq 0 \\ 0 = G(x,y) \end{array}$$

where $F : R^{n_1} \times R^{n_2} \to R^{n_1}$, $G : R^{n_1} \times R^{n_2} \to R^{n_2}$ and in general. (These problems are monotone in the context of this paper.)

More generally, one may want to optimize a certain function $\Pi(x,y,z)$ of three sets of variables $x \in R^{n_1}, y \in R^{n_2}, z \in R^{n_3}$. The z vector represents the "first stage" variables whereas x and y represent the "second stage" variables. A typical constraint set consists of two sets of restrictions. First, there are regular constraints on the upper level variables of the form $z \in S$. Secondly, the second stage variables must satisfy some mixed complementarity problem for fixed values of the first stage variables z. This problem is given as follows.

3. Mathematical Programming Problem Subject to Equilibrium Constraints (MPEC):

$$\begin{aligned} &\max_{x,y,z} \Pi(x,y,z) \\ &\text{s.t. } 0 \leq F(x,y;z) \perp x \geq 0 \\ &\qquad 0 = G(x,y;z) \\ &\qquad z \in S \end{aligned}$$

which is in general a non-convex problem and computationally challenging. A well-known example of an MPEC is the bilevel programming problem in which the lower level constraints are the optimality conditions for a second-stage problem.

MPEC problems can be generalized to equilibrium problems with equilibrium constraints (EPEC). A specific example of an EPEC is as follows: Let K agents have first stage decision variables $z_k, k = 1, \cdots, K$. Each of these agents seeks to maximize an objective function $\Pi^k(x, y, z_k, z^*_{-k})$ where z^*_{-k} represents the optimal but fixed values for the other players. This objective function is optimized subject to the constraint that $z_k \in S_k$ and equilibrium constraints such as specified in the MPEC problem. The full problem is thus to find $z^*_k, k = 1, \cdots, K, x, y$ as follows.

4. Equilibrium Problems subject to Equilibrium Constraints:

$$\begin{aligned} z^*_k \text{ solves } &\max_{x,y,z_k} \Pi^k(x,y,z_k,z^*_{-k}) \\ \text{s.t. } &0 \leq F(x,y;z_k,z^*_{-k}) \perp x \geq 0 \\ &0 = G(x,y;z_k,z^*_{-k}) \\ &z_k \in S_k \end{aligned}$$

This problem, like the MPEC, is computationally difficult given that it is in general a non-convex problem and existence of a solution (here a pure strategy equilibrium) is not guaranteed even under standard compactness assumptions on the feasible region. The solution of an EPEC problem, if it exists, is a subgame perfect equilibrium [17].

3 A Perfect Competition Model

3.1 Demand Functions

Market models commonly assume that demand reacts to prices. Short-term (real time) demand of electricity is the exception where demand is commonly assumed to be insensitive to price. This reaction is represented by a demand function, which, for concreteness, we assume to be affine and downward sloping. We let

$$d_{js}(p_{js}) \text{ and } p_{js}(d_{js}) \tag{7}$$

be, respectively, the demand and inverse demand functions of consumer j, in season s. Using the inverse demand function, one introduces the willingness to pay function given as

$$WP_{js}(d_{js}) = \int_0^{d_{js}} p_{js}(\xi)d\xi. \tag{8}$$

We assume, in order to simplify the discussion, that the prices will automatically turn out positive.

3.2 Basic Assumptions

Perfect competition assumes that all agents are price-takers. This means that agents optimize their profit or utility subject to prices that they take as given. This assumption does not imply that these prices are exogenous to the system, but simply that these agents see them as such. An expanded version of problem (6) more amenable to an interpretation in terms of an equilibrium is given in problem (9) in which all variables are taken to be nonnegative. In preparation for its interpretation in terms of an equilibrium model, this version also assumes that the marketing/sales activity of the integrated company has been split in several marketing/sales activities k each under the responsability of a different independent marketer k, with its own activity variable and cost. Problem (9) differs from problem (6) in two respects. First it explicitly introduces the demand functions (7) via the willingness to pay function (8). Second it reformulates the constraints by introducing new variables that are easier to interpret in terms of unbundled gas activities. Specifically this latter difference between the two formulations allows for an explicit representation of all the transactions of the marketers and the introduction of a possibly different unit marketing cost mn^k of each marketer k in the objective function. It also separates the production, transportation and storage activities.

$$\begin{aligned}\max\ & \textstyle\sum_j \sum_s WP_{js}(d_{js}) - \sum_\ell \sum_s EC_{\ell s}(q_{\ell s}) - \sum_s TC_s(f_s) \\ & - IC(i) - WC(w) - \textstyle\sum_k mn^k(\sum_j \sum_s md^k_{js})\end{aligned} \tag{9.1}$$

s.t.

$$q_{\ell s} - \textstyle\sum_k mq^k_{\ell s} \geq 0 \quad (wp_{\ell s}) \quad \text{wellhead price} \tag{9.2}$$

$$\textstyle\sum_\ell mq^k_{\ell s} - mf^k_s \geq 0 \quad (bp^k_s) \quad \text{border price} \tag{9.3}$$

$$mf^k_1 - mi^k - \textstyle\sum_j md^k_{j1} \geq 0 \quad (cg^k_1) \quad \text{citygate price} \tag{9.4}$$

$$mf^k_2 + mw^k - \textstyle\sum_j md^k_{j2} \geq 0 \quad (cg^k_2) \quad \text{citygate price} \tag{9.5}$$

$$\textstyle\sum_k md^k_{js} - d_{js} \geq 0 \quad (p_{js}) \quad \text{price paid by consumer} \tag{9.6}$$

$$f_s - \textstyle\sum_k mf^k_s \geq 0 \quad (\tau_s) \quad \text{transmission price} \tag{9.7}$$

$$mi^k - mw^k \geq 0 \quad (\mu^k) \quad \text{value of gas in storage for marketer } k \tag{9.8}$$

$$i - \textstyle\sum_k mi^k \geq 0 \quad (ip) \quad \text{injection charge} \tag{9.9}$$

$$w - \textstyle\sum_k mw^k \geq 0 \quad (wp) \quad \text{withdrawal charge} \tag{9.10}$$

$$\bar{f}_s - f_s \geq 0 \quad (\rho_s) \quad \text{transmission congestion} \tag{9.11}$$

$$\bar{i} - i \geq 0 \quad (\lambda) \quad \text{storage congestion charge} \tag{9.12}$$

Dual variables are written to the right of each constraint together with their interpretation. The dual variables of constraints (9.2) to (9.5) are respectively,

the wellhead prices (wp), border prices (bp), and citygate prices in summer and in winter (cg). The other dual variables can also be usefully interpreted. Specifically p_{js} is the price paid by consumer j in season s; τ_s is the transmission charge in season s, ip and wp respectively the injection and withdrawal charges into and from storage; ρ_s is the congestion charge of the pipeline, λ the congestion charge of storage facilities, and μ^k is the implicit price of gas in storage for marketer k. Note that except for the introduction of the possibly different transaction costs of the marketing activity and the addition of different marketing variables, this model is equivalent to problem (6). As we argue next, because of the new variables, it is amenable to an interpretation in terms of the behavior of the agents in the market. Note also that balance inequalities are written under the "free disposal assumption" i.e., they hold as equalities when the commodity/service price is positive.

3.3 *KKT* Conditions, Complementarity Formulations and Agent Behavior

We now proceed to establish the KKT conditions of problem (9) and interpret them in terms of agent behavior in perfect competition. This interpretation paves the way to the introduction and formulation of different assumptions of imperfect competition.

3.3.1. Producers' Behavior

The relation

$$0 \leq \frac{\partial EC_{\ell s}(q_\ell)}{\partial q_{\ell s}} - wp_{\ell s} \perp q_{\ell s} \geq 0. \tag{10}$$

expresses that each producer maximizes its profit at the prevailing price in the season. If producer ℓ is active in season s ($q_{\ell s} > 0$), then the wellhead price is equal to the marginal cost.

3.3.2. Pipeline Operator Behavior

The conditions

$$\begin{aligned} &0 \leq \frac{\partial TC_s(f_s)}{\partial f_s} - \tau_s + \rho_s \perp f_s \geq 0 \\ &0 \leq \overline{f}_s - f_s \perp \rho_s \geq 0 \end{aligned} \tag{11}$$

state that the pipeline operator maximizes the profit accruing from the use of the pipeline at the prevailing price. If the pipeline is used ($f_s > 0$), this price is equal to the sum of a marginal transportation cost and a congestion cost ($\tau_s = \frac{\partial TC_s(f_s)}{\partial f_s} + \rho_s$). The congestion cost ρ_s is only different from zero when the pipeline is full ($f_s = \overline{f}_s$).

3.3.3. Storage Operator Behavior

The conditions describing the behavior of the storage operator can be stated as follows.

$$\begin{aligned} 0 &\leq \frac{\partial IC(i)}{\partial i} - ip + \lambda \perp i \geq 0 \\ 0 &\leq \frac{\partial WC(w)}{\partial w} - wp \perp w \geq 0 \\ 0 &\leq (\bar{i} - i) \perp \lambda \geq 0 \end{aligned} \tag{12}$$

These define storage operation charges and can be interpreted as follows. There is a charge λ on injection facilities only when there are congested, i.e. $\lambda > 0$ implies $i = \bar{i}$. If the injection facilities are used ($i > 0$) the injection charge is equal to the sum of the marginal injection cost and the congestion charge: $ip = \frac{\partial IC(i)}{\partial i} + \lambda$. The withdrawal charge is equal to the marginal withdrawal cost when $w > 0$: $wp = \frac{\partial WC}{\partial w}$.

3.3.4. Consumer Behavior

The condition

$$0 \leq -\frac{\partial WP_{js}}{\partial d_{js}} + p_{js} \perp d_{js} \geq 0 \tag{13}$$

expresses that the marginal willingness to pay for gas is equal to the price when there is consumption, that is, $d_{js} > 0$ implies $\frac{\partial WP_{js}}{\partial d_{js}} = p_{js}$.

3.3.5. Marketers' Behavior

The appearance of marketers is a key element of the restructuring of the gas industry. Marketers emerge from the optimization models as agents that take on former coordination activities that involved procuring the commodity and transportation and storage services as well as marketing the gas. They compete against each other, and as a result put competitive pressure on other agents that are not in a monopoly position (e.g. producers in the EU). Each of the marketer's tasks is described in complementarity form as follows.

3.3.6. Procuring the Gas

$$0 \leq wp_{\ell s} - bp_s^k \perp mq_{\ell s}^k \geq 0. \tag{14}$$

When $mq_{\ell s}^k > 0$, the border price charged to marketer k is equal to the wellhead price.

3.3.7. Shipping the Gas

$$0 \leq bp_s^k - cg_s^k + \tau_s \perp mf_s^k \geq 0. \tag{15}$$

When $mf_s^k > 0$, the citygate price of marketer k is equal to the sum of the border price charged to marketer k and the transmission price.

3.3.8. Procuring Storage Services

$$\begin{aligned} 0 &\leq ip + cg_1^k - \mu^k \perp mi^k \geq 0 \\ 0 &\leq \mu^k + wp - cg_2^k \perp mw^k \geq 0, \\ mi^k &= mw^k. \end{aligned} \tag{16}$$

Note that relation (9.8), $mi^k \geq mw^k$, must hold with equality. Indeed, suppose $mi^k > mw^k \geq 0$, then $\mu^k = 0$ by (9.8). This also implies $ip = cg_1^k = 0$ (by the first complementarity condition of (16)). $i \geq \sum_{k'} mi^{k'} > mi^k > 0$, (12) would then imply $\frac{\partial IC}{\partial i} = 0$, ... which means that the cost of the whole supply chain vanishes to zero in season 1. We exclude this case for economic reasonableness.

The difference of citygate prices between seasons 2 and 1 ($cg_2^k - cg_1^k$) for marketer k is equal to what it has to pay for storage services ($ip + wp$) when it uses these services ($mi^k = mw^k > 0$). This is an intertemporal arbitrage condition.

3.3.9. Marketing the Gas

$$0 \leq cg_s^k + mn^k - p_{js}(d_{js}) \perp md_{js}^k \geq 0. \tag{17}$$

When $md_{js}^k > 0$, the price offered to consumer j in season s is equal to the sum of the citygate price of marketer k and the marketing cost mn^k.

4 Imperfect Competition: Market Power of the Marketers

4.1 Background and Definition of the Agents

The above discussion is rather straightforward both in mathematical and economic terms. It is well known that KKT conditions of convex problems can be expressed as complementarity conditions and that they can be interpreted in economic terms under our assumptions of convexity (see Section 2.1.6). But this economic interpretation is very specific. It only refers to perfect competition, that is to conditions where all agents are price takers. The interest of the KKT conditions in this model stems from the fact that we would like to modify each of these complementarity conditions in order to better represent the reality of the market. Indeed, European producers do not necessarily behave as price-taking agents. Transmission may be regulated both in the US and Europe resulting in their charging their average cost. Marketers may have a dominant position in their home market in Europe or in some large fraction of the market in the US and hence not behave according to the perfect competition paradigm. Similarly storage owners could be regulated or be in a position to exert market power. In short, one would like to construct a model that resembles the above KKT model at least in terms of its structure, but differs from it in specific market aspects. We begin by briefly motivating this approach.

4.1.1. Unbundling of the Transportation and Merchant Activities

It is commonly assumed, but by no means proved in theory or practice, that the transportation infrastructure is a natural monopoly. This implies that

one should not expect competition or, at least much competition to develop, in transportation. We take the extreme view (which is true in Spain and France but not in Germany) that there is a single transportation company operating the infrastructure. Transportation, because it is a monopoly should be regulated both in terms of the conditions of access and its pricing. In other words one cannot expect that competition will naturally lead to relation (11). One would thus need to impose some pricing regulation on the transportation activity.

4.1.2. Unbundling of Storage and Merchant Activity

Storage is essential for gas operation. Storage can only be developed at certain sites and the incumbent European companies currently already operate most sites. It would seem natural to also unbundle storage from the marketing operations. This can be done in two ways: one is to make storage competitive, that is either to transfer ownership to other agents or to auction its capacity; an alternative is to regulate the access to storage. For the sake of brevity we shall not elaborate here on the regulation of the storage activity or on the market power that storage owners can exert. For the sake of simplicity we retain the perfect competition assumption model in (16).

4.1.3. Making Marketing Competitive

In contrast with storage and transportation, there is no restriction on having several marketers operating in a given territory. Specifically all former gas companies can have a marketing activity. Because they know the producers of gas and the main characteristics of the gas consumers, this implies that they can compete with each other in different geographic segments of the market. Needless to say, the incumbent in some European country is likely to know more about the demand sector of his country than about other countries, at least in a first stage. But this is not sufficient to refrain from entering other markets or from trying to team up with smaller agents operating in other markets. This justifies unbundling the marketing activities and allowing for different marketers in every market.

In short, we thus assume in the following that there is a single pipeline company and a single storage company. The transportation activity is regulated. We do not make any special assumption on storage that remains ruled by (12), that is, at marginal cost pricing. We suppose that there are several marketers that buy and resell gas and procure transportation and storage services, possibly exerting market power.

4.2 Price Discrimination and Arbitrage

Even though there may be several marketers in a single market, it is unlikely that it will immediately become perfectly competitive. This implies that one looks for a Nash equilibrium with respect to some strategic variables. It is common and easy to use quantities as strategic variables (à la Cournot). We

shall later use a similar Cournot assumption for representing producers. This will lead to a much more difficult EPEC problem. According to this assumption, each marketer optimizes its profit, assuming the quantitative actions of the others given. In order to illustrate the principle, consider for a moment the simpler problem of marketer k buying gas in season s at citygate prices cg_s^k respectively. These marketers incur marketing costs mn^k. In perfect competition they will sell the gas to segment j at the price p_{js} satisfying

$$p_{js}(d_{js}) = cg_s^k + mn^k, \quad k = 1, 2. \tag{18}$$

Both marketers will sell to segment j if the quantities $cg_s^1 + mn^1$ and $cg_s^2 + mn^2$ are equal. If not, only the marketer with the smallest $cg_s^k + mn^k$ will remain in that market segment.

The situation is different with a Nash-Cournot assumption. We adopt the standard notation to let $-k$ designate marketers other than k. Under this assumption and with this notation, marketer k solves the problem

$$\max_{md_s^k} p_{js}(md_s^k + md_s^{-k})md_s^k - (cg_s^k + mn^k)md_s^k. \tag{19}$$

Assuming a positive sale ($md_s^k > 0$), one sees that the pricing condition (18) is replaced by

$$p_{js}(d_{js}) + md_s^k \frac{\partial p_{js}}{\partial d_{js}} = cg_s^k + mn^k, \quad k = 1, 2. \tag{20}$$

The only difference between the perfect and Nash-Cournot competition is thus the replacement of $p_{js}(d_{js})$ by $p_{js}(d_{js}) + md_s^k \dfrac{\partial p_{js}(d_{js})}{\partial d_{js}}$.

Applying this reasoning to the previously derived KKT conditions, the Nash-Cournot behavior of the marketers can be inserted into the above model by simply replacing relation (17) by

$$0 \leq cg_s^k + mn^k - p_{js}(d_{js}) - md_{js}^k \frac{\partial p_{js}(d_{js})}{\partial d_{js}} \perp md_{js}^k \geq 0. \tag{21}$$

In this relation the gas price collected by the marketer from customer j in time segment s is replaced by the marginal revenue from the same client in that period. The rest of the KKT conditions remain unchanged.

This model is amenable to some variations. One can assume that all marketers behave à la Cournot. Alternatively, one can suppose that the incumbent marketer retains a dominant position and that the entering marketers behave competitively, that is that they are price-takers. One would then have a mix of relations (17) for the entrants and (21) for the incumbent. This could be justified for instance if an entrant believes that it is too small to try to exert market power in this new market. The entrant therefore prefers to leave the task of maintaining a relatively high gas price to the incumbent and simply behaves as a price-taker.

The possibility of having this mix of behaviors introduces alternative possible formulations. One may simply combine the competitive relations describing the Cournot (21) and competitive (17) behaviors. This is the situation where the incumbent naively considers the actions of the entrant as given. Alternatively, one could assume that the incumbent takes the actions of the entrant into account when planning its strategy. It then chooses its marketing action taking into account the reaction of the entrant. This latter interpretation complicates the problem as shown below.

Price discrimination does not occur in perfect competition but is a standard outcome of market power. In order to analyze this phenomenon, consider again the perfect competition model and suppose that marketer k supplies both consumers 1 and 2 in season s $(md_{1s}^k > 0, md_{2s}^k > 0)$. Relation (17) becomes

$$p_{1s} = p_{2s} = mn^k + cg_s^k. \tag{22}$$

One sees that the prices paid by the two customers in season s are identical. Consider now the Cournot model and make the similar assumption that marketer k supplies both consumers in season s. Relation (21) becomes

$$\begin{aligned} &p_{1s}(d_{1s}) + md_{1s}^k \frac{\partial p_{1s}}{\partial d_{1s}}(d_{1s}) - mn^k - cg_s^k = 0 \text{ and} \\ &p_{2s}(d_{2s}) + md_{2s}^k \frac{\partial p_{2s}}{\partial d_{2s}}(d_{2s}) - mn^k - cg_s^k = 0. \end{aligned} \tag{23}$$

This time one cannot conclude that $p_{1s} = p_{2s}$. The prices to the two consumers could be different and therefore price discrimination could occur. An interesting question is whether price discrimination can persist in an open market. This is where new agents, namely arbitrageurs, intervene.

Arbitrageurs are new agents that take advantage of price differences existing in a market. They buy where the price is lower and sell where it is higher if the difference exceeds their transaction costs. Suppose, in order to simplify the problem, different consumer prices as results from the Cournot pricing of the marketers, zero transportation costs between the customers and negligible transaction costs are present. The following modeling of arbitrageurs has been introduced by [37] for the electricity sector and is presented here for natural gas. Suppose an arbitrageur that buys a quantity a from a first consumer paying a lower price and sells this amount to a second consumer with a higher price. The arbitrageur can make a profit and will expand this trading until the prices of both consumers are equal. This can be formalized by imposing that an arbitrageur solves the following problem

$$\max_{as} [p_{1s}(d_{1s} + a_s^*) - p_{2s}(d_{2s} - a_s^*)]a_s \qquad (a_s \text{ unconstrained}) \tag{24}$$

where a_s represents the amount that is arbitraged. It is important to note that the a_s^* in $p_{1s}(d_{1s} + a_s^*)$ and $p_{2s}(d_{2s} - a_s^*)$ is not a decision variable to the arbitrageur (based on the perfect competition assumption). The arbitrageur is supposed to be a price taker. He/she trades as long as $p_{1s} \neq p_{2s}$ but does

not take the impact of his/her trade on the price into account. This is the usual assumption of a competitive agent: it implies that the market settles at a value a_s for which

$$p_{1s}(d_{1s} + a_s^*) - p_{2s}(d_{2s} - a_s^*) = 0. \tag{25}$$

This effect can be readily inserted in the model (10) to (17) by adding both the variables a_s and the constraints (25), $s = 1, 2$ to the set of complementarity conditions.

Price discrimination can also take place between seasons. Suppose that the marketer uses storage services. In perfect competition (17) implies that the difference between the prices charged to a given consumer is equal to the difference between the citygate prices in these seasons (see the discussion of storage operations in section 3.3.5). This difference is itself equal to the sum of the marginal injection and withdrawal costs, to which one also adds a congestion cost in case the storage capacity is full. This is expressed in the following relation

$$\begin{aligned} p_{j1} - cg_1^k - mn^k &= 0 \\ p_{j2} - cg_2^k - mn^k &= 0 \end{aligned} \tag{26}$$

which imply

$$p_{j1} - p_{j2} = cg_1^k - cg_2^k.$$

Taking the Cournot assumption where the prices charged to a consumer in the two seasons have been replaced by the marginal revenues accruing from these consumers, one obtains

$$\begin{aligned} p_{j1}(d_{j1}) + md_{j1}^k \frac{\partial p_{j1}}{\partial d_{j1}}(d_{j1}) - cg_1^k - mn^k &= 0 \\ p_{j2}(d_{j2}) + md_{j2}^k \frac{\partial p_{j2}}{\partial_{j2}}(d_{j2}) - cg_2^k - mn^k &= 0. \end{aligned} \tag{27}$$

This does not imply that $p_{j2} - p_{j1} = cg_2^k - cg_1^k$. The price difference between the two seasons is not necessarily equal to the difference between the citygate prices and hence to the sum of the marginal injection and withdrawals charges and a possible congestion cost. In other words, there may be price discrimination. This price discrimination between seasons has been pointed out for the case of reservoir management in electricity in [7]. It also appears here in natural gas. The implication of the market power here is a non-optimal use of the storage compared to the perfect competition case. Arbitrageurs can again intervene to reduce the price discrimination between seasons. An arbitrageur here is an agent who buys a quantity in the first, low-price period and releases it in the higher price period. The arbitrageur does not buy gas from the producers (it would be an other marketer in that case); he/she simply takes a position between the two periods. Needless to say the arbitrageur incurs the storage costs, in this case the sum of the marginal injection and withdrawal costs and the possible congestion charge in case storage facilities are full. The arbitrageur therefore solves the following problem

$$\max \left[p_{j2}(d_{j2} + a^*) - p_{j1}(d_{j1} - a^*) - \left(\frac{\partial IC(i + a^*)}{\partial} + \frac{\partial WC(w + a^*)}{\partial w} + \lambda \right) \right] a \tag{28}$$

where he/she takes $\frac{\partial IC}{\partial i}$, $\frac{\partial WC}{\partial w}$ and λ as given. Solving the problem will imply that the prices between two seasons will satisfy the relation

$$p_{j2}(d_{j2} + a^*) - p_{j1}(d_{j1} - a^*) = \frac{\partial IC(i + a^*)}{\partial i} + \frac{\partial WC(w + a^*)}{\partial w} + \lambda. \tag{29}$$

Again this effect can be readily inserted in the model (10)-(17) by adding both the variables a and relation (29) to the set of complementarity conditions at least if one assumes that one has an analytic expression of both $\frac{\partial IC}{\partial i}$ and $\frac{\partial WC}{\partial w}$. One also needs to replace $\bar{i} \geq i$ be $\bar{i} \geq i + a$. We saw before that the Cournot marketer could anticipate the actions of the spatial arbitrageurs expressed in relation (25) (clairvoyant marketer) or take them as given (myopic marketer). The same distinction can be made here with respect to the behavior of marketers vis à vis the seasonal arbitrageurs (relation (29)). The case of the naïve arbitrageur is straightforward to model: one simply replaces relation (21)-(22) by the pair

$$\begin{aligned} &0 \leq cg_s^k + mn^k - p_{is}(d_{js} + a) - md_{js}^k \tfrac{\partial p_{js}}{\partial d_{js}} \perp md_{js}^k \geq 0 \\ &p_{1s}(d_{1s} + a) - p_{2s}(d_{2s} - a) = 0. \end{aligned} \tag{30}$$

In contrast with the naïve Cournot marketer, the clairvoyant marketer foresees the action of the arbitrageur and takes them into account in its sales. Metzler et al. [37] have shown that both assumptions lead to the same outcome in electricity markets. It is conjectured that the same result holds here. The reader is referred to [37] for an in-depth discussion of this question.

5 Regulated Transportation

5.1 Background

It was argued before that there will likely remain a single transportation company in each EU Member State after restructuring has taken place. This transportation company therefore has a dominant position in the transportation market and hence needs to be regulated. Germany is the only proponent of an alternative approach and argued for a long time that transportation of natural gas is a competitive activity. And indeed some competition developed. But Directive 2003/55/EC applies to all Member States and Germany will need to comply with the common approach which is to regulate gas transportation. It remains to be seen how it will meet the regulation requirement. Regulation should facilitate the proper access to transportation infrastructure. The exact meaning of "proper" has been extensively discussed in the literature on access

pricing in network industries (mainly in telecommunication). We note that our formulation (11) implements a marginal cost pricing of transportation services and a congestion charge when the capacity of the pipeline is saturated. This congestion cost is charged to all marketers. Marginal cost pricing has been vigorously discussed in the context of access to the electric power network where it gave rise to the famous disputes between proponents of the flowgate and nodal models and to the discussion of zonal/nodal pricing in the United States. It also gives rise to various issues of market power in the transportation of electricity. We shall not discuss these questions here because in contrast with electricity, congestion in natural gas transport does not seem yet to be a major issue. Besides marginal cost pricing we consider two other approaches to transportation pricing, namely average cost and Ramsey-Boiteux pricing. Average cost pricing is the most widely accepted tariff structure in practice even though it has little economic virtue. By contrast, Ramsey Boiteux pricing is a sophisticated way to allocate costs. Its application to utilities was made famous by Boiteux's seminal contribution to electricity pricing. It has been extensively discussed in the context of access to telecommunication infrastructure. Its application to natural gas is due to [9]. We model these approaches without any attempt to summarize the extensive discussions that they generated.

5.2 Average Cost Pricing

Average cost pricing is the preferred access pricing method in practice. It consists of setting a price that allows the network owner to cover its cost including a proper rate of return on capital. To illustrate the principle, consider the simple situation depicted in Figure 2 with two marketers. One assumes that the charge is set at regular time intervals by the regulator on the basis of the transportation cost and on some historical or prospective view of the flow in the pipeline.

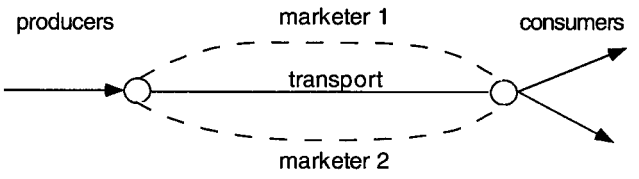


Fig. 2. two marketers and a transporter

Let tc and F be respectively the variable and the fixed cost of the network (see Figure 3). Assume two marketers who respectively ship f_1 and f_2 through the network. A plausible average cost access tariff is given by the unit rate τ_s

$$\tau_s = \frac{F}{f_1 + f_2} + tc \tag{31}$$

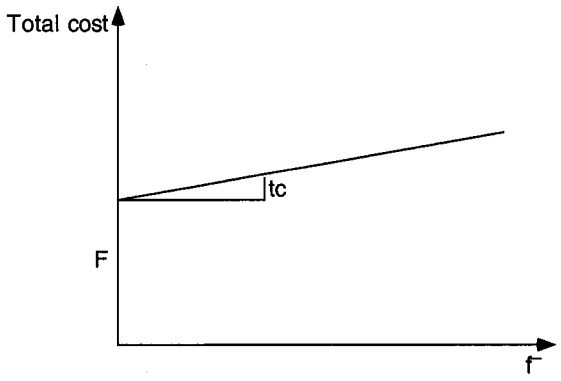


Fig. 3. Cost function of the transportation activity

One can again think of two possible implementations of this tariff. In a first "naïve" implementation, the marketers do not foresee that increasing the amount of demanded transmission service will decrease the unit rate τ_s. In another interpretation, they anticipate this change. Replacing (11) by (31) and keeping the rest of the KKT conditions unchanged models the naïve interpretation.

5.3 Infeasible Problems and Multiple Equilibria

All models covered up to this point can be converted into convex optimization problems, at least under standard conditions on the cost function (convexity), the demand curve (downward sloping) and for the Cournot model, revenue function (concavity). They are thus guaranteed to have a convex set of solutions. In contrast the introduction of average cost pricing prevents this conversion into a convex optimization problem. The complementarity problem becomes nonlinear and ceases to be monotone as a result of the decreasing unit rate τ_s (31) replacing (11). This may make the model infeasible or introduce multiple equilibria. This is illustrated in Figure 4 for the cost function of the transportation activity shown in Figure 3. The example assumes a single consumer, no storage, zero marginal gas production cost and a marketer who needs to pay for transportation priced at average cost. The figure illustrates two situations that correspond to different levels of the fixed charge F. Curve (1) corresponds to the case of relatively low value of F; the average cost curve intersects with the demand curve at two points so that here are two equilibria.

When the fixed charges of the pipeline are too high (curve (2)), the transporter cannot find a demand level that pays for the cost of the network. This lack of equilibrium may seem unrealistic if the fixed charges are limited to the sole cost of the network. This phenomenon proved dramatically relevant before the restructuring of the US gas sector in the 1980's when the fixed charges to be recovered by the marketers (at that time the pipelines companies) included the take or pay commitments of long-term contracts.

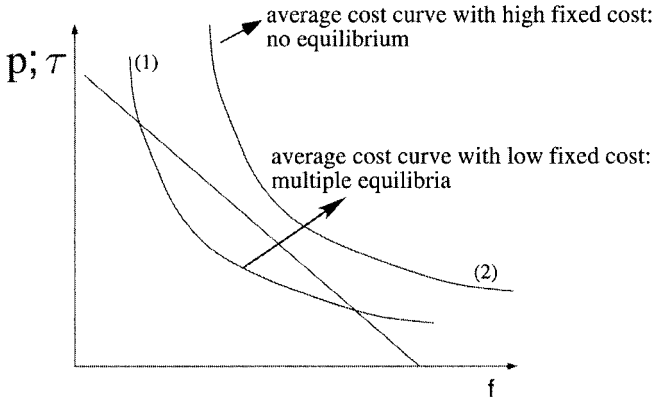


Fig. 4. Non existing and multiple equilibria

5.4 Ramsey-Boiteux Problem Statement

Economists working on access pricing in the telecommunication area have extensively promoted the application of Ramsey-Boiteux pricing for access to the infrastructure. Cremer et al.[9] converted this approach to transportation of natural gas. We first introduce the method in a simplified context and then discuss the problem that it raises in the more realistic context (even though extremely simplified) of our example.

Consider the simplified case where there is no storage, a single marketer, one gas producer and two customers as shown in Figure 4.

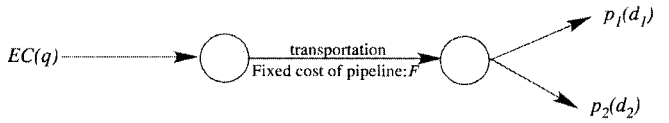


Fig. 5. One producer, one marketer, no storage

Assume the charge to recover through access prices amounts to a single fixed cost of the pipeline ($tc = 0$ in Figure 3). One wants to find access charges for the two customers that maximize economic welfare and allow one to cover the revenue requirement of the pipeline.

Supposing that the whole economy is in perfect competition except for the transportation of natural gas. We note q the production quantity and use τ^j to denote the transport charge to consumer j in the example. τ^j is then given by

$$\tau^j = p_j(d_j) - \frac{\partial EC}{\partial q}(q) \tag{32}$$

where $\frac{\partial EC}{\partial q}$ is the price charged by the producer for its gas in perfect competition. The transport charge is equal to the difference between the price paid by the consumer and the marginal extraction cost of gas. Relation (32) implies that q can be written as a function of d_j and τ^j: let $q(d, \tau)$ be this function. The resulting welfare maximization problem in simplified form is stated as

$$\begin{array}{l} \max \int_0^{d_1} p_1(\xi_1)d\xi_1 + \int_0^{d_2} p_2(\xi_2)d\xi_2 - EC(q) \\ \text{s.t. } \tau^1 d_1 + \tau^2 d_2 \geq F \\ \quad 0 \leq q \leq \overline{f}. \end{array} \tag{33}$$

Note that the KKT conditions of (33) are similar but not identical to (32). (33) is indeed the regulator's problem while (32) represents the equilibrium conditions in a perfectly competitive market.

This formulation assumes that there exists a benevolent regulator that tries to maximize the overall welfare while simultaneously covering the fixed charge of the network. The formulation assumes that the marketer procures the gas at marginal cost which corresponds to a perfectly competitive production market. Alternative assumptions are possible. A perfectly competitive gas production is a quite reasonable in North America, but not in Europe. Whatever the assumption of competition on the production side, Ramsey-Boiteux introduces access charges that are specific to the consumer segment. The consumer which values gas more pays more. This is price discrimination but it is accepted in this context because of the objective pursued, namely an efficient pricing of the infrastructure. In U.S. parlance, the discrimination is not undue. We do not discuss this legal and economic issue here.

Consider the formulation given in (33) and the transmission charges τ^1 and τ^2. The (perfect competition) equilibrium conditions of the rest of the gas market can be written as

$$\begin{array}{ll} p_j(d_j) & = bp + \tau^j \\ \sum d_j & = q \end{array} \tag{34}$$

where $bp = \partial E/\partial q$ is the border price in the one producer case. This is a square system, which means that the production and demand are entirely

determined by bp, τ^1 and τ^2. The Regulator is only responsible for choosing τ^1 and τ^2 while the market will select bp on the basis of q. The Regulator therefore optimizes a criterion that effectively depends on d_j and q by playing on the τ^j. Assume a quadratic cost function EC, then bp is affine. Because we also assumed affine demand functions, the dependence of all variables d on τ^1 and τ^2 is affine. The maximization problem of the Regulator is thus convex. Economists have elaborated at length on the analytic solution of this problem.

The same reasoning could have been made if marketers behaved à la Cournot. The relationship (32) would have been replaced by Cournot equilibrium equations, that is, by replacing the price by the marginal revenue. These resulting expressions would have been affine. The problem of the Regulator would have been different from an economic point of view, but its mathematical structure would have remained unchanged. In both cases, Ramsey pricing is amenable to an analytic solution. Things become much more complex when one turns to a more detailed physical model where the square system of equation (34) is replaced by a complementarity problem.

5.5 Applying Ramsey-Boiteux to the Example

The above reasoning can be considered in the more general case of our example. Assume as before that the transport charges (τ_{js}) differentiated by customer and season are known. All the other variables of the market are determined by the equilibrium conditions that describe the behavior of all agents except the transporter. Specifically one defines a restricted equilibrium subproblem RESP(τ) consisting of the following complementarity conditions

- Producer's behavior (10)
- Storage operation behavior (12)
- Consumer behavior (13)
- Marketers behavior (14) to (17)
- All balance inequalities (9.2) to (9.10) holding as equalities.

One notes that the pipeline operator equations (11) that involved the transport charges τ_s are not part of the subproblem. They have been replaced in RESP(τ) by exogenous assumptions on the τ. The result is a well defined restricted equilibrium subproblem RESP(τ) parametrized by the τ_{js}.

RESP(τ) is a complementarity problem, which in this case is equivalent to an optimization problem. It has a convex set of solutions which reduces to a single point when the marginal cost of the producers and the demand functions are affine and non-constant. It is thus possible to define the Ramsey pricing problem using the same philosophy as before: the Regulator selects the τ_{js} in order to maximize a function that depends on the d_{js} and $q_{\ell s}$. While the objective function is concave in these variables, it is no longer concave in the τ_{js}. The relation between the former and the latter is indeed piecewise affine in this case because it is the solution of a linear complementarity problem that

is parametrized in τ. This problem is a mathematical programming problem subject to equilibrium constraints (MPEC) as discussed above (see [4]). Note that the formulation can encompass different variants of the restricted equilibrium subproblem. Specifically, there is no difficulty accommodating Cournot marketers instead of perfectly competitive marketers. The variants on arbitrageurs that we discussed in this problem can also be included.

6 Cournot Producers

6.1 A First Model (see [36])

Both the former "gas companies" and the gas producers had market power in the pre-restructuring European market. In contrast gas producers can be seen as largely competitive in the US. The study of market power in the European gas sector through complementarity problems began in Norway and combined both economic analysis and computational methods. Specifically, [36] modeled the European gas market under three assumptions of competition, namely perfect competition, monopoly, and the now standard Cournot assumption. By comparing the results obtained to observation, they concluded that the Cournot model was a realistic representation of the European market of the time. Mathiesen [34, 35] also showed how complementarity problems could be used to solve equilibrium models. We begin our discussion of the market power of the producers by casting this early work in our example that we simplify somewhat further. Consider a hypothetical gas company (that is, a company that bundles merchant, transmission and storage activities) operating in the pre-restructured period. It is regulated at cost and can only charge the sum of the procurement cost and a fixed mark-up that represents its average costs and some previously agreed upon margin. We let $ac^{\ell j}$ be this mark-up when the company procures gas at producer ℓ's location and sells it to market j. Neglect storage operation and assume a single season. Let p_j be the price in market j. A producer ℓ selling to the consumer market j receives a netback $p_j - ac^{\ell j}$ as shown in Figure 6.

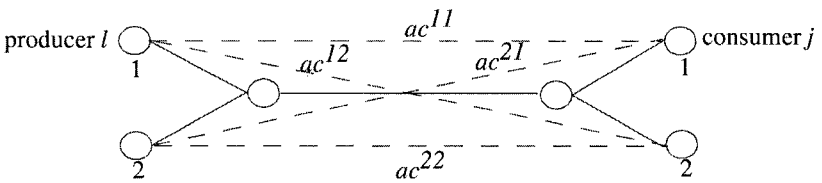


Fig. 6. No storage, fixed gas company margin

The behavior of the Cournot producer 1 can then be described by the following optimization problem

$$\begin{aligned} \max_{md_1^1, md_2^1} \; & p_1(md_1^1 + md_1^{-1})md_1 + p_2(md_2^1 + md_2^{-1})md_2^1 \\ & -ac^{11}md_1^1 - ac^{12}md_2^1 - EC_1(md_1^1 + md_2^1) \\ & \text{s.t. } md_1^1 \geq 0, md_2^1 \geq 0 \end{aligned} \tag{35}$$

where the optimization is carried out on the variables md_1^1 and md_2^1, keeping the sales md_1^{-1} and md_2^{-1} fixed. Mathiesen [36] formulated and solved this problem on a bipartite transportation network with European producers ℓ being the left-hand nodes and European markets being the right hand nodes. The ac gave the transportation costs. Different extensions of Mathiesen et al.'s work were made that ultimately led to the GASTALE model mentioned in the introduction ([3]). Golombeck et al. [25] examined the impact of the introduction of the first European Gas Directive by assuming that it would lead to arbitrage between gas prices inside the border of the European Union. In other words, arbitrageurs would trade gas between the different border points so as to eliminate the price differences that would not be justified by transporting costs. Golombeck et al. [26] also examined the impact of abolishing export monopolies in the exporting countries. In all these studies, the marketing company was represented by an exogenously given overall cost and margin that we noted ac. In contrast, GASTALE introduced market power both at the producer and marketer side. The representation of the latter was simplified with respect to Gabriel et al. [21] in order to make the example more tractable. It does so by implementing an oligopolistic version of the economic notion of double marginalization [27, 28]. See [41] for a discussion for the monopoly case.

6.2 Double Marginalization and the GASTALE Model

The structure of the European gas market suggests that both the producers and the marketers have market power. The question is whether this duality of market power can be accommodated in computational models. GASTALE extends Golombeck's model to account for this phenomena [3].

Mathiesen et al.'s original model briefly recalled above assumes that the marketers simply add a mark-up to the price that they get from the producers. In other words the margin between the price paid by the consumer and the marginal cost of the producers is shared by the producers and the marketers but the part of the latter is fixed. This is the case when one assumes that all transportation and storage costs are exogenously given and the profit of the marketer is regulated. The price charged by a marketer to a consumer is thus equal to the price at the wellhead plus the sum of the price of transportation and storage including some regulated margins. [3] consider an extension of this view where the marketers behave competitively or à la Cournot. Specifically these authors assume a given number of identical marketers in each market

that equally share the demand in that market. All segments are served and hence each marketer sells an equal quantity to each segment. Using this property, Boots et al. can relate the prices charged to the different segments of the final demand to the price charged by the producer to the marketers. Their model is a mix of computational and analytical modeling. The derivation of the demand curve seen by the producers is analytical and relies on the assumption of symmetry of the marketers. The exertion of market power by the producers is computational and directly related to the previous work of [36] and [25]. An interesting objective is to remove the analytical part of this model to make it purely computational. This is necessary if we want to do away with the assumption of symmetric marketers. We shall see that Boots et al.'s approach can in principle be extended by assuming non-identical marketers that behave à la Cournot but at the price of additional computational difficulties. We consider two cases depending on whether the producers behave à la Cournot or à la Bertrand.

6.3 Bertrand Producers and Competitive or Cournot Marketers

Following the standard reasoning of double marginalization we assume that the marketers take the border price bp_s^k as given and that they can buy unlimited quantities at that price. Natural gas is normally considered as an homogeneous product after pretreatment at the well or at the beach (the "border" in bp). This suggests representing the competition of the producers à la Bertrand. The producer ℓ that sells to a marketer k at the lowest price gets all the demand in that market. If several producers sell to a marketer they do it at the same price and equally share the demand of that market. This complies with our noting bp_s^k as the price paid by marketer k "at the border" in season s. Given the bp_s^k, one can define a restricted equilibrium subproblem RESP(bp) that represents the behavior of the rest of the market by assembling the complementarity conditions that describe

- The pipeline operator behavior (11)
- The storage operator behavior (12)
- The consumer behavior (13)
- The marketer behavior (14) to (17)
- All balance inequalities (9.2) to (9.10) holding as equalities.

Only the relation (10) describing the behavior of the producers is left out. It is replaced by taking the bp_s^k as parameters. Because of the integrability of the demand functions, RESP(bp) is a complementarity problem that is equivalent to an optimization problem. Introduce the notation

$$mq_s^k = \sum_\ell mq_{\ell s}^k$$

to denote the total demand of gas by marketer k in season s. This value can be derived from the solution of RESP(bp). It is unique for each vector bp_s^k if

one assumes affine demand functions as we did throughout the paper (affine demand functions are a sufficient but non necessary condition for this result). Then, let $mq_s^k(bp)$ be the demand of gas of marketer k in season s, as a function of the prices bp_s^k found as part of the solution of RESP(bp). Inserting this solution in the profit function of the producers, it is possible to define a new Nash equilibrium problem whereby the producers select the border prices $bp_{\ell s}^k$ at which they sell gas to the marketers in order to maximize their profit. The resulting problem is an equilibrium problem subject to equilibrium constraints but of a type that to the best of our knowledge has not been mentioned let alone studied in the literature. The first-stage competition is of the Bertrand type while the second stage is Cournot. The natural question is whether this model would be relevant in practice. We already indicated that production is competitive in the US so that this model would thus add very little. In contrast there is much talk of the emergence of "gas to gas competition" in the oligopolistic European gas market. A Bertrand competition, where, producers compete in price would thus be worth exploring. The interest of the problem is that, in contrast with all the other models discussed in this paper, Bertrand competition for homogeneous products cannot be modeled through complementarity formulations.

We do not explore this problem any further and turn instead to the more standard formulation where producres have limited possibilities for exerting market power through prices but do so through quantities. This leads to a full two-stage Cournot model.

6.4 Cournot Producers and Competitive or Cournot Marketers

In order to adapt the above formulation to arrive at a problem where both stages are Cournot, consider the case where marketers are not given a certain border price bp_s^k but an import quantity $mq_{\ell s}^k$. In other words producers behave strategically by restricting their sales to marketers. It is easy to see that one can restate the restricted equilibrium subproblem to accommodate this new situation where quantities are the strategic variables. Consider the restricted equilibrium subproblems RESP(mq) consisting of

- The pipeline operator behavior (11)
- The storage operator behavior (12)
- The consumer behavior (13)
- The marketer behavior (14) to (17)
- All balance inequalities (9.2) to (9.10) holding as equalities.

Again the relation (10) describing the behavior of the producer is left out and is replaced by an assignment of $mq_{\ell s}^k$.

This subproblem is again a complementarity problem, which is equivalent to an optimization problem. It has a convex set of solutions which is unique when the marginal cost of the producers are affine and non-constant. Let bp_s^k be the price of the gas found in relation (15). This price, bp_s^k is the marginal

value of the gas sold to marketer k in season s, that comes out as a solution of that suproblem. It is thus the price at which marketer k is willing to pay for the gas,when offered the quantities $mq^k_{\ell s}$. We can thus define the mapping $bp^k_s(mq_s)$. This allows one to define a new Nash equilibrium problem for the producers whereby they select the quantities $mq^k_{\ell s}$ that they sell to the marketers. This is stated in

$$\max \sum_s [\sum_k bp^k_s(mq^k_{\ell s}, mq^k_{-\ell s})mq^k_{\ell s} - EC_{\ell s}(\sum_k mq^k_{\ell s})] \qquad (36)$$
$$mq^k_{\ell s} \geq 0, mq^k_{-\ell s} \text{ fixed.}$$

There is one such intertemporal problem for each producer. The collection of these problems for the different producers and the search of a set of $mq^k_{\ell s}$ that simultaneously solves all of them is an equilibrium problem subject to equilibrium constraints (EPEC). Again, there is no real difficulty accommodating perfectly competitive marketers instead of Cournot marketers or any mix of assumptions that we have seen. The difficulty is indeed to solve such a problem.

7 Conclusions

This paper surveys some work as well as points out work that remains to be done. It considers essential problems brought about by the restructuring of the gas industry in Europe and North America for which one has relatively little knowledge and understanding. We can improve our insight of this market by modeling it on the basis of standard economic assumptions. Models of industrial organization raising questions of direct relevance to the gas market flourish in industrial economics. As it is often the case, their results differ drastically depending on their assumptions. This is confirmed by numerical experiments. As one says "the devil is in the details". The problem is that the devil has considerable potential in the important area of natural gas. It is important to add to the insight provided by economists by also exploring these questions experimentally, in this case computationally. Because of the novelty of the market, there are currently little data in Europe to validate these models. In contrast the restructured US gas market has accumulated several years of experience. This validation process is especially interesting since many of the models arising from industrial economic concepts also turn out to be quite difficult in mathematical programming terms.

References

1. B.H. Ahn and W. W. Hogan. On convergence of the PIES algorithm for computing equilibria. Oper. Res., 30: 281–300, 1982.

2. M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. Nonlinear Programming Theory and Algorithms. New York, John Wiley & Sons, 1993.
3. M.G. Boots, F.A.M. Rijkers, and B.F. Hobbs. Modelling the role of trading companies in the downstream European gas market: A successive oligopoly approach. The Energy Journal, 25(3): 73–102, 2004.
4. J.Boucher and Y. Smeers. Optimal development planning of gas reserves. Energy Economics, 18:25–47, 1996.
5. BP (2003). BP Statistics 2003, www.bp.com.
6. N. Breton and Z. Zaccour. Equilibria in an asymmetric duopoly facing a security constraint. Energy Economics, 23:457–475, 2001.
7. J. Bushnell. A mixed complementarity model of hydro-thermal electricity competition in the Western U.S., Oper. Res., 51(1):81–93, 2003.
8. Cedigaz. Energie Bulletin, n° 4145, p. 5., 2002.
9. H. Cremer, F. Gasmi, and J.-J. Laffont. Access to pipelines in competitive gas market. J. Regulatory Economics, 24(1):5–33, 2003.
10. D. De Wolf and Y. Smeers. Optimal dimensioning of pipe networks with application to gas transmission networks. Oper. Res., 44(4):596–608, 1996.
11. D. De Wolf and Y. Smeers. A stochastic version of a Stackelberg-Nash-Cournot equilibrium model. Management Science, 43(2):190–197, 1997.
12. EC. COM (2000) 769 final. Towards an European strategy for the security of energy supply. Nov.29, 2000.
13. EC. Directive 2003/55/EC concerning common rules for the internal market in natural gas. June 26, 2003.
14. R. Egging and S.A. Gabriel. Examining market power in the European natural gas market. Accepted, *Energy Policy.*
15. EIA. The national energy modeling system: an overview. Office of Integrated Analysis and Forecasting. U.S. Department of Energy, Washington, D.C., 1998.
16. F. Facchinei and J.-S. Pang. Finite-dimensional Variational Inequalities and Complementarity Problems, Volumes I and II. New York, Springer, 2003.
17. D. Fudenberg and J. Tirole. Game Theory, Cambridge, MIT Press, 2000.
18. S.A. Gabriel, A.S. Kydes, and P. Whitman. The national energy modeling system: a large-scale energy-economic equilibrium model. Oper. Res., 49:14–25, 2001.
19. S.A. Gabriel, J. Manik, and S. Vikas. Computational experience with a large-scale, multiperiod, spatial equilibrium model of the North American natural gas system. Networks and Spatial Economics, 3: 97–122, 2003.
20. S.A. Gabriel, S. Vikas, and D.M. Ribar . Measuring the influence of Canadian carbon stabilization programs on natural gas exports to the United States via a 'bottom-up' intertemporal spatial price equilibrium model. Energy Economics, 22:497-525, 2000.
21. S.A. Gabriel, S. Kiet, and J.F. Zhuang. A competitive equilibrium model for the natural gas market based on a mixed complementary formulation. Forthcoming in *Oper. Res.*.
22. S.A. Gabriel, , J.-F. Zhuang, and S. Kiet. A large-scale linear complementarity model of the North American natural gas market. Energy Economics, in press.
23. S.A. Gabriel,J.-F. Zhuang, and S. Kiet. A Nash-Cournot Model for the North American natural gas market. IAEE Conference Proceedings, Zurich, Switzerland, 2004.
24. Global Competition Review. Gas regulation in 26 jurisdictions worldwide, www.globalcompetitionreview.com, 2003.

25. R. Golombek, E. Gjelsvik, and K.E. Rosendahl. Effects of liberalising the natural gas markets in Western Europe. The Energy Journal, 16(1):85–111, 1995.
26. R. Golombek, E. Gjelsvik, and K.E. Rosendahl. Increased competition on the supply side of the Western European natural gas market. The Energy Journal, 19(3):1–18, 1998.
27. M.L. Greenhut and H. Ohta. Related market conditions and interindustrial mergers. Amer. Economic Review, 66:267–277, 1976.
28. M.L. Greenhut and H. Ohta. Vertical integration of successive oligopolists. Amer. Economic Review, 69(1):137–141, 1979.
29. G. Gürkan, A.Y. Özge, and S.M. Robinson. Sample-path solution of stochastic variational inequalities, Math. Programming, 84: 313-333, 1999.
30. P.T. Harker and J.-D. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Math. Programming, 48:161-220, 1990.
31. A. Haurie, G. Zaccour, J. Legrand, and Y. Smeers. A stochastic dynamic Nash-Cournot model for the European gas market, Tech. Rep. G-86-24, GERAD, Ecole des Hautes Etudes Commerciales, Montréal, Québec, Canada, 1987.
32. IEA. IEA Energy Statistics 2001, www.iea.org/Textbase/stats/index.asp, 2002.
33. IPCC. IPCC Guidelines for national Green House Gas inventories, Vol. 3, Ref Manual., 1995.
34. L. Mathiesen. Computational experience in solving equilibrium models by a sequence of linear complementarity problems. Oper. Res., 33:1225–1250, 1985.
35. L. Mathiesen. Computation of economic equilibria by a sequence of linear complementarity problems. Math. Programming Study, 24:144-162, 1985.
36. L. Mathiesen, K. Roland, and K. Thonstad. The European natural gas market: Degrees of market power on the selling side. In R. Golombek, M. Hoel and J. Vislie (eds.), Natural Gas Markets and Contracts. North-Holland, Amsterdam, 1987.
37. C. Metzler, B.F. Hobbs, and J.-S. Pang (2003). Nash Cournot equilibria in power markets on liberalized D.C. network with arbitrage: formulations and properties. Network and Spatial Economics, 3(2), 123–150, 2003.
38. F.H. Murphy. An overview of the intermediate future forecasting system. In A.S. Kydes et al. (eds.), Energy Modeling and Simulation, North Holland Publishing Company, 66-73, 1983.
39. F.H. Murphy, J.J. Conti, S.H. Shaw, and R. Sanders. Modeling and forecasting energy markets with the Intermediate Future Forecasting System. Oper. Res. 36:406-420, 1998.
40. T. Takayama and G. Judge. Spatial and Temporal Price and Allocation Models, North-Holland Publishing Company, London, 1971.
41. J. Tirole. The Theory of Industrial Organization. MIT Press, Cambridge (MA), 1989.
42. UNFCCC. Kyoto Protocol to the United Nations Framework Convention on Climate Change. COP 3 report, document FCCC/CP/1997/7/Add.1, Mar18, 1998.

Reconciling Franchisor and Franchisee: A Planar Biobjective Competitive Location and Design Model*

José Fernández[1], Boglárka Tóth[1,2], Frank Plastria[3], and Blas Pelegrín[1]

[1] Dpt. Statistics and Operations Research, University of Murcia, Spain.
[2] Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and the University of Szeged, Hungary.
[3] MOSI - Dpt. of Mathematics, Operational Research and Information Systems for Management, Vrije Universiteit Brussel, Belgium.

Summary. This paper deals with a hard nonlinear biobjective optimization problem: finding the optimal location and design for a new franchised facility within a region where facilities (both of the franchise and not) already exist and compete for the market. The franchisor and the new franchisee both want to maximise their own profit in the market, but these two objectives are in conflict. Customers patronize all the facilities, old and new, proportionally to their attraction to them. Both resulting objective functions are neither convex nor concave. An interval branch and bound method is proposed to obtain an outer approximation of the whole set of efficient solutions. Computational experiments highlight the different kinds of information provided by this method and by a variation of the lexicographic method.

1 Introduction

Multiobjective optimization problems are ubiquitous. Many real-life problems require taking several conflicting points of view into account. In fact, although the origins of the multiobjective optimization literature are linked to utility theory, game theory, linear production theory and economics (see [20]), we now can find applications in many and diverse fields, such as portfolio optimization [10], jury selection [42], airline operations [11], radiation therapy [33], manpower planning [43] or reservoir management [1], among others. In [49], White mentions more that 500 applications between 1955 and 1986. Classical references on multiobjective optimization are the books [4, 5, 45, 50, 51]. Other more recent books are [17, 35].

* This paper has been supported by the Ministry of Science and Technology of Spain under the research project BEC2002-01026, in part financed by the European Regional Development Fund (ERDF).

In this paper we restrict ourselves to the biobjective case, that is, to the problem

$$\begin{array}{l} \min \{f_1(y), f_2(y)\} \\ \text{s.t. } y \in S \subseteq \mathbb{R}^n \end{array} \tag{1}$$

where $f_1, f_2 : \mathbb{R}^n \longrightarrow \mathbb{R}$ are two real-valued functions. Let us denote by $f(y) = (f_1(y), f_2(y))$ the vector of objective functions, and by $Z = f(S) \subseteq \mathbb{R}^2$ the image of the feasible region. Some widely known definitions to explain the concept of solution of (1) are the following.

Definition 1. *A feasible vector $y^* \in S$ is said to be* efficient *iff there does not exist another feasible vector $y \in S$ such that $f_i(y) \leq f_i(y^*)$ for all $i = 1, 2$, and $f_j(y) < f_j(y^*)$ for at least one index j. The set S_E of all the efficient points is called the* efficient set.

Efficiency is defined in the decision space. The corresponding definition in the criterion space is as follows.

Definition 2. *An objective vector $z^* = f(y^*) \in Z$ is said to be* nondominated *(or also* efficient*) iff y^* is efficient. The set of all nondominated vectors will be denoted by Z_N.*

Ideally, solving (1) means obtaining the whole efficient set, although in practice we will be satisfied if we obtain some representative (in some sense) efficient points. There is a great and rich variety of methods with that aim, as can be seen in the references mentioned above (weighting method, constraint method, lexicographic method,...). However, most of the literature on multiobjective optimization deals with either *discrete* problems or with continuous multiobjective *linear* problems, whereas the interest in this paper is in *nonlinear* multiobjective optimization. Although less studied, we can also find in the literature many references dealing with this last topic (see [35] and the references therein).

In particular, here we consider a competitive facility location problem. Competitive location deals with the problem of locating facilities to provide a service (or goods) to the customers (or consumers) of a given geographical area where other competing facilities offering the same service are already present (or will enter to the market in the near future). Many competitive location models are available in the literature, see for instance the survey papers [8, 9, 19, 29, 38, 41, 48] and the references therein. However, the literature on multiobjective competitive location models is rather scarce. In fact, to our knowledge, [21, 47] seem to be the only references in this field. This is in part due to the fact that single objective competitive location problems are difficult to solve, and considering more than one objective makes the problem near intractable.

We study the case of a franchise which wants to enlarge its presence in a given geographical region by opening one new facility. Both the franchisor (the owner of the franchise) and the franchisee (the actual owner of the new facility

to be opened) have the same objective: maximize their own profit. However, the maximization of the profit obtained by the franchisor is in conflict with the maximization of the profit obtained by the franchisee. This suggests to use a biobjective model to obtain the efficient solutions for this problem, so that later on the franchisor and the franchisee can agree in both location and design for the new facility, taking the corresponding economical implications of their selection into account.

With the aim of obtaining a close view of the complete efficient set S_E of the resulting continuous nonlinear biobjective problem, we introduce in this paper a new method, an interval branch-and-bound algorithm, which is able to obtain a superset of S_E (a non-interval branch-and-bound method with the same aim was proposed in [24] for solving another biobjective location problem). This superset of S_E maps into a superset of the nondominated set Z_N, which may be made as tight as required (up to the precision provided by the inclusion functions used) by reducing the tolerances employed in the termination rules (see Section 3.6). To the extent of our knowledge, this is the first *general* method proposed in the literature with that purpose. The reason for this lack of methods is that even obtaining a single efficient point can be a difficult task. That is why some authors have proposed to present to the decision-maker a 'representative set' of efficient points which suitably represent the whole efficient set, either by modifying the definition of efficiency [3] or by selecting a finite set of efficient points with the criteria of coverage, uniformity and cardinality as quality measures [2, 40]. Notice that the approach in this paper is completely different. Instead of offering a (small) subset of the efficient set to the decision-maker, we offer a superset which tightly contains it. By drawing in the image space that superset the decision-maker can easily see the trade-off between the two objectives, i.e., how one objective improves as the other gets worse. Something similar can be done in the decision space, by drawing the superset in a color scale depending on the objective value of one of the objectives.

The interval B&B method deals with the multiple objectives directly. It starts with an initial box containing the feasible set. The box considered is either sent to the solution list, it is removed from further consideration by a 'discarding test', or it is split into several subboxes which are considered later. This process is repeated by choosing a new box until no box remains to be considered. We also briefly describe (and slightly modify) an interval lexicographical-like method recently proposed in [47], and show the possibilities offered by both methods for obtaining different information about the biobjective problem.

The paper is organized as follows. In the following section we present our biobjective competitive continuous location problem. In Section 3 we introduce the new interval branch-and-bound method for obtaining the whole efficient set. In the next section, the lexicographical-like method is presented, as well as a slight modification of it. Some computational studies are reported

in Section 5. The paper ends with conclusions and pointing lines for future research.

2 The Biobjective Model

A franchise wants to locate a new single facility in a given area of the plane, where there already exist m facilities offering the same good or product. The first k ($k \geq 1$) of those m ($m > k$) facilities are part of the franchise. The *demand*, supposed to be *inelastic* (which means that the goods are essential for the customers, that is, they will satisfy their full demand), is concentrated at n demand points, whose locations p_i and *buying power* w_i are known. The location f_j and quality of the existing facilities is also known.

In the spirit of Huff [26], and later generalized in [36] and [28], we consider that the *patronising behaviour* of customers is *probabilistic*, that is, demand points split their buying power among all the facilities proportionally to the *attraction* they feel for them. The attraction (or utility) function of a customer towards a given facility depends on the distance between the customer and the facility, as well as on other characteristics of the facility which determine its *quality*. The location and the quality of the new facility are the variables of the problem.

The following notation will be used throughout this paper:

x	location of the new facility, $x = (x_1, x_2)$.
α	quality of the new facility ($\alpha > 0$).
n	number of demand points.
p_i	demand points, $p_i = (p_{i1}, p_{i2})$ $(i = 1, \ldots, n)$.
w_i	demand (or buying power) at p_i.
m	number of existing facilities.
f_j	existing facilities $(j = 1, \ldots, m)$.
k	number of existing facilities that are part of one's own chain (the first k of the m facilities are assumed in this category, $0 < k < m$).
d_{ij}	distance between demand point p_i and facility f_j.
d_{ix}	distance between demand point p_i and the new facility x.
α_{ij}	quality of facility f_j as perceived by demand point p_i.
$g_i(\cdot)$	a non-negative non-decreasing function.
$\frac{\alpha_{ij}}{g_i(d_{ij})}$	attraction that demand point p_i feels for facility f_j.
γ_i	weight for the quality of x as perceived by demand point p_i.
$\frac{\gamma_i \alpha}{g_i(d_{ix})}$	attraction that demand point p_i feels for the new facility x.

These particular attraction functions generalize the proposals in [7, 26, 28, 36]. We may assume that $g_i(d_{ij}) > 0 \ \forall i, j$. Some particular cases already proposed in the literature for the g_i functions are $g_i(d_{ix}) = e^{\lambda_i d_{ix}}$ (see [25]) or $g_i(d_{ix}) = (d_{ix})^{\lambda_i}$ (see [6]), with $\lambda_i > 0$ a given parameter.

2.1 First Objective: Maximization of the Market Share Captured by the Franchisor

From the previous assumptions, the total market share attracted by the franchise is

$$M(x,\alpha) = \sum_{i=1}^{n} w_i \frac{\dfrac{\gamma_i \alpha}{g_i(d_{ix})} + \displaystyle\sum_{j=1}^{k} \dfrac{\alpha_{ij}}{g_i(d_{ij})}}{\dfrac{\gamma_i \alpha}{g_i(d_{ix})} + \displaystyle\sum_{j=1}^{m} \dfrac{\alpha_{ij}}{g_i(d_{ij})}}.$$

We assume that the operating costs for the franchisor due to the new facility are fixed, that is, they are independent from the final location and quality chosen. This is usually the case, since the operating costs for the franchisor are mainly due to the advertising of the franchise's trademark. In the same way, we also disregard the operating costs of the existing facilities that are part of the franchise.

In this way, the profit obtained by the franchisor is an increasing function of the market share captured by the franchise. Thus, maximizing the profit obtained by the franchisor is equivalent to maximizing the market share captured by the franchise. This will be the first objective of our problem.

2.2 Second Objective: Maximization of the Profit Obtained by the Franchisee

The second objective of our problem is the maximization of the profit obtained by the franchisee, to be understood as the difference between the revenues obtained from the market share captured by the new facility minus its operational costs (see [16]). The market share captured by the new facility (franchisee) is given by

$$m(x,\alpha) = \sum_{i=1}^{n} w_i \frac{\dfrac{\gamma_i \alpha}{g_i(d_{ix})}}{\dfrac{\gamma_i \alpha}{g_i(d_{ix})} + \displaystyle\sum_{j=1}^{m} \dfrac{\alpha_{ij}}{g_i(d_{ij})}}$$

and the profit is given by the following expression,

$$\pi(x,\alpha) = F(m(x,\alpha)) - G(x,\alpha)$$

where $F(\cdot)$ is a strictly increasing function which determines the expected sales (i.e., income generated) for a given market share $m(x,\alpha)$ and $G(x,\alpha)$ is a function which gives the operating cost of a facility located at x with quality α.

The function F will sometimes be linear (in problems without economies of scale), $F(m(x,\alpha)) = c \cdot m(x,\alpha)$, where c is the income per unit of good

sold. Of course, other functions can be more suitable depending on the real problem considered.

As for the function $G(x, \alpha)$, it should increase as x approaches to one of the demand points, since it is rather likely that around those locations the operational cost of the facility will be higher (due to the value of land and premises, which will make the cost of buying or renting the location higher). On the other hand, G should be a nondecreasing and convex function in the variable α, since the more quality we require of the facility, the higher the costs will be, at an increasing rate. In our computational studies we have considered G to be separable, of the form $G(x, \alpha) = G_1(x) + G_2(\alpha)$, where $G_1(x) = \sum_{i=1}^{n} \Phi_i(d_{ix})$, with $\Phi_i(d_{ix}) = w_i/((d_{ix})^{\phi_{i0}} + \phi_{i1})$, $\phi_{i0}, \phi_{i1} > 0$ and $G_2(\alpha) = e^{\frac{\alpha}{\alpha_0}+\alpha_1} - e^{\alpha_1}$, with $\alpha_0 > 0$ and α_1 given values (other possible expressions for $G(x, \alpha)$ can be found in [16]).

2.3 The Problem

The problem to be solved is then

$$\begin{cases} \max M(x, \alpha) \\ \max \pi(x, \alpha) \\ \text{s.t.} \quad d_{ix} \geq d_i^{\min} \ \forall i \\ \qquad \alpha \in [\alpha_{\min}, \alpha_{\max}] \\ \qquad x \in R \subset \mathbb{R}^2 \end{cases} \tag{2}$$

where the parameters $d_i^{\min} > 0$ and $\alpha_{\min} > 0$ are given thresholds, which guarantee that the new facility is not located over a demand point and that it has a minimum level of quality, respectively. The parameter $\alpha_{\max}$ is the maximum value that the quality of a facility may take in practice. By R we denote the region of the plane where the new facility can be located. To carry out the interval methods proposed in the next two sections we just need to be able to write R through a set of analytical constraints.

To clarify the biobjective nature of (2), consider Figure 1. Dotted circles with numbers 1 to 5 denote the forbidden regions around the existing demand points (supposed to be at the center of the forbidden regions, and all with demand 1), the cross $\times$ denotes the location of an existing facility owned by the chain and the solid circle point $\bullet$ the location of a competitor's facility. The franchisor would like the new facility to be located close to demand point 5 (he/she already captures the market of demand points 1 to 4, and in this way he/she can win a part of the market of demand point 5), whereas the franchisee would like the facility to be located close to the existing chain-owned facility (in this way, he/she can capture nearly half of the market of demand points 1 to 4, which is much more than he/she can get by locating close to demand point 5). In a grey scale we can see the efficient set for this problem (the lighter the better for the franchisor, and darker the better for the franchisee).

Fig. 1. Conflicting objectives.

In order to have problem (2) written in the form of problem (1), in what follows we will use the following notation: $y = (x, \alpha)$, $f_1(y) = -M(x, \alpha)$, $f_2(y) = -\pi(x, \alpha)$ and S will denote the feasible set of problem (2).

3 An Interval Branch-and-Bound Method for Obtaining the Whole Efficient Set

Problem (2) is very hard to solve: its objective functions are neither concave nor convex, thus, the optimization of one of them alone leads to a global optimization problem. Furthermore, we are interested in obtaining its whole efficient set. To cope with it, we need to use global optimization techniques. Among these, only a branch-and-bound scheme seems to be appropriate for our purposes, although the computation of bounds is a difficult task, too. In this paper we have used such a method, which makes use of the *Interval Analysis* (see the books [23, 30, 39], which are excellent introductions to the topic).

To our knowledge, and with only two exceptions [27, 47], all the publications on interval methods for optimization deal with single objective problems. In [27] it is simply proposed to use the classical interval B&B global optimization methods for solving the single objective problems to which the multiobjective problem is reduced when using the weighting method or the minimax method. A more sophisticated use of interval techniques is the modification of the classical lexicographic method proposed in [47] and shortly described (and slightly modified) in Section 4.

The purpose of this section is to present a new interval B&B method able to obtain a superset of the complete efficient set with values within a given precision. The method deals with the multiple objectives directly, that is, it does not convert the problem into a single objective optimization problem or a family of such kind of problems, as most of the multiobjective optimization methods do. We briefly summarize the fundamental concepts of interval analysis which are needed for this paper. For more details, the interested reader

is referred to [23, 30, 39]. Other applications of interval analysis to location problems can be found in [12, 14, 15, 16, 34, 47].

3.1 Interval Analysis

Following the notation suggested by Kearfott *et al.* [31] as standard, boldface will denote intervals, lower case will be used for scalar quantities or vectors (vectors are then distinguished from components by use of subscripts), and upper case for matrices. Brackets '$[\cdot]$' will delimit intervals, while parentheses '$(\cdot)$' are used for vectors and matrices. Underlines will denote lower bounds of intervals and overlines indicate upper bounds of intervals. For example, we may have the interval vector $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, where $\boldsymbol{x}_i = [\underline{x_i}, \overline{x_i}]$. The *width* of an interval $\boldsymbol{x}_i$ is denoted by $\mathrm{wid}(\boldsymbol{x}_i) = \overline{x_i} - \underline{x_i}$ and its relative width by $\mathrm{wid}_{\mathrm{relat}}(\boldsymbol{x}_i) = \mathrm{wid}(\boldsymbol{x}_i)/\min\{|x_i| : x_i \in \boldsymbol{x}_i\}$ if $0 \notin \boldsymbol{x}_i$ and $\mathrm{wid}(\boldsymbol{x}_i)$ otherwise. The width of an interval vector $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is to be understood as $\mathrm{wid}(\boldsymbol{x}) = \max\{\mathrm{wid}(\boldsymbol{x}_i) : i = 1, \ldots, n\}$. The set of intervals will be denoted by $\mathbb{IR}$, and the set of n-dimensional interval vectors, also called *boxes*, by $\mathbb{IR}^n$.

The *interval arithmetic operations* are defined by

$$\boldsymbol{x} * \boldsymbol{y} = \{x * y : x \in \boldsymbol{x}, y \in \boldsymbol{y}\} \text{ for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{IR}, \tag{3}$$

where the symbol $*$ stands for $+, -, \cdot$ and $/$, and where $\boldsymbol{x}/\boldsymbol{y}$ is only defined if $0 \notin \boldsymbol{y}$. Definition (3) is equivalent to simple constructive rules (see [23, 30, 39]). The algebraic properties of (3) are different from those of real arithmetic operations, but the main properties from the operational point of view still hold, as for instance the *inclusion isotonicity*,

$$\boldsymbol{x} \subseteq \boldsymbol{y}, \boldsymbol{z} \subseteq \boldsymbol{t} \Longrightarrow \boldsymbol{x} * \boldsymbol{z} \subseteq \boldsymbol{y} * \boldsymbol{t} \text{ (if } \boldsymbol{y} * \boldsymbol{t} \text{ is defined) for } \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{t} \in \mathbb{IR}.$$

Inclusion isotonicity is implicitly used in the construction of *inclusion functions*, which are the main interval arithmetic tool applied to optimization methods.

Definition 3. *A function $\boldsymbol{f} : \mathbb{IR}^n \to \mathbb{IR}$ is said to be an* inclusion function *of $f : \mathbb{R}^n \to \mathbb{R}$ provided $\{f(y) : y \in \boldsymbol{y}\} \subseteq \boldsymbol{f}(\boldsymbol{y})$ for all boxes $\boldsymbol{y} \subset \mathbb{IR}^n$ within the domain of f.*

Observe that if $\boldsymbol{f}$ is an inclusion function for f then we can directly obtain lower bounds and upper bounds of f over any box $\boldsymbol{y}$ within the domain of f just by taking $\underline{\boldsymbol{f}}(\boldsymbol{y})$ and $\overline{\boldsymbol{f}}(\boldsymbol{y})$, respectively.

For a function h predeclared in some programming language (like sin, exp, etc.), it is not too difficult to obtain a *predeclared* inclusion function $\boldsymbol{h}$. For a general function $f(y)$, $y \in \mathbb{R}^n$, several methods can be employed to obtain inclusion functions. The easiest method to obtain an inclusion function is the *natural interval extension*, which is obtained by replacing each occurrence of the variable y with a box, $\boldsymbol{y}$, including it; each occurrence of a predeclared function h by its corresponding inclusion function $\boldsymbol{h}$; and the real arithmetic operators by the corresponding interval operators.

3.2 The Prototype Method

The prototype interval B&B algorithm for solving (1) is described in pseudocode form in Algorithm 1. In that algorithm, $\boldsymbol{y}_0$ denotes an initial box containing the feasible set S, $\mathcal{L}_{\mathcal{W}}$ is the working list and $\mathcal{L}_{\mathcal{S}}$ the solution list.

Algorithm 1 The prototype interval B&B algorithm

```
L_W ← y_0, L_S ← ∅
while ( L_W ≠ ∅ ) do
  Select an interval y from L_W                          Selection Rule
  Divide y into subintervals y_1, ..., y_s               Division Rule
  if (y_i cannot be discarded)                           Discarding Tests
    if (y_i satisfies the termination criterion)         Termination Rule
  Store y_i in L_S
    else
  Store y_i in L_W
return L_S
```

At the end of the algorithm, the boxes in the solution list $\mathcal{L}_{\mathcal{S}}$ contain the whole efficient set S_E. Different methods can be derived depending on the actual selection, division and termination rules and discarding tests employed. We describe below the details of each of these rules we used.

3.3 Selection Rules

Several rules for the selection of the next box to be processed can be used. In particular, we have worked with the following ones:

Rule 1: Select the box $\boldsymbol{y} \in \mathcal{L}_{\mathcal{W}}$ with minimum lower bound $\underline{\boldsymbol{f}}_1(\boldsymbol{y})$. In the criterion space, this implies that Z_N will be generated from left-top to right-bottom.

Rule 2: Select the box $\boldsymbol{y} \in \mathcal{L}_{\mathcal{W}}$ with minimum lower bound $\underline{\boldsymbol{f}}_2(\boldsymbol{y})$. In the criterion space, this implies that Z_N will be generated from right-bottom to left-top.

Rule 3: We first rescale the objective functions so that their objective values are of approximately the same magnitude. Ideally, the normalization process is done with the help of the ideal and nadir objective vectors.

Definition 4. *The* ideal objective vector *of problem (1) is a vector* $z^* \in \mathbb{R}^2$ *whose components* z_i^* *are obtained by minimizing each of the objective functions individually subject to the original constraints of problem (1).*

Definition 5. *The* nadir objective vector *of problem (1) is a vector* $z^{nad} \in \mathbb{R}^2$ *giving the upper bounds of* Z_N.

Once the ideal and nadir objective vectors are known, we replace $f_i(y)$ by

$$\frac{f_i(y) - z_i^*}{z_i^{nad} - z_i^*},$$

whose range is within $[0,1]$. However, since in practice those vectors are not known in advance, we propose the following normalization:

$$n_i(y) = \frac{f_i(y) - \tilde{z}_i^*}{\tilde{z}_i^{nad} - \tilde{z}_i^*},$$

where

$$\tilde{z}_i^* = \min\{\underline{f}_i(\boldsymbol{y}) : \boldsymbol{y} \in \mathcal{L}_{\mathcal{W}} \cup \mathcal{L}_{\mathcal{S}}\},$$
$$\tilde{z}_i^{nad} = \max\{\underline{f}_i(\boldsymbol{y}) : \boldsymbol{y} \in \mathcal{L}_{\mathcal{W}} \cup \mathcal{L}_{\mathcal{S}}\}.$$

The new selection rule that we propose is to select the box $\boldsymbol{y} \in \mathcal{L}_{\mathcal{W}}$ with minimum lower bound $\lambda \underline{\boldsymbol{n}}_1(\boldsymbol{y}) + (1-\lambda)\underline{\boldsymbol{n}}_2(\boldsymbol{y})$, with $\lambda \in [0,1]$ a given value. Depending on the actual value of λ we explore first a given area of Z_N. Notice that this rule generalizes the previous ones (we can obtain the first rule by setting $\lambda = 1$ and the second one when $\lambda = 0$, see Figure 2).

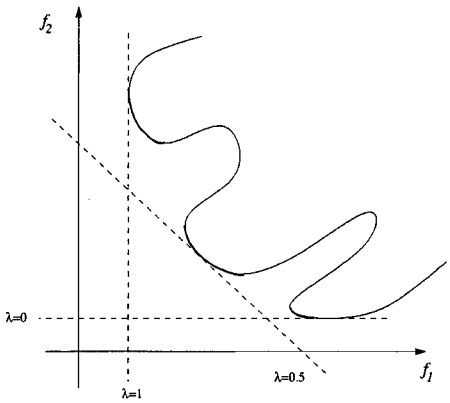


Fig. 2. Selection rules.

Rule 4: We may alternate between different λ-values in the previous rule from iteration to iteration. In this way, all the parts of the criterion space are generated in a more or less uniform way. In particular, to consider p different values, at iteration k we may set $\lambda = \frac{(k-1) \bmod p}{p-1}$.

3.4 Division Rules

The most widely used division rule in interval B&B methods is the bisection of the box perpendicular to a direction of maximum width. We have used this same rule here.

3.5 Discarding Tests

The tests for verifying that a box contains no efficient point form the most important part of Algorithm 1. We next briefly discuss the ones we have used here.

Feasibility Test: Let us suppose that S is given by $S = \{y : h_l(y) \leq 0, l = 1, \ldots, r\}$. We say that a box $\boldsymbol{y}$ *certainly satisfies* the constraint $h_l(y) \leq 0$ if $\overline{\boldsymbol{h}}_l(\boldsymbol{y}) \leq 0$ and that $\boldsymbol{y}$ *does certainly not satisfy* it if $\underline{\boldsymbol{h}}_l(\boldsymbol{y}) > 0$. A box $\boldsymbol{y} \subseteq \boldsymbol{y}_0$ is said *certainly feasible* if it certainly satisfies all the constraints, *certainly infeasible* if it does certainly not satisfy at least one of the constraints, and *undetermined* otherwise. A box $\boldsymbol{y} \subseteq \boldsymbol{y}_0$ is said *certainly strictly feasible* if $\overline{\boldsymbol{h}}_l(\boldsymbol{y}) < 0, l = 1, \ldots, r$. The 'feasibility test' [39] discards boxes that are certainly infeasible. It also provides information about the feasibility of a box. Notice that to apply this test we just need an inclusion function $\boldsymbol{h}_l$ for each of the functions h_l defining the constraints. To avoid unnecessary feasibility checks, we store with each box $\boldsymbol{y}$ a Boolean vector stating whether the box certainly (strictly) satisfies a constraint or not; when the box is split, the subboxes inherit this vector and in this way it is only necessary to do the checking for the constraints not certainly (strictly) satisfied by $\boldsymbol{y}$.

Multi-objective cut-off test: Every time a box $\boldsymbol{y}$ is chosen from the list $\mathcal{L}_{\mathcal{W}}$, and provided that its midpoint c (as a point interval) is certainly feasible, we compute $\overline{\boldsymbol{f}}(c) = (\overline{\boldsymbol{f}}_1(c), \overline{\boldsymbol{f}}_2(c))$ and update (if possible, i.e., if $\overline{\boldsymbol{f}}(c)$ is not dominated by any point in $\mathcal{L}_{PES}$) the list $\mathcal{L}_{PES}$ of 'provisional' efficient solution points available so far. The multiobjective cut-off test is the natural generalization to the multiobjective case of the classical 'cut-off test' (also called 'midpoint test' [39]), and discards boxes whose points are not efficient, i.e., a box $\boldsymbol{y}$ is removed if $\underline{\boldsymbol{f}}(\boldsymbol{y}) > z^*$ for some $z^* \in \mathcal{L}_{PES}$. Also, a new box $\boldsymbol{y}$ must only be entered into $\mathcal{L}_{\mathcal{W}}$ if $\underline{\boldsymbol{f}}(\boldsymbol{y}) \not> z^*$ holds for all $z^* \in \mathcal{L}_{PES}$. A similar idea has already been suggested in [3, 24, 44] for other non-interval B&B methods.

3.6 Termination Rules

As termination rule, we sent a box $\boldsymbol{y}$ to $\mathcal{L}_S$ whenever

$$\left(\mathrm{wid}_{\mathrm{relat}}\boldsymbol{f}_1(\boldsymbol{y}) < \epsilon_1 \text{ and } \mathrm{wid}_{\mathrm{relat}}\boldsymbol{f}_2(\boldsymbol{y}) < \epsilon_2\right) \quad \text{or} \quad \mathrm{wid}(\boldsymbol{y}) < \epsilon_3.$$

4 An Interval Lexicographical-Like Method

The lexicographic method (see [18]) is used to obtain efficient solutions when the decision-maker has clear preferences on the objectives and can arrange the objective functions according to their absolute importance. In that case, the decision-maker does not want the whole efficient set, but just a small part of it which reflects his/her preferences.

Although the classical method has good properties (any lexicographic solution is efficient) it presents several drawbacks. In particular, it is very likely that the less important objective is not taken into consideration at all and it does not allow a small increment of the first objective to be traded off with a decrement of the second objective. Recently, in [47], a modification of that method has been proposed which solves those drawbacks. The new method is called δ-lexicographic method, and works as follows.

δ-lexicographic method

1. The decision maker arranges the objective functions according to their absolute importance, f_1, f_2.
2. The region R_δ^1 of δ-optimality of problem

$$(P_1) \quad \min f_1(y) \text{ s.t. } y \in S$$

is calculated, where

$$R_\delta^1 = \{y \in S : f_1(y) - f_1^* \leq \delta f_1^*\},$$

$\delta > 0$ is a given value and f_1^* denotes the optimal value of (P_1).

3. The second objective function is optimized subject to the original constraints and to a new one to guarantee the δ-optimality of f_1, that is, we solve the problem

$$(P_2^\delta) \quad \min f_2(y) \text{ s.t. } y \in R_\delta^1$$

Any solution to this second problem is a *δ-lexicographic solution*, to be presented to the decision-maker as a solution of (1).

As we can see, instead of solving problem (P_1) till optimality, as the classical lexicographic method does, now its region of δ-optimality is obtained. In this way, a relative increment of $\delta\%$ in the first objective is allowed to be traded off with a decrement of the second objective. The theoretical properties of the new method are rather appealing. In particular, any efficient solution can be obtained by the δ-lexicographic method (see [47]).

On the other hand, from the locational point of view, for a decision-maker it is better to have a small region within which to choose the final location for the facility (taking other aspects not included in the formulation of the problem into account) than just a single efficient point. In order to be able to offer to the decision-maker such a small region, we propose that instead of solving (P_2^δ) till optimality, we obtain its region of θ-optimality,

$$R_{\theta(\delta)}^2 = \{y \in R_\delta^1 : f_2(y) - f_2^*(\delta) \leq \theta f_2^*(\delta)\},$$

where $f_2^*(\delta)$ denotes the optimal value of (P_2^δ) and $\theta > 0$ is a given value.

Definition 6. *A feasible vector $y^* \in S$ is said to be* (δ, θ)-efficient *iff there exists an efficient vector $\check{y}$ such that $f_1(y^*) - f_1(\check{y}) \leq \delta f_1(\check{y})$ and $f_2(y^*) - f_2(\check{y}) \leq \theta f_2(\check{y})$.*

Notice that the points in $R^2_{\theta(\delta)}$ are not necessarily efficient, but they are (δ, θ)-efficient. For any fixed δ, as used in the δ-lexicographic method, closeness of a (δ, θ)-efficient point to an efficient point is determined by the parameter θ alone. Thus, the smaller the θ, the better.

In [47] the authors present an interval B&B method to carry out the δ-lexicographic method. The same method can be used to obtain $R^2_{\theta(\delta)}$. The only changes to be done are:

1. In the 'cut-off test due to the second objective function', now we can discard a box $\boldsymbol{y}$ provided that $\underline{f}_2(\boldsymbol{y}) > \tilde{f}_2(1+\theta)$ (where $\tilde{f}_2$ is the best upper bound on the optimal value of (P_2^δ) known by the algorithm), and
2. In the second call of the main procedure, we have to pass the parameter θ instead of 0, as the size for the region of θ-optimality for problem (P_2^δ).

5 Computational Experiments

All the computational results presented in this paper have been obtained on a PC with an Intel Pentium IV 2.33GHz processor and with 1 Gbyte RAM running under the Linux operating system. For the implementation we have used the interval arithmetic in the PROBIL/BIAS library [32] and the automatic differentiation of the C++ toolbox library described in [22].

5.1 Usefulness of the Selection Rules

The first part of the computational experiments studies the efficiency of the selection rules presented in Subsection 3.3. To do it, we have generated different types of problems, varying the number of demand points (n =20, 40 or 60), the number of existing facilities (m =2, 3, 4 or 5) and the number among these belonging to the chain ($k = 1, \ldots, m-1$). For every type of setting 5 problems were generated, by randomly choosing the parameters of the problems uniformly within the following intervals:

- $f_j, p_i \in [0, 10]^2$,
- $\omega_i \in [1, 10]$,
- $\gamma_i \in [0.75, 1.25]$,
- $\alpha_{ij} \in [0.5, 5]$,
- $\phi_{i0} = \phi_0 = 2, \phi_{i1} \in [0.5, 2]$, the parameters for $\Phi_i(d_{ix}) = w_i \frac{1}{(d_{ix})^{\phi_{i0}+\phi_{i1}}}$, and $G_1(x) = \sum_{i=1}^n \Phi_i(d_{ix})$,
- $\alpha_0 \in [7, 9], \alpha_1 \in [4, 4.5]$, the parameters for $G_2(\alpha) = e^{\frac{\alpha}{\alpha_0}+\alpha_1} - e^{\alpha_1}$,
- $c \in [1, 2]$, the parameter for $F(m(x, \alpha)) = c \cdot m(x, \alpha)$,
- $b_1, b_2 \in [1, 2]$, parameters for $d_{ix} = \sqrt{b_1(x_1 - p_{i1})^2 + b_2(x_2 - p_{i2})^2}$ (see [13]).

The searching space for every problem was

$$x \in [0, 10]^2, \quad \alpha \in [0.5, 5].$$

The tolerances used for the interval B&B method were $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0.05$.

Table 1. Efficiency of the selection rules

	CPU Time				Max list length			
(n,m,k)	R1	R2	R3	R4	R1	R2	R3	R4
20,2,1	1053.2	90.7%	89.3%	143.1%	4964	199.5%	318.7%	388.7%
20,3,1	803.0	97.9%	110.4%	96.0%	6531	177.3%	257.3%	228.6%
20,3,2	1486.0	109.4%	113.4%	102.7%	7572	195.2%	293.5%	242.8%
20,4,1	1248.2	103.7%	111.0%	101.7%	10955	151.9%	219.6%	203.7%
20,4,2	849.6	113.4%	128.0%	117.4%	11113	212.8%	239.6%	243.0%
20,4,3	749.0	112.6%	132.6%	120.5%	8439	203.1%	266.7%	260.2%
Average	1031.5	104.5%	112.8%	112.7%	8262	189.0%	258.3%	249.9%
40,2,1	301.5	107.3%	123.9%	106.6%	4124	211.4%	317.6%	254.9%
40,3,1	492.9	107.8%	123.1%	104.3%	5320	210.6%	288.7%	240.3%
40,3,2	489.9	97.7%	123.1%	102.7%	4999	220.8%	315.9%	264.2%
40,4,1	280.3	90.1%	122.1%	93.4%	10411	151.5%	182.2%	161.1%
40,4,2	524.3	100.0%	119.7%	104.7%	9347	172.8%	222.6%	194.0%
40,4,3	1287.0	109.1%	131.5%	114.6%	10341	229.9%	309.5%	282.0%
40,5,1	981.9	89.2%	112.9%	93.1%	20612	134.8%	172.5%	144.7%
40,5,2	703.1	93.4%	110.0%	96.0%	9836	154.5%	191.7%	182.8%
40,5,3	321.1	94.2%	117.7%	93.7%	9197	182.1%	205.8%	187.2%
40,5,4	691.5	141.4%	166.5%	168.6%	10042	306.4%	313.7%	372.9%
Average	607.3	104.2%	126.1%	110.0%	9423	188.0%	234.4%	215.5%
60,2,1	104.8	99.0%	114.0%	105.0%	3424	135.5%	163.8%	164.3%
60,3,1	339.5	108.3%	120.1%	103.4%	9183	143.5%	174.1%	149.8%
60,3,2	223.3	103.2%	135.7%	102.9%	4592	196.8%	272.0%	216.9%
60,4,1	190.1	87.4%	106.9%	88.3%	9503	117.6%	142.1%	128.1%
60,4,2	833.8	107.1%	125.2%	109.9%	7412	184.8%	250.4%	224.4%
60,4,3	888.0	111.1%	138.9%	111.0%	4697	246.2%	401.8%	311.7%
60,5,1	660.7	85.3%	119.1%	92.6%	17794	130.2%	179.2%	143.3%
60,5,2	304.0	80.2%	108.6%	87.8%	14020	132.2%	148.3%	140.6%
60,5,3	360.7	88.9%	111.9%	94.1%	14112	137.1%	159.0%	151.9%
60,5,4	610.7	123.3%	133.5%	129.3%	7695	282.9%	292.1%	314.2%
Average	451.6	102.5%	125.1%	105.6%	9243	158.1%	197.6%	177.0%
Glob.Av	645.3	103.9%	120.9%	109.8%	9086	176.5%	225.0%	207.6%

All the problems were solved four times by the interval B&B method: using each selection rule in turn. The results are shown in Table 1. The columns 'Ri' refer to the results obtained when using selection rule 'i'. In rule 3, we have set $\lambda = 0.5$. In the fourth selection rule, we have used $p = 10$ different values of λ. The values in the table are averages over the five problem instances in

each setting (n, m, k). We give first results about CPU time. In column R1 we give CPU times in seconds observed when using R1, while for the other selection rules we give their relative times as compared to R1. Next, we give the maximum number of boxes stored in memory at any time by the algorithm, following the same structure as for the CPU time.

CPU times obtained by rules R1 and R2 are very similar, and they both are slightly better than for rule R4. R3 is significatively the worst rule. As for the maximum number of boxes stored in memory at any time by the algorithm, we see clearly that R1 is the best rule: it uses half the space of R4, and the difference is even larger with R3, which is the worst rule. R2 also needs 176.5% more space than R1. The explanation for this last fact seems to be that the width of the range of f_2 at the solution boxes is usually greater than the corresponding range of f_1 (see Table 2), thus the multi-objective cut-off test will more easily discard boxes when using R1 than when using R2 (or any other rule).

This clearly shows that rule R1 is the best for handling our competitive location problems. In the rest of the experiments in this paper, we have always used rule R1.

5.2 Branch and Bound vs. Lexicographic Method

Let us now try to compare the information obtained by the interval branch-and-bound method introduced in Section 3 and the modification of the δ-lexicographic method described in Section 4. Observe first that these methods are *not* directly comparable, since they do very different things: whereas the former obtains the 'whole' efficient set, the latter just gives a 'small part' of it. That is why we highlight the different kind of information that they provide.

Figure 3 shows the solution boxes (projected in the location space) and their image in criterion space obtained by each algorithm in one of the instances. As we can see, the nondominated set in criterion space can have a disconnected shape. Accordingly, the efficient set may also be disconnected. This clearly illustrates the difficulty of finding the complete efficient set to this kind of location problems. Notice also that (most of) the region $R^2_{\theta(\delta)}$ offered by the modified δ-lexicographic method is included in the solution set offered by the interval B&B method (which, we recall, contains the complete efficient set); in fact, in all our instances, it was a proper subset of it. This confirms that by choosing an appropriate θ the points in $R^2_{\theta(\delta)}$ are either efficient or lie very near efficient points.

In our computational study we have used the same problems as in previous subsection. The tolerances used for the interval B&B method were $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0.05$, and for the modified δ-lexicographic method $\delta = 0.1$ and $\theta = 0.01$. The results are summarized in Table 2. As before, the given values are averages over the five problems in each setting (n, m, k). For each algorithm ('Lexic' refers the the modified δ-lexicographic method and 'B&B' to the interval B&B method) we give the CPU time (in seconds), the area covered by the

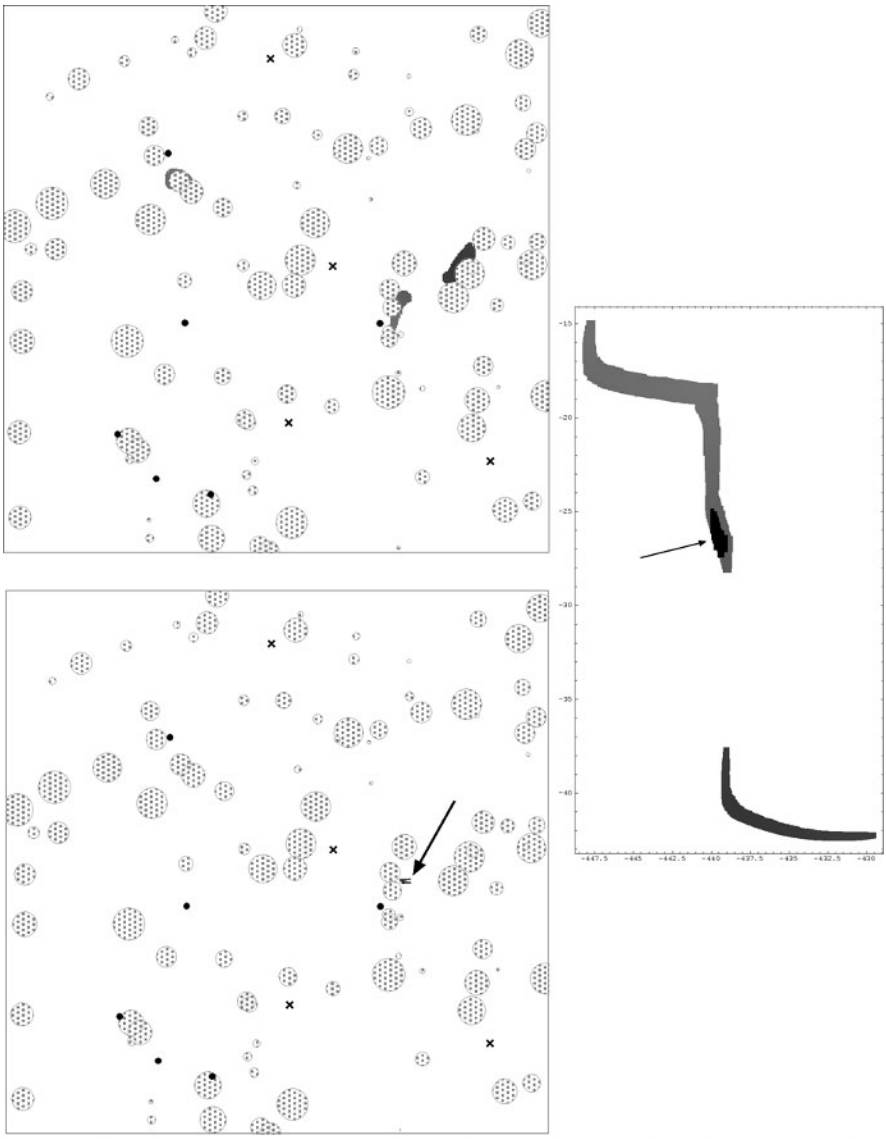


Fig. 3. Solution of an instance by the methods. The dotted circles correspond to the forbidden areas surrounding the demand points; the crosses × and the solid points • represent the locations of the existing facilities. The top-left figure is the projection on the location space of the efficient set (in a grey scale) obtained by the interval B&B; the bottom-left figure is the projection on the location space of the solution boxes (in black) offered by the modified δ-lexicographic method (they are pointed with an arrow); in the right picture, we have the corresponding images in the solution space.

Table 2. Study of the information provided by the modified δ-lexicographic method and the interval B&B method.

	CPU time		size of solution boxes				width of range at $\cup_{y \in \mathcal{L}_S} y$			
			x-space		α-space		f_1		f_2	
(n,m,k)	Lexic	B&B	Lexic	B&B	Lexic	B&B	Lexic	B&B	Lexic	B&B
20,2,1	5.18	1053.2	0.18	1.20	0.16	3.59	1.00	17.99	0.88	39.89
20,3,1	6.36	803.0	0.04	0.71	0.15	2.86	0.94	18.62	1.10	21.50
20,3,2	4.90	1486.0	0.06	1.33	0.15	3.48	0.62	13.79	1.41	35.88
20,4,1	8.05	1248.2	0.09	0.92	0.15	3.64	1.45	20.71	2.60	28.07
20,4,2	5.14	849.6	0.03	0.65	0.06	3.70	0.71	21.37	1.26	28.33
20,4,3	6.49	749.0	0.10	0.85	0.13	3.48	0.72	10.54	1.64	27.18
Average	6.02	1031.5	0.08	0.94	0.13	3.46	0.91	17.17	1.48	30.14
40,2,1	21.48	301.5	0.77	1.13	0.75	1.07	4.88	12.51	1.77	14.05
40,3,1	20.13	492.9	0.73	1.18	0.66	1.24	6.84	19.79	2.91	13.87
40,3,2	18.81	489.9	0.32	1.83	0.41	1.20	1.57	13.66	1.18	27.88
40,4,1	34.09	280.3	0.59	0.98	0.52	0.87	6.16	16.93	2.12	9.17
40,4,2	20.45	524.3	0.16	1.02	0.30	2.03	1.43	16.19	1.15	18.92
40,4,3	14.43	1287.0	0.10	1.86	0.35	1.78	1.16	12.15	1.45	26.86
40,5,1	30.47	981.9	0.03	1.01	0.27	2.38	4.06	30.52	1.87	18.78
40,5,2	26.34	703.1	0.29	1.06	0.25	1.78	2.66	23.09	1.26	20.58
40,5,3	22.04	321.1	0.17	1.38	0.65	0.77	4.66	17.78	1.39	12.23
40,5,4	14.22	691.5	0.03	2.32	0.26	1.45	0.81	18.87	1.14	33.74
Average	22.25	607.3	0.32	1.38	0.44	1.46	3.42	18.15	1.63	19.61
60,2,1	30.11	104.8	1.15	1.15	0.61	0.40	10.80	10.71	5.90	18.57
60,3,1	44.81	339.5	0.73	0.71	0.33	0.41	8.20	9.35	3.27	7.21
60,3,2	20.31	223.3	0.08	1.59	0.15	0.38	1.46	20.87	1.92	22.43
60,4,1	57.84	190.1	1.00	0.84	0.45	0.46	9.82	10.52	2.93	8.06
60,4,2	31.75	833.8	0.24	1.82	0.43	1.11	4.74	24.67	1.82	22.31
60,4,3	20.72	888.0	0.20	1.68	0.21	1.17	1.39	12.83	2.01	37.31
60,5,1	51.69	660.7	0.27	0.99	0.56	1.50	8.09	25.98	1.90	14.20
60,5,2	48.85	304.0	0.38	0.60	0.61	1.20	8.05	19.15	2.32	12.81
60,5,3	45.36	360.7	0.19	1.19	0.69	0.88	5.67	15.72	2.07	18.81
60,5,4	25.78	610.7	0.06	2.36	0.26	0.85	1.92	11.21	1.57	37.28
Average	37.72	451.6	0.43	1.29	0.43	0.84	6.01	16.10	2.57	19.90
Glob.Av	24.45	645.3	0.31	1.24	0.37	1.68	3.84	17.13	1.96	22.15

solution boxes when they are projected to the location space (i.e., the area of $\cup_{\boldsymbol{y} \in \mathcal{L}_S} proj_x(\boldsymbol{y})$), the width of the interval containing the possible values of α (obtained by projecting the solution boxes to the α-space, i.e., the width of $\cup_{\boldsymbol{y} \in \mathcal{L}_S} proj_\alpha(\boldsymbol{y})$), and the width of the range of f_1 and f_2 at the solution set offered by each algorithm.

As was to be expected, Lexic is much faster than B&B. However, notice that CPU-time increases with n for Lexic, while the converse holds for B&B. For Lexic, and for a fixed pair (n,m), the CPU time decreases as k increases; this is so, because usually the region of δ-optimality R^1_δ becomes smaller as k

increases, as we can see from the columns giving the size of the solution boxes and the width of the range. Notice also that for Lexic, for a fixed n, the CPU times for the problems with $k = m - 1$ are very similar.

As for the area covered by the solution boxes, notice that it increases with n both in x-space and in α-space for Lexic. However, the area in x-space is more or less constant for B&B, whereas in α-space it decreases as n increases. Notice also that with B&B, for fixed (n, m), the x-space area usually increases with k. In all cases, we can see that the area covered by the solution boxes, both in x-space and α-space, is much smaller (from 3 to 10 times) for Lexic than for B&B, which clearly shows that Lexic only finds a small part of the efficient set.

Something similar can be said about the width of the ranges of both functions at the solution boxes offered by the algorithms. Again, the range for Lexic is from 3 to 20 times smaller than for B&B for f_1, and the difference is even higher for f_2. With Lexic the ranges increase with n; for a fixed (n, m) the ranges decrease as k increases; and for a fixed n the ranges for $k = m - 1$ are similar. With B&B, the range of f_1 is more or less constant for all the settings, whereas the range of f_2 seems to decrease; and for a fixed (n, m) the ranges of f_1 usually decrease as k increases, while the converse holds for f_2. This means that the greater the presence of the chain in the area, the more variable can be the profit obtained by the franchisee in the efficient set.

5.3 Economical Analysis of the Model

In this subsection, we try to obtain supplementary sensitivity information when modifying some of the parameters in the model or when adding some additional constraints to it. This may be viewed as an exploratory analysis of some of the economical implications of the model proposed. The aim is to get a deeper knowledge of the interactions between the different elements of the model that can adjust it to real applications. We use here some of the instances of the previous subsection.

First, we have studied the variation in the range of one of the objectives in a region of δ-optimality of the other objective for four different values of δ. The results are given in Table 3. Each pair of columns refers to one of the instances. In the columns called $f_2(R_\delta^1)$ we give the exact range of f_2 at R_δ^1. The lower bound of this interval is just the solution of (P_2^δ) provided by the δ-lexicographic method; the upper bound has been estimated by solving the problem

$$\max f_2(x) \text{ s.t. } x \in R_\delta^1$$

To solve that problem we need to obtain the whole region of δ-optimality; for this, we have used the main procedure of the δ-lexicographic method without the 'cut-off test due to the second objective function' [47]. Analogously, $f_1(R_\delta^2)$ gives the range of f_1 at the region of δ-optimality R_δ^2 of problem

$$\min f_2(x) \text{ s.t. } x \in S$$

Table 3. Study of the variation in the range of one of the objectives in a region of δ-optimality of the other objective.

	(20,4,3)		(40,5,4)		(60,5,4)	
δ	$f_2(R^1_\delta)$	$f_1(R^2_\delta)$	$f_2(R^1_\delta)$	$f_1(R^2_\delta)$	$f_2(R^1_\delta)$	$f_1(R^2_\delta)$
0.01	[13.7, 40.0]	[166.7,168.6]	[35.1, 42.2]	[358.5,361.7]	[82.6, 94.7]	[537.3,540.6]
0.05	[11.8, 43.3]	[165.4,169.7]	[31.5, 47.6]	[356.9,362.9]	[78.3,116.0]	[534.8,541.9]
0.10	[10.2, 46.7]	[164.3,171.3]	[28.6, 92.8]	[355.4,363.5]	[73.4,127.1]	[532.5,543.7]
0.15	[8.5, 52.7]	[163.6,171.7]	[25.9,101.3]	[354.3,363.7]	[66.2,137.7]	[530.6,545.4]

As should be, the ranges increase with δ: e.g. $R^1_\delta \subset R^1_{\delta'}$ when $\delta < \delta'$. However, notice that the $f_2(R^1_\delta)$ ranges increase much more than the $f_1(R^2_\delta)$ ranges. Also, the width of the $f_2(R^1_\delta)$ ranges are much bigger than the width of the $f_1(R^2_\delta)$ ranges. This indicates that there are efficient points offering similar values for the first objective but quite different for the second. Thus, in a negotiation between the franchisor and the franchisee to decide the final location, the franchisor has a much more comfortable situation.

Table 4. Sensitivity of the objectives to the size of the forbidden regions

	(20, 4, 3)		(40, 5, 4)		(60, 5, 4)	
I	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$
			$r = 0$			
[0%, 25%]	[32.0,43.1]	[167.4,170.6]	[44.7,61.8]	[402.0,415.0]	[86.8,117.8]	[539.5,543.8]
[25%, 50%]	[42.8,44.6]	[170.5,171.6]	[61.5,70.3]	[414.8,421.8]	[117.1,130.7]	[543.6,547.4]
[50%, 75%]	[44.3,46.8]	[171.6,171.8]	[70.0,77.9]	[421.6,427.4]	[129.9,142.4]	[547.2,549.1]
[75%, 100%]	[46.5,47.3]	[171.8,171.9]	[77.7,79.6]	[427.1,428.8]	[141.5,152.0]	[549.0,550.3]
			$r = 1$			
[0%, 25%]	[32.0,43.1]	[167.4,170.6]	[44.7,61.8]	[402.0,415.0]	[86.8,116.9]	[540.2,544.2]
[25%, 50%]	[42.8,44.6]	[170.5,171.6]	[61.5,70.3]	[414.8,421.8]	[116.1,129.2]	[544.0,547.6]
[50%, 75%]	[44.3,46.8]	[171.6,171.8]	[70.0,77.9]	[421.6,427.4]	[128.4,140.2]	[547.4,549.1]
[75%, 100%]	[46.5,47.3]	[171.8,171.9]	[77.7,79.6]	[427.1,428.8]	[139.4,149.6]	[549.0,550.3]
			$r = 2$			
[0%, 25%]	[40.2,43.7]	[167.7,169.3]	[42.2,62.1]	[401.5,415.7]	[86.3,119.1]	[538.7,547.4]
[25%, 50%]	[43.4,44.6]	[169.2,170.6]	[61.1,70.8]	[415.1,423.0]	[117.3,128.0]	[547.0,548.8]
[50%, 75%]	[44.3,46.5]	[170.4,171.1]	[69.9,78.1]	[422.2,428.2]	[126.7,130.3]	[548.6,549.3]
[75%, 100%]	[46.2,46.9]	[171.1,171.4]	[77.7,79.3]	[427.6,428.8]	[129.4,130.7]	[548.9,550.4]

In our second analysis we investigated how the addition of circular forbidden regions around the existing franchise-owned facilities affects both objectives, depending on the radii of the forbidden regions. The addition of this kind of constraints is a common technique in franchise systems to avoid that the existing facilities loose too much of their market share [47]. We have used the same instances as in the previous study. Table 4 gives a summary of the results obtained for three different radii r of the circular forbidden regions.

The columns $f_2(f_1^{-1})$ give the range for the values of f_2 on those efficient points at which f_1 takes values within $I = [a\%, b\%]$ of its efficiency range $f_1(S_E)$. The columns $f_1(f_2^{-1})$ give similar figures, but interchanging both

functions. For any given r, once we have obtained the whole efficient set of the constrained problem with the interval B&B method this information can be obtained easily for any choice of $I = [a\%, b\%]$. Thus, each pair of 4-row subcolumns of Table 4 is obtained after a single B&B run, and shows only part of the full information provided by the run. In fact using the $f_1(f_2^{-1})$ and $f_2(f_1^{-1})$ ranges given allows to reconstruct a coarser approximation of the efficient set, as shown in Figure 4.

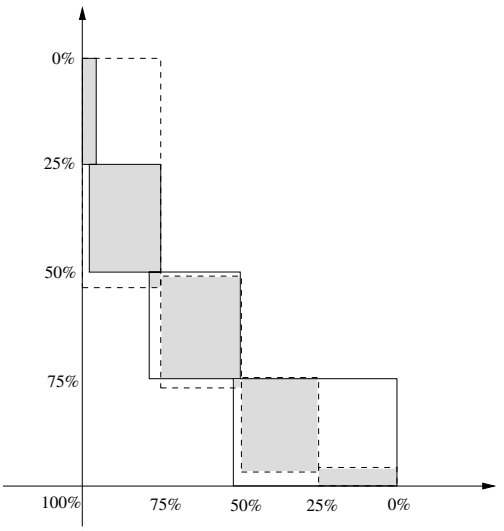


Fig. 4. Approximate efficient set for the instance $(40, 5, 4)$, when $r = 2$

As we can see from Table 4, in the chosen instances, the influence of the forbidden regions around the existing chain-owned facilities is quite small. In these instances the efficient set in the unconstrained problems lies far from the existing chain-owned facilities; thus, the addition of the constraints does not affect the solution much. When the constraints do have some influence (specially for the case $r = 2$), it mainly reduces the range of $f_1(f_2^{-1})$.

Our last experiment dealt with the sensitivity of both objectives when the franchisee has a limited budget, that is, we now have an additional constraint of the form $G(x, \alpha) \leq B$, where B is the available budget. We have again used the same instances as before and the results are presented in Table 5. Now, 'Reduction of $B = b\%$' refers to the reduction (in percentage) in the budget B as compared to the original budget needed for locating and running the facility in the unconstrained case (which corresponds to the case $r = 0$ of the previous table). The meaning of the columns is the same as in Table 4. As before this information can be obtained easily once we have obtained the whole efficient set of the constrained problem with the interval B&B method.

Table 5. Sensitivity of the objectives to the limited budget

	(20, 4, 3)		(40, 5, 4)		(60, 5, 4)	
I	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$	$f_2(f_1^{-1})$	$f_1(f_2^{-1})$
Reduction of $B = 10\%$						
[0%, 25%]	[14.5,43.3]	[167.4,171.7]	[42.3,62.3]	[401.6,414.0]	[86.8,120.5]	[537.6,544.3]
[25%, 50%]	[43.0,44.6]	[171.6,171.7]	[61.9,68.4]	[413.6,422.9]	[119.8,135.1]	[544.1,547.6]
[50%, 75%]	[44.3,46.8]	–	[68.1,77.8]	[422.6,427.3]	[134.2,143.7]	[547.5,549.1]
[75%, 100%]	[46.6,47.3]	[171.6,171.8]	[77.6,78.3]	[427.0,427.9]	[142.9,149.2]	[549.0,550.3]
Reduction of $B = 20\%$						
[0%, 25%]	[14.3,43.3]	[167.4,171.1]	[41.6,58.4]	[404.7,411.1]	[86.8,122.6]	[534.6,544.4]
[25%, 50%]	[43.0,44.6]	–	[57.8,63.8]	[410.8,421.0]	[121.9,131.4]	[544.1,548.2]
[50%, 75%]	[44.3,46.8]	–	[63.5,68.3]	[420.7,425.5]	[130.7,136.6]	[548.1,549.1]
[75%, 100%]	[46.6,47.3]	[171.0,171.8]	[67.7,75.2]	[425.2,426.5]	[135.8,141.6]	[549.0,550.3]
Reduction of $B = 30\%$						
[0%, 25%]	[16.9,44.1]	[167.6,170.4]	[39.0,53.9]	[402.3,407.5]	[84.9,115.7]	[530.7,543.9]
[25%, 50%]	[44.0,44.6]	–	[53.3,59.6]	[407.1,418.6]	[114.8,123.2]	[543.5,547.0]
[50%, 75%]	[44.5,46.8]	–	[59.2,61.4]	[418.2,423.0]	[122.5,125.6]	[546.8,548.3]
[75%, 100%]	[46.7,47.3]	[170.4,171.1]	[60.7,69.9]	[422.6,424.6]	[124.9,128.3]	[548.1,549.2]
Reduction of $B = 40\%$						
[0%, 25%]	[18.7,22.4]	[167.4,169.5]	[33.4,47.2]	[398.6,403.4]	[79.5,104.1]	[532.7,543.4]
[25%, 50%]	[22.0,44.6]	–	[46.5,53.2]	[402.8,415.7]	[103.2,109.5]	[543.0,545.0]
[50%, 75%]	[44.5,45.7]	–	[52.7,54.3]	[415.3,419.2]	[108.7,110.6]	[544.8,545.5]
[75%, 100%]	[45.4,47.2]	[169.5,170.4]	[53.8,61.7]	[418.7,421.7]	[110.0,110.7]	[545.3,546.3]
Reduction of $B = 50\%$						
[0%, 25%]	[19.5,22.5]	[166.3,168.4]	[21.9,39.5]	[395.0,405.6]	[70.2, 89.9]	[533.3,541.0]
[25%, 50%]	[22.1,43.6]	–	[38.6,43.7]	[397.6,412.4]	[88.9, 92.4]	[540.6,542.1]
[50%, 75%]	[43.5,44.6]	–	[43.0,44.0]	[412.0,415.1]	[91.6, 93.5]	[541.9,542.2]
[75%, 100%]	[44.3,46.0]	[168.4,169.5]	[43.3,50.9]	[414.6,416.3]	[92.8, 93.6]	[542.0,542.4]

The influence of the budget constraint is now clearly noticeable. For example, in the $(20, 4, 3)$ problem, the two first quarter ranges for $f_1(f_2^{-1})$ are quite different when we the budget cut increases from $B = 30\%$ to 40%. See also in the problem $(41, 50, 5)$ the ranges $f_1(f_2^{-1})$ for $I = [50\%, 75\%]$ and $[75\%, 100\%]$ when we change from $B = 10\%$ to 20% or the changes in the $f_1(f_2^{-1})$ ranges in the $(60, 5, 4)$ example.

Observe also from Table 5 that some of the $f_2(f_1^{-1})$ ranges for the $(20, 4, 3)$ problem are empty (they are represented by a '–'). This is because in that problem the efficient set is disconnected (similar to figure 3): there are no efficient points at which f_2 take values within interval I.

6 Conclusions and Future Research

We have shown how to tackle a particular biobjective location problem with difficult nonlinear objective functions. Interval-analysis based B&B methods designed for standard single objective global optimization were adapted to obtain good approximations of either some efficient points, or of the whole efficient set. Clearly the presented interval B&B method and the δ-lexicographic method may be applied to any biobjective problem, not only to the particular one presented here.

It should be stressed that our interval B&B method is a practical method able to obtain a superset containing the complete efficient set of any (nonlin-

ear) biobjective problem. In the particular 3-dimensional setting used here the computational burden remained very acceptable. How the method will behave in higher dimension remains to be investigated. It is to be expected that the efficient set will grow in size, however, with a corresponding growth in computation time and memory usage. This also holds for multiple objective settings. Therefore the tests used in our prototype B & B method should be improved, and more sophisticated pruning tests are currently under investigation.

The analysis in our last section gives an indication of the wealth of information obtained by the methods. For the particular application in competitive location used here this kind of analysis should be very useful, e.g. in a negotiation between the franchisor and the franchisee, in which the former might want to convince the latter through some payment to compensate for a lesser income in order for the whole chain to have a larger profit. Having the complete efficient set, and its corresponding two-dimensional graph in the criterion space, may help to reach an agreement, since each can know by how much one of them improves while the other loses. Further exploitation of this type of information is currently under study.

References

1. P.J. Agrell, B.J. Lence, and A. Stam. An intereactive multicriteria decision model for multipurpose reservoir management: the Shellmouth reservoir. Journal of Multi-Criteria Decision Analysis 7:61–86, 1998.
2. H.P. Benson and S. Sayin. Towards finding global representations of the efficient set in multiple objective mathematical programming. Naval Research Logistics 44:47–67, 1997.
3. E. Carrizosa, E. Conde, and D. Romero-Morales. Location of a semiobnoxious facility. A biobjective approach. In: Advances in Multiple Objective and Goal Programming, Springer, pp. 338–346, 1997.
4. V. Chankong and Y.Y Haimes. Multiobjective Decision Making Theory and Methodology, Elsevier Science Publishing Co., Inc., New York, 1983.
5. J.L. Cohon. Multiobjective Programming and Planning, Academic Press, Inc., New York, 1978.
6. T. Drezner. Optimal continuous location of a retail facility, facility attractiveness, and market share: an interactive model. Journal of Retailing 70:49–64, 1994.
7. T. Drezner and Z. Drezner. Validating the gravity-based competitive location model using inferred attractiveness. Annals Oper. Res. 111:227–237, 2002.
8. H.A. Eiselt and G. Laporte. Sequential location problems. European J. Oper. Res., 96:217–231, 1996.
9. H.A. Eiselt, G. Laporte, and J.F. Thisse. Competitive location models: a framework and bibliography. Transportation Science 27:44–54, 1993.
10. M. Ehrgott, K. Klamroth and S. Schwehm. An MCDM approach to portofolio optimization. European J. Oper. Res., 155:752–770, 2004.
11. M. Ehrgott and D.M. Ryan. Constructing robust crew schedules with bicriteria optimization. J. Multi-Criteria Decision Analysis 11:139–150, 2002.

12. J. Fernández, P. Fernández P. and B. Pelegrín. A continuous location model for siting a non-noxious undesirable facility within a geographical region. European J. Oper. Res., 121:259–274, 2002.
13. J. Fernández, P. Fernández and B. Pelegrín. Estimating actual distances by norm functions: a comparison between the $l_{k,p,\theta}$-norm and the $l_{b_1,b_2,\theta}$-norm and a study about the selection of the data set. Computers and Operations Research 29:609-623, 2002.
14. J. Fernández and B. Pelegrín. Sensitivity analysis in continuous location models via interval analysis. Studies in Locational Analysis 14:121–136, 2000.
15. J. Fernández and B. Pelegrín. Using interval analysis for solving planar single-facility location problems: new discarding tests. J. Global Optimization 19:61–81, 2001.
16. J. Fernández, B. Pelegrín, F. Plastria, and B. Tóth. Solving a Huff-like competitive location and design model for profit maximization in the plane, European J. Oper. Res., to appear.
17. J. Figueira, S. Greco, and M. Ehrgott (eds). Multiple Criteria Decision Analysis: State of the Art Surveys. Kluwer, 2004.
18. P.C. Fishburn. Lexicographic orders, utilities and decision rules: a survey. Management Science 20:1442–1471, 1974.
19. T.L. Friesz, T.C. Miller, and R.L. Tobin. Competitive network facility location models: a survey. Papers of the Regional Science Association 65:47–57, 1998.
20. T. Gal and T. Hanne. On the development and future aspects of vector optimization and MCDM. In: Multicriteria Analysis, J. Clímaco (ed.), Springer-Verlag, Berlin, Heidelberg, pp.130–145, 1997.
21. A. Ghosh and C.S. Craig. FRANSYS: a franchise distribution system location model. Journal of Retailing 67:466–495, 1991.
22. R. Hammer, M. Hocks, U. Kulisch and D. Ratz. C++ Toolbox for Verified Computing. Springer-Verlag, Berlin, 1995.
23. E. Hansen. Global Optimization Using Interval Analysis. Marcel Dekker, New York, 1992.
24. P. Hansen and J.F. Thisse. The generalized Weber-Rawls problem. In J.P.Brans (Ed.), Operations Research. North Holland, 1981.
25. M.J. Hodgson. A location-allocation model maximizing consumers' welfare. Regional Studies 15:493–506, 1981.
26. D.L. Huff. Defining and estimating a trading area. Journal of Marketing 28:34–38, 1964.
27. K. Ichida and Y. Fujii. Multicriterion optimization using Interval Analysis. Computing 44:47–57, 1990.
28. A.K. Jain and V. Mahajan. Evaluating the competitive environment in retailing using multiplicative competitive interactive models. In: J. Sheth (Ed). Research in Marketing, JAI Press, pp.217–235, 1979.
29. M. Kilkenny and J.F. Thisse. Economics of location: a selective survey. Computers and Operations Research 26:1369–1394, 1999.
30. R.B. Kearfott. Rigorous Global Search: Continuous Problems. Kluwer, Dordrecht, 1996.
31. R.B. Kearfott, M.T. Nakao, A. Neumaier, S.M. Rump, S.P. Shary, and P. van Hentenryck. Standardized notation in interval analysis, 2002, submitted to Reliable Computing.
32. O. Knüppel. PROFIL/BIAS – a fast interval library. Computing 53:277–287, 1994.

33. K.H. Küfer, A. Scherrer, M. Monz, F. Alonso, H. Trinkaus, T. Bortfeld, and C. Thieke. Intensity-modulated radiotherapy - a large scale multi-criteria programming problem. OR Spectrum 25:223–249, 2003.
34. M.C. Markót, J. Fernández, L.G. Casado, and T. Csendes. New interval methods for constrained global optimization. Mathematical Programming, to appear.
35. K.S. Miettinen. Nonlinear Multiobjective Pptimization. Kluwer, Boston, 1998.
36. N. Nakanishi and L.G. Cooper. Parameter estimate for mutiplicative interactive choice model: least square approach. J. Marketing Research 11:303–311, 1974.
37. F. Plastria. GBSSS: The generalized big square small square method for planar single-facility location. European J. Oper. Res., 62:163–174, 1992.
38. F. Plastria. Static competitive facility location: an overview of optimisation approaches. European J. Oper. Res., 129:461–470, 2001.
39. H. Ratschek and J. Rokne. New Computer Methods for Global Optimization. Ellis Horwood, Chichester, 1988.
40. S. Sayin. Measuring the quality of discrete representations of efficient sets in multiple objective mathemathical programming. Mathematical Programming 87:543–560, 2000.
41. R. Schmalensee and J.F. Thisse. Perceptual maps and the optimal location of new products: an integrative essay. Int. J. Research in Marketing 5: 225–249, 1988.
42. M.J. Schniederjans and E. Hollcroft. A multi-criteria modeling approach to jury selection. Socio-Economic Planning Sciences 39:81–102, 2005.
43. J. Silverman, R.E. Steuer, and A.W. Whisman. A multi-period, multiple criteria optimization system for manpower planning. European J. Oper. Res., 34:160–170, 1988.
44. A.J.V. Skriver and K.A. Andersen. The bicriterion semi-obnoxious location (BSL) problem solved by an ϵ-approximation. European J. Oper. Res., 146:517–528, 2003.
45. R.E. Steuer. Multiple Criteria Optimization: Theory, Computation, and Applications. John Wiley & Sons, Inc., 1986.
46. B. Tóth and T. Csendes. Empirical investigation of the convergence speed of inclusion functions. Reliable Computing 11:253-273, 2005.
47. B. Tóth, J. Fernández, F. Plastria and B. Pelegrín. Location and design of a new facility in a competitive planar market: a biobjective analysis and an interval lexicographical-like solution procedure, 2004, submitted.
48. R.E. Wendell and R.D. McKelvey. New perspectives in competitive location theory. European J. Oper. Res., 6:174–182, 1981.
49. D.J. White. A bibliography on the applications of Mathematical Programming multiple-objective methods. J. Oper. Res. Soc., 41:669–691, 1990.
50. P.L. Yu. Multiple-Criteria Decision Making Concepts, Techniques and Extensions. Plenum Press, New York, 1985.
51. M. Zeleny. Multiple Criteria Decision Making. McGraw-Hill, Inc., 1982.

Tools for Robotic Trajectory Planning Using Cubic Splines and Semi-Infinite Programming

A. Ismael F. Vaz and Edite M.G.P. Fernandes

Departamento de Produçcção e Sistemas, Escola de Engenharia, Universidade do Minho Campus de Gualtar, 4710-057 Braga, Portugal.
{aivaz,emgpf}@dps.uminho.pt

Summary. In this paper we describe how robot trajectory planning, using cubic splines to generate the trajectory, can be formulated as standard semi-infinite programming (SIP) problems and efficiently solved by a discretization method. These formulated problems were coded in the publicly available SIPAMPL environment and to allow the codification of these problems a cubic splines dynamic library for AMPL was developed. The discretization method used to solve the formulated problems is implemented in the NSIPS solver and numerical results with four particular problems are shown.

1 Introduction

In the last decades many engineering problems have been formulated as semi-infinite programming problems (cf. [9]), and robot trajectory planning ([7, 8, 10, 11, 12, 14]) is not an exception.

In robot trajectory planning two major approaches can be considered. In the first one, the robot trajectory is known *a priori* and the optimization consists of computing the best reparametrization of the trajectory, by minimizing the total travel time (or energy consumption), while some robot physical limitations are taken into consideration (maximum velocity, acceleration and torque of the links). The reparametrization is obtained by using B-Splines that depend on a set of control points. The second approach consists of finding the robot trajectory, where only a set of via points are known. These points are interpolated by the trajectory of the robot and since B-Splines are not interpolation functions the use of cubic splines (C-Splines) is more appropriate. The optimization consists in minimizing the total travel time when passing through the given trajectory points, subject to the robot physical limitations.

Some recent available tools [16, 18] have provided an easy and fast way of coding and solving SIP problems, by taking advantage of available modeling software (AMPL [9]). The SIPAMPL package relies on the AMPL modeling software, which provides the most used mathematical functions (exp, sin, cos,

etc) for coding mathematical problems as well as automatic differentiation. To allow the codification of the reparametrization optimal trajectory planning problems, an external B-Spline dynamic library was built for AMPL [15].

In this paper we will describe how robot trajectory planning problems (using C-Splines to generate the trajectory) formulated as semi-infinite programming problems can be coded in AMPL and efficiently solved by a discretization method. This formulation follows the ideas in [7, 10, 11].

The B-Splines library publicly available in SIPAMPL is now extended to include the C-Splines. The new library is now called Splines and includes both the B and C-Splines.

We start in Section 2 with a brief introduction on the trajectory definition. Section 3 will be devoted to presenting the robotics problems formulated as SIP problems. Section 4 presents the C-Splines and Section 5 shows the coded problems for which we present some numerical results (in Section 6). Some conclusions are presented in the last section.

2 Trajectory Definition

For a brief review of robotics mechanics and control we refer to [2].

A robot can be schematically represented as a sequence of rigid links (arms) connected by joints. In the most schematic form, a link can be represented as a rigid body of length d and mass m, thus d_i, m_i represent the length and mass of the ith robot link. Some mechanical devices provide the joint torque which move the links. The number of joints which can be independently actuated defines the manipulator *degrees of freedom* (d.o.f.). Often, all joints can be independently controlled, so that the manipulator d.o.f. coincide with the number of joints. Each degree of freedom is associated to an independent variable $\theta_i, i = 1, 2, \ldots, l$, where l is the number of d.o.f..

Let us define the manipulator joint space as the space where joint variables θ_i can span

$$\theta := [\theta_1 \theta_2 \ldots \theta_l]^T.$$

Evidently, vector θ completely defines the manipulator position in the cartesian space. Since the robot position (d.o.f. values) varies, we can define the path as a curve

$$\theta(\tau) = [\theta_1(\tau), \theta_2(\tau), \ldots, \theta_l(\tau)]^T, \quad \tau \in [0, \tau_f], \tag{1}$$

parametrized by τ, where τ_f is the total travel time.

Possible constraints applied to the parametric curve are:

- given initial and final velocities,

$$\frac{d\theta}{d\tau}(0) = v_i \quad \text{and} \quad \frac{d\theta}{d\tau}(\tau_f) = v_f$$

- given initial and final accelerations,

$$\frac{d^2\theta}{d\tau^2}(0) = a_i \quad \text{and} \quad \frac{d^2\theta}{d\tau^2}(\tau_f) = a_f.$$

3 Optimal Cubic Polynomial Joint Trajectories

The problem can be formulated as follows. Let us suppose that the designed trajectory has to cross n assigned via (or knot) points in the cartesian space. These points can be converted into joint space via points by solving an inverse kinematics problem, see [3], thus obtaining

$$\theta(\tau_0), \theta(\tau_1), \ldots, \theta(\tau_n) \ ,$$

where the generic $\theta(\tau_i)$ is defined according to (1). The optimization consists of finding the minimal total displacements time that fits the joint trajectory by using cubic splines constrained to velocity, acceleration, jerk and torque bounds. Let $t_0 < t_1 < \cdots < t_n$ be a time sequence where t_i represents the time instant associated to the manipulator position $\theta(\tau_i)$. Let $h_1 = t_1 - t_0$, $h_2 = t_2 - t_1$, ..., $h_n = t_n - t_{n-1}$ be the time displacements. Introduce a different cubic spline segment Q_{ij} for each time displacement h_j and for each joint variable θ_i. Identify by $Q_i(t)$, $t \in [t_0, t_n]$ the overall trajectory of joint i, obtained by concatenating Q_{ij} for $j = 1, 2, \ldots, n$.

We will use the notation $Q'(t) = \frac{dQ(t)}{dt}$ for the derivative.

We are thus led to consider the optimization problem:

$$\begin{aligned} \min_{h_j} \ & \sum_{j=1}^{n} h_j \\ s.t. \ & |Q_i'(t)| \leq C_{i,1} \\ & |Q_i''(t)| \leq C_{i,2} \\ & |Q_i'''(t)| \leq C_{i,3} \\ & |F_i(t)| \leq C_i, \ \ i = 1, ..., l \\ & h_j > 0, \ \ j = 1, ..., n \\ & \forall t \in [t_0, t_n] \end{aligned} \tag{2}$$

where $C_{i,1}$, $C_{i,2}$, $C_{i,3}$ and C_i are the bounds for the velocity, acceleration, jerk and torque, respectively, on joint i.

The optimization (2) falls into the category of what is usually called a generalized semi-infinite programming problem (cf. [15]). Notice that t ranges over $[t_0, t_n]$, an interval that depends implicitly on the decision variables h_j's.

The expression for the ith joint torque is

$$F_i(t) = J_i n_i Q_i''(t) + B_i n_i Q_i'(t) + \frac{1}{n_i}\left(\sum_{j=1}^{l} I_{ij}(Q(t))Q_j''(t) + \sum_{j=1}^{l}\sum_{k=1}^{l} C_{ijk}(Q(t))Q_j'(t)Q_k'(t) + d_i(Q(t))\right)$$

where J_i, n_i, B_i, $I_{ij}(\cdot)$, $C_{ijk}(\cdot)$ and $d_i(\cdot)$ have the usual meaning (cf. [15]).

In [10], an optimization of cubic polynomial joint trajectory is proposed, based on this formulation with bound constraints on velocity, acceleration and jerk. Initial and final velocities and accelerations were given. The optimization problem was solved with a Nelder and Mead [19] polyhedron search applied to the objective function and a technique that converts an infeasible vertex into a feasible one by a simple procedure that consists of multiplying the infeasible vertex by a suitable scalar. In a later paper, De Luca *et al.* [11] proposed an approach to solve a similar problem where the bound constraints were on velocity and torque and only the initial and final accelerations are imposed. De Luca *et al.* applied the algorithm from Gonzaga *et al.* [6] to the generalized SIP problem. Since the implemented algorithm needs first derivatives, De Luca *et al.* developed a sensitivity approach to the C-Splines. Guarino Lo Bianco and Piazzi [7] proposed an approach based on the generalized SIP problem with bound constraints on linear and angular velocities, and torque. The proposed genetic algorithm is applied to a penalty function computed by means of interval analysis.

By using the linear transformation $t = \tau \sum_{k=1}^{n} h_k + t_0$, the generalized SIP problem (2) can be reformulated in the form

$$\begin{aligned}
&\min_{h_j} \sum_{j=1}^{n} h_j \\
&s.t. \quad \left|Q_i'\left(\tau \sum_{k=1}^{n} h_k + t_0\right)\right| \le C_{i,1} \\
&\left|Q_i''\left(\tau \sum_{k=1}^{n} h_k + t_0\right)\right| \le C_{i,2} \\
&\left|Q_i'''\left(\tau \sum_{k=1}^{n} h_k + t_0\right)\right| \le C_{i,3} \\
&\left|F_i\left(\tau \sum_{k=1}^{n} h_k + t_0\right)\right| \le C_i, \quad i = 1, ..., l \\
&h_j > 0, \quad j = 1, ..., n, \\
&\forall \tau \in [0, 1].
\end{aligned} \tag{3}$$

This new problem fits into the category of what is usually called a standard SIP problem. Notice that τ ranges over the fixed interval $[0, 1]$. For SIP problems of this type one can apply the tools presented recently in [16, 18].

4 Cubic Splines

In this section, we assume that we are dealing with only one joint of the robot, i.e., $Q(t)$ is the function used to interpolate the trajectory at the $\theta(\tau)$ knots and $Q_j(t)$, $j = 1, \ldots, n$, are the cubic spline segments in $[t_{j-1}, t_j]$.

Given a finite number of data points, $\theta_0 = \theta(\tau_0)$, $\theta_1 = \theta(\tau_1)$, $\ldots$, $\theta_n = \theta(\tau_n)$, a C-Spline is formed by n cubic polynomials ($Q_j(t)$, $i = j, \ldots, n$) that interpolate the given data points. The set of the $Q_j(t)$, $j = 1, \ldots, n$ will provide a cubic interpolation at the given trajectory knots. Since $Q_j(t)$ are cubic polynomials, the second derivative with respect to t can be expressed as

$$Q_j''(t) = \frac{t_j - t}{t_j - t_{j-1}} M_{j-1} + \frac{t - t_{j-1}}{t_j - t_{j-1}} M_j, \quad j = 1, \ldots, n$$

where M_j is the second derivative of $\theta(\tau)$ at τ_j. Integrating $Q_j''(t)$ twice and imposing the conditions $Q_j(t_{j-1}) = \theta_{j-1}$ and $Q_j(t_j) = \theta_j$ results in the following interpolating functions:

$$\begin{aligned} Q_j(t) = & \frac{M_{j-1}}{6h_j}(t_j - t)^3 + \frac{M_j}{6h_j}(t - t_{j-1})^3 \\ & + \left(\frac{\theta_{j-1}}{h_j} - \frac{h_j M_{j-1}}{6}\right)(t_j - t) \\ & + \left(\frac{\theta_j}{h_j} - \frac{h_j M_j}{6}\right)(t - t_{j-1}), \quad j = 1, \ldots, n, \end{aligned}$$

where $h_j = t_j - t_{j-1}$.

The C-Spline is completely defined if the M_j, $j = 0, \ldots, n$, are known. Problems to generate trajectories need to specify the first and second derivatives at the extreme of the spline (initial and final velocities, and accelerations). Imposing the continuity of the first derivative, $Q_j'(t_j) = Q_{j+1}'(t_j)$, $j = 1, \ldots, n-1$, results in a tridiagonal system from where the M_j, $j = 1, \ldots, n-1$, can be obtained. A natural C-Spline would be completely specified by considering $M_0 = a_i$, $M_n = a_f$ and therefore imposing the initial and final conditions on the velocities ($\theta'(\tau_0) = v_i$ and $\theta'(\tau_n) = v_f$) would not be possible.

Two more degrees of freedom are necessary and the goal is achieved by considering two extra knots where the θ values are not specified. Consider, without loss of generality, that t_1 and t_{n-1} are the extra knots. Solving

$$Q_1'(t_0) = v_i, \quad Q_n'(t_n) = v_f$$

for the two unknowns θ_1 and θ_{n-1} results in

$$\theta_1 = \theta_0 + h_1 v_i + \frac{h_1^2 M_0}{3} + \frac{h_1^2 M_1}{6}$$
$$\theta_{n-1} = \theta_n - v_f h_n + \frac{h_n^2 M_n}{3} + \frac{h_n^2 M_{n-1}}{6}.$$

Replacing θ_1 and θ_{n-1} in the natural C-Spline tridiagonal system results in the new tridiagonal system

$$A \begin{pmatrix} M_1 \\ M_2 \\ \dots \\ M_{n-2} \\ M_{n-1} \end{pmatrix} = B$$

where the matrix A and the vector B are given respectively by:

$$\begin{pmatrix} 3h_1 + 2h_2 + \frac{h_1^2}{h_2} & h_2 & & & & & 0 \\ h_2 - \frac{h_1^2}{h_2} & 2(h_2+h_3) & h_3 & & & & \\ & & \vdots & & & & \\ & h_j & 2(h_j + h_{j+1}) & h_{j+1} & j = 3, \dots, n-3 & & \\ & & \vdots & & & & \\ & & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} - \frac{h_n^2}{h_{n-1}} & \\ 0 & & & & h_{n-1} & 3h_n + 2h_{n-1} + \frac{h_n^2}{h_{n-1}} & \end{pmatrix}$$

$$\begin{pmatrix} 6\left(\frac{\theta_2}{h_2} + \frac{\theta_0}{h_1}\right) - 6\left(\frac{1}{h_1} + \frac{1}{h_2}\right)\left(\theta_0 + h_1 v_i + \frac{h_1^2 M_0}{3}\right) - h_1 M_0 \\ \frac{6}{h_2}\left(\theta_0 + h_1 v_i + \frac{h_1^2 M_0}{3}\right) + \frac{6\theta_3}{h_3} - 6\left(\frac{1}{h_2} + \frac{1}{h_3}\right)\theta_2 \\ \dots \\ 6\left(\frac{\theta_{j+1}-\theta_j}{h_{j+1}} - \frac{\theta_j - \theta_{j-1}}{h_j}\right), \quad j = 3, \dots, n-3 \\ \dots \\ \frac{6}{h_{n-1}}\left(\theta_n - \theta_{n-2} - v_f h_n + \frac{h_n^2 M_n}{3}\right) - 6\left(\frac{\theta_{n-2}-\theta_{n-3}}{h_{n-2}}\right) \\ -6\left(\frac{\theta_n - \theta_{n-2} - v_f h_n}{h_{n-1}}\right) + 6\left(v_f - \frac{h_n M_n}{3} - \frac{h_n^2 M_n}{3h_{n-1}}\right) - h_n M_n \end{pmatrix}.$$

When formulating an optimization problem involving cubic splines we may need to impose limits on the velocity, acceleration, jerk and torque, and therefore the derivatives of the C-Splines are requested. When solving the formulated problem the first and second derivatives (with respect to the t and h_j variables) of the constraints involving the C-Spline may be requested by AMPL. The C-Splines dynamic library, built for AMPL, only provides the first derivatives with respect to the t and h_j variables and these formulae can be consulted in [17].

5 Robotics Problems Coded in AMPL

A library for AMPL to allow the codification of mathematical problems that use C-Splines is available and will be described in the following subsections. The C-Splines library can be used to code the optimal trajectory planning (SIP) problems that appear in robotics.

In this section a brief introduction to AMPL is given. We also describe the C-Splines library for AMPL and give an example on how this library is used for coding the robotics problems.

5.1 (SIP)AMPL

AMPL [9] is a modelling language that allows the codification of mathematical programming problems. AMPL is a descriptive language and does not allow the direct programming of functions (no recursion is allowed) and the automatic differentiation is used only to obtain the derivatives of the objective function and constraints.

The SIPAMPL package contains a database of more than one hundred and sixty SIP problems coded in (SIP)AMPL format and a set of interface routines to connect AMPL to any SIP solver (in particular to the NSIPS [18] solver). See [17] for the internet address where SIPAMPL is publicly available, and [16] for details concerning the (SIP)AMPL format.

5.2 The Splines Library

Some problems in mathematical programming resort to function approximation by splines [1]. The codification of such problems in AMPL can be very tedious if one needs to use several splines and their derivatives.

AMPL has the ability to load a dynamic library and we made an extension to a previous B-Splines library [15] to include the C-Splines. The Splines library and installation procedure are available from the internet along with the SIPAMPL package.

The library provides three functions. `bspline` and `dbspline` refer to the B-Splines functions (see [15]). `cspline` provides the C-Spline and its derivatives to AMPL. The function syntax is:

$$\texttt{cspline}\ (t, d, n, h_1, h_2, \ldots, h_n, \theta_1, \theta_2, \ldots, \theta_{n-1}, v_i, v_f, a_i, a_f)$$

where t is the time instant where the cubic spline is to be computed and h_1, $\ldots$, h_n are the time displacements, all defined as AMPL variables. θ_1, $\ldots$, θ_{n-1} are the assigned knots, d is the derivative order (0 for the C-Spline, 1 for the first, 2 for the second and 3 for the third derivative with respect to t), n is the number of subintervals, v_i, v_f are the first and a_i, a_f the second derivatives at the boundaries (initial and final velocities, and accelerations). The latter arguments are defined as AMPL parameters.

`cspline` returns to AMPL the C-Spline (or the d order derivative) value at t and, if requested by the solver, the first derivatives of the C-Spline with respect to t, and h_j variables. When an external function is required the `function` AMPL command must be used, before using the function itself.

5.3 A Robotics Example

Consider the following cubic splines trajectory planning problem. The total travel time is to be minimized while the velocity is to be bounded by constants (only one simple constraint is imposed, since the purpose is to illustrate how C-Splines can be used in AMPL, and not how to solve a real robotics problem)

$$\begin{aligned} \min_{h \in R^7} \quad & \sum_{j=1}^{7} h_j \\ \text{s.t.} \quad & -2 \leq Q'\left(\tau \sum_{j=1}^{7} h_j\right) \leq 2 \\ & h_j > 0, \quad \forall \tau \in [0,1]\,. \end{aligned}$$

The trajectory knots are 0, 0.5, 0.75, 1, 0.75, 1.5, $v_i = v_f = 0$, $a_i = 13.880794$ and $a_f = -0.415203$.

A possible codification of this problem in (SIP)AMPL format is the following:

```
#declareusedexternalfuntion
functioncspline;
#numberofcoefficients
paramn:=7;
#numberofknots(nk=n-1)
paramnk:=6;
#knotsvector
paramknots{1..nk};
#initialguessfortimedisplacements
paramhinit{1..n};
#timedisplacements
varh{jin1..n}:=hinit[j];
#Infinitevariable
vartau:=0;
#initialandfinalvelocities,and
#accelerations
paramvi:=0;
paramvf:=0;
paramai:=13.880794;
paramaf:=-0.415203;
#definenewvariableforC-spline
varg=cspline(tau*(sum{jin1..n}(h[j])),
 1,n,{jin1..n}h[j],{jin1..nk}knots[j],
 vi,vf,ai,af);
minimizeobj:
 (sum{jin1..n}(h[j]));
subjecttotcons:
 -2<=g<=2;
subjecttobounds{jin1..n}:
 h[j]>=0.01;
subjecttotbounds:
 0<=tau<=1;
```

```
data;
#knots
paramknots:=
 10
 20.5
 30.75
 41
 50.75
 61.5;
#initialguess
paramhinit:=
 11
 21
 30.5
 40.5
 50.5
 60.5
 70.5;
#auxiliaryfilesforNSIPS
optionnsips_auxfilesrc;
#selectdiscretizationmethod
#andinitialgridspace
optionnsips_options
 'method=disc_hettdisc_h=0.01';
#selectNSIPS
optionsolvernsips;
#solveproblem
solve;
#printthesolution
printf"Solutionfound\n";
displayh;
displayobj;
```

5.4 Coded SIP Problems

The constraints $h_j > 0$, $j = 1, \ldots, n$, in (3) were replaced by $h_j \geq 0.01$, $j = 1, \ldots, n$ in all problems, to avoid the proximity of zero.

The problem described in [10] was coded in the `lin2.mod` file. It describes an Unimate PUMA 560 type robot with 6 revolute joints were no torque limits are imposed. The problem is a minimum time trajectory planning with bound constraints on velocity, acceleration and jerk. The robot starts and ends in rest position ($v_i = v_f = a_i = a_f = 0$). The initial time intervals are $h = [3.607, 3.607, 2.878, 4.275, 5.612, 2.915, 5.879, 1.336, 1.336]$, giving a total time of 31.445 seconds ($t_0 = 0$, $t_n = 31.445$) and the remaining data is presented in Table 1.

Table 1. Data for the `lin2` problem

knot	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
	position in *degrees*					
1	10	15	45	5	10	6
2	60	25	180	20	30	40
3	75	30	200	60	-40	80
4	130	-45	120	110	-60	70
5	110	-55	15	20	10	-10
6	100	-70	-10	60	50	10
7	-10	-10	100	-100	-40	30
8	-50	10	50	-30	10	20
Bounds	Joint 1	Joint 2	Joint 3	Joint 4	Joint 5	Joint 6
Velocity (*degrees/sec*)	100	95	100	150	130	110
Acceleration ($degrees/sec^2$)	45	40	75	70	90	80
Jerk ($degrees/sec^3$)	60	60	55	70	75	70

The two problems reported in [11] were also coded. `deluca1.mod` is a light robot with 2 joints and `deluca2.mod` is a planar motion of an IBM 7535 robot with 2 joints. The robot parameters are reported in Table 2, where l_i and J_i ($i = 1, 2$) are the length and moment of inertia, with respect to the axis of the driving joint for link i, m_2 is the mass of link 2, while m_p and J_p are the mass and centroidal inertia of the payload. d_2 is the distance between the axis of the second link joint and the center of mass of the second link. The dynamic equation can be consulted in the model files `deluca1.mod` or `deluca2.mod` in [17] or in [11].

These problems contain velocity and torque bound constraints. The velocity limit was 2 rad/sec for both joints and 7 Nm and 2 Nm are the torque limits in joint 1 and joint 2, respectively. The cubic splines used by De Luca

Table 2. Robot parameters for the `deluca1-2` problems

	l_1 (m)	l_2 (m)	d_2 (m)	m_2 (kg)	m_p (kg)	J_1 $(kg\ m^2)$	J_2 $(kg\ m^2)$	J_p $(kg\ m^2)$
`deluca1`	0.5	0.5	0.25	1	0	0.084	0.084	0
`deluca2`	0.4	0.25	0.125	15	6	1.6	0.34	0.01

et al. in [11] only consider initial and final zero velocities and we have computed the initial and final accelerations that result from the solution presented by the authors. The initial and final velocities, and accelerations used in the coded problems are reported in Table 3.

Table 3. Initial and final velocities, and accelerations for the `deluca1-2` problems

Problem	Joint	v_i (rad/sec)	v_f (rad/sec)	a_i (rad/sec^2)	a_f (rad/sec^2)
`deluca1`	1	0	0	13.880794	-0.415203
	2	0	0	-11.067942	-4.186542
`deluca2`	1	0	0	2.5207742	-2.1966904
	2	0	0	2.5207742	-2.1966904

The initial time intervals, in seconds, considered were $h = [1, 1, 0.5, 0.5, 0.5, 0.5, 0.5]$ and $h = [0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3]$ for `deluca1` and `deluca2`, respectively. Since De Luca *et al.* did not use the acceleration constraints, in order to get the necessary extra freedom, we have repeated the first and the last time intervals. The via points in joint space are reported in Table 4.

Table 4. Via points for the `deluca1-2` problems

Problem	Joint	Position in *radians* (knots)					
		1	2	3	4	5	6
`deluca1`	1	0	0.5	0.75	1	1.25	1.5
	2	0	-0.5	-1	-1.5	-1	0.5
`deluca2`	1 and 2	0.1	0.2	0.25	0.3	0.35	0.4

Another coded problem (`lobianco1.mod`) refers to a robot with 2 joints. The robot arm is considered in initial and final rest position ($v_i = v_f = a_i = a_f = 0$). The problem considers torque, linear and angular velocity limits of 260 Nm, 50 Nm, 0.7 m/sec and 1.5 rad/sec, respectively. The robot parameters are $l_1 = 1.0(m)$, $l_2 = 0.5(m)$, $m_1 = 15.0(kg)$ and $m_2 = 7.0(kg)$, where l_i and m_i, $i = 1, 2$, are the link lengths and masses, respectively.

Since Guarino Lo Bianco and Piazzi [7] proposed a genetic algorithm to solve the problem an initial guess was not provided. We used $h = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]$ as initial guess which gives a total time travel of 5.5 *sec.*

The dynamic equations, linear and angular velocity equations can be consulted in the coded problem (`lobianco1.mod` in [17]) and the trajectory via points are presented in Table 5.

Table 5. Via points for the `lobianco1` problem

Knot	Joint 1 (*rad*)	Joint 2 (*rad*)
1	0.0000	-1.5708
2	0.1253	-1.6804
3	0.2517	-1.7594
4	0.3789	-1.8074
5	0.5054	-1.8235
6	0.5837	-1.7087
7	0.6119	-1.4581
8	0.4263	-1.1040
9	0.3903	-1.1124
10	0.3526	-1.1152

6 Numerical Experiments

We have used the recent publicly available NSIPS [18] solver connected with the SIPAMPL [16] interface that provides the Splines library. The NSIPS solver software package can be obtained freely from the same internet address as SIPAMPL. The discretization method is the only one that supports simple bounds on the finite variables ($h_j > 0, j = 1, \ldots, n$, in (3)). The discretization method consists of replacing the infinite set T by a sequence of finite grids $T_0 \subset T_1 \subset \cdots \subset T_k \subset T$. Finite nonlinear subproblems are solved (using the NPSOL software package [4]) in each grid with a selected number of points.

All plots were obtained with the MATLAB [13] `plot` function together with the SIPAMPL interface to MATLAB [16, 17].

The numerical results were obtained in a computer with a Pentium III 450Mhz processor, 128MB of RAM, Linux operating system (Red Hat 5.2) and with AMPL Student Version number 19991027 (Linux 2.0.18).

The solution found for the problem described in subsection 5.3 was 0.281396, 0.0650184, 0.128647, 0.178947, 0.226906, 0.621131 and 0.01 in a total travel time of 1.51205. The trajectory and velocity are plotted in Fig.1.

Table 6 presents the user time in seconds for all the solved problems.

The time solution vector for all the problems are shown in Tables 7 and 8. In these tables, "NSIPS" is the solution obtained with the NSIPS solver and "Prev." is the solution obtained by the authors where the problem was proposed. In Table 8, the "Extra knot" entry means that the considered knot was not present in the authors proposed problem, since the extra degree of freedom in the C-Spline was not used. The discretization method computes a

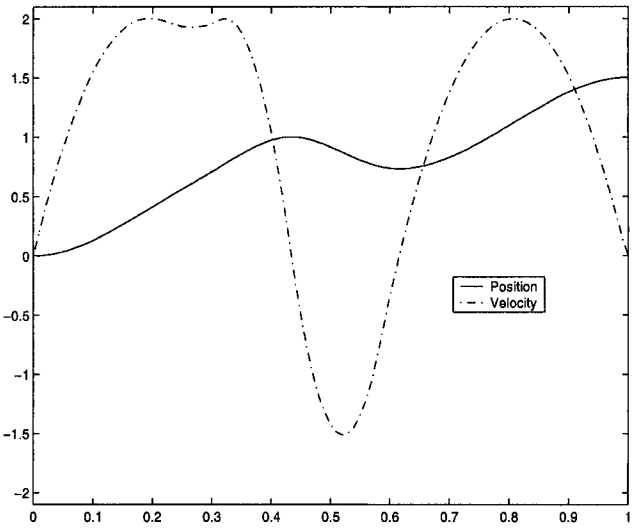


Fig. 1. Results for position and velocity of the example problem

Table 6. User time, in seconds, for solved problems

Problem	Time (*sec*)
`lin2`	6.59
`deluca1`	6.57
`deluca2`	17.71
`lobianco1`	3.11

solution in the finest grid and if the grid is not fine enough some infeasibility in the infinite constraints can occur. If a finer grid is required then the dimension of the first grid or the number of grid refinements should be changed from their default values in the solver. For problem `deluca2` the solution we found is different from the one reported by De Luca *et al.*. While in the solution found by De Luca *et al.* the velocity constraints are inactive and the torque are active, in our solution the velocity constraints are active and the torque inactive.

Except for the `lobianco1` problem, an improvement in the total trajectory time was obtained for all the considered problems.

Figure 2 shows the position, velocity, acceleration and jerk in the solution found for the first joint in `lin2` problem. Figures 3-5 show the position, linear and angular velocities and torque for problem `lobianco1`.

Table 7. Solutions of `lin2` and `lobianco1` problems

	`lin2`		`lobianco1`	
	NSIPS	Prev.	NSIPS	Prev.
h_1	1.125150	1.131000	0.010000	0.020000
h_2	2.039520	2.004000	0.348599	0.364290
h_3	1.635940	2.068000	0.156699	0.184190
h_4	2.158020	2.016000	0.150559	0.183860
h_5	2.046600	2.714000	0.154683	0.184230
h_6	2.510830	1.973000	0.138140	0.167350
h_7	3.781200	3.807000	0.191483	0.223100
h_8	1.831450	1.971000	0.391619	0.365390
h_9	0.803105	0.767000	0.106903	0.099450
h_{10}			0.022616	0.238180
h_{11}			0.385888	0.020050
Total	17.931800	18.45100	2.057190	2.050090

Table 8. Solutions of `deluca1` and `deluca2` problems

	`deluca1`		`deluca2`	
	NSIPS	Prev.	NSIPS	Prev.
h_1	0.010000	0.370000	0.010000	0.290000
h_2	0.348255	Extra knot	0.134696	Extra knot
h_3	0.260631	0.250000	0.053838	0.070000
h_4	0.361528	0.340000	0.050615	0.070000
h_5	0.351404	0.430000	0.051988	0.080000
h_6	0.010000	Extra knot	0.066819	Extra knot
h_7	1.061350	1.070000	0.010000	0.200000
Total	2.403170	2.460000	0.377956	0.710000

7 Conclusions

Previous works on trajectory planning have solved the optimization problem under the generalized SIP formulation shown in equation (2) and new algorithms for addressing this problem were proposed. In this paper we present a reformulation of the generalized SIP into a standard SIP and applied publicly available tools for coding and solving this type of problems.

To allow the codification of the robotics problems a C-Splines dynamic library for AMPL was developed. The robotics problems herein presented are coded and freely available via the web in the SIPAMPL [17] database (see subsection 5.4).

We have shown that the formulated robot trajectory planning problems can be easily solved with the new available tools for semi-infinite programming problems.

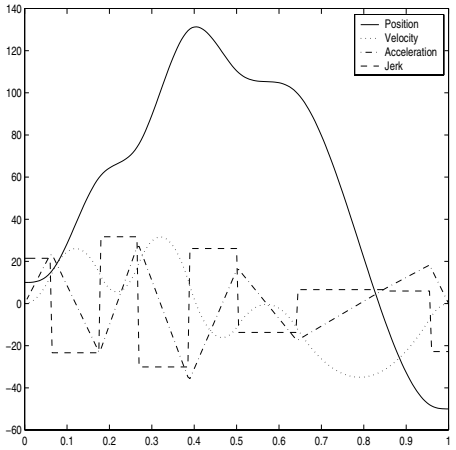


Fig. 2. Results for joint 1 of `lin2`

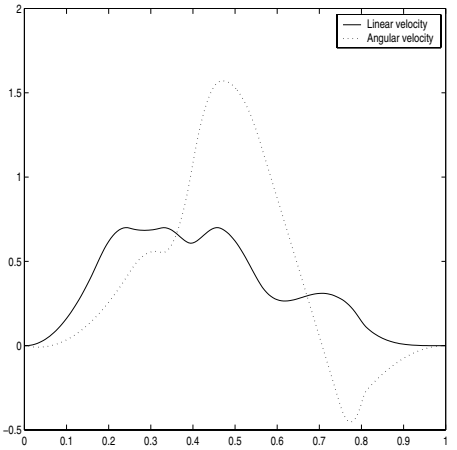


Fig. 3. Results for linear and angular velocity of `lobianco1`

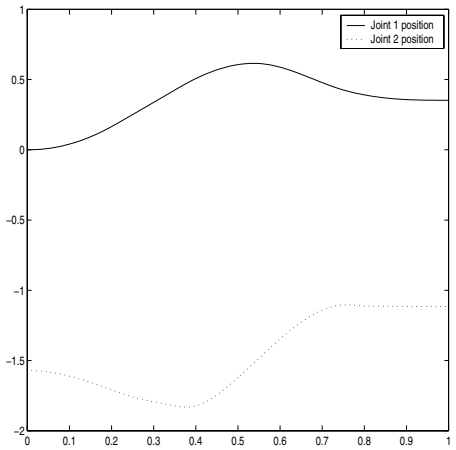


Fig. 4. Results for position of `lobianco1`

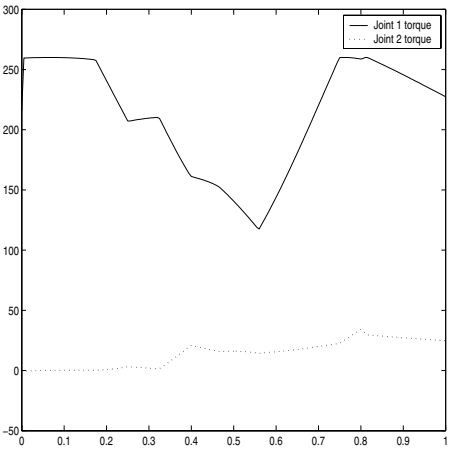


Fig. 5. Results for torque of `lobianco1`

Acknowledgments. The authors are grateful to two anonymous referees for their interest and their useful comments and suggestions that greatly improved the paper.

References

1. C. Boor. A Pratical Guide to Splines. Springer-Verlag, 1978.
2. J.J. Craig. Introduction to Robotics, Mechanics and Control. Addison-Wesley, second edition, 1989.

3. R. Fourer, D.M. Gay, and B.W. Kernighan. A modeling language for mathematical programming. Management Science, 36(5):519–554, 1990.
4. P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright. User's Guide for NPSOL: A Fortran Package for Nonlinear Programming. Stanford University, 1986.
5. M. Goberna and M. López (Eds.). Semi-Infinite Programming: Recent Advances, volume 57 of Nonconvex Optimization and its Applications. Kluwer Academics Publishers, Dordrecht, The Netherlands, 2001.
6. C. Gonzaga, E. Polak, and R. Trahan. An improved algorithm for optimization problems with functional inequality constraints. IEEE Trans. Automatic Control, AC-25(1):49–54, 1980.
7. C. Guarino Lo Bianco and A. Piazzi. A semi-infinite optimization approach to optimal spline trajectory planning of mechanical manipulators. In [5], pages 271–297, 2001.
8. E. Haaren-Retagne. A Semi-Infinite Programming Algorithm for Robot Trajectory Planning. PhD thesis, Univ. of Trier, 1992.
9. R. Hettich and K.O. Kortanek. Semi-infinite programming: Theory, methods, and applications. SIAM Review, 35(3):380–429, 1993.
10. C.-S. Lin, P.-R. Chang, and J. Y. S. Luh. Formulation and optimization of cubic polynomial joint trajectories for industrial robots. IEEE Trans. Automatic Control, AC-28(12):1066–1074, 1983.
11. A. De Luca, L. Lanari, and G. Oriolo. A sensitivity approach to optimal spline robot trajectories. Automatica, 27(3):535–539, 1991.
12. S.P. Marin. Optimal parametrization of curves for robot trajectory design. IEEE Trans. Automatic Control, 33(2):209–214, 1988.
13. MathWorks. MATLAB. The MathWorks Inc., 1999. Version 5.4, Release 11.
14. O. Von Stryk and M. Schlemmer. Optimal control of the industrial robot manutec r3. In: R. Bulirsch, D. Kraft (eds.), Computational Optimal Control, Int. Series of Numerical Mathematics, 115:367–382, 1994.
15. A.I.F. Vaz, E.M.G.P. Fernandes, and M.P.S.F. Gomes. Robot trajectory planning with semi-infinite programming. European J. Operational Research, 53(3):607–617, 2004.
16. A.I.F. Vaz, E.M.G.P. Fernandes, and M.P.S.F. Gomes. SIPAMPL: Semi-infinite programming with AMPL. ACM Trans. Mathematical Software, 30(1):47–61, 2004.
17. A.I.F. Vaz, E.M.G.P. Fernandes, and M.P.S.F. Gomes. SIPAMPL v2.1: Semi-Infinite Programming with AMPL. Tech. Rep. ALG/EF/2-2003, Univ. do Minho, Braga, Portugal, Dec. 2004. http://www.norg.uminho.pt/aivaz/.
18. A.I.F. Vaz, E.M.G.P. Fernandes, and M.P.S.F. Gomes. NSIPS v2.1: Nonlinear Semi-Infinite Programming Solver. Tech. Rep. ALG/EF/5-2002, Univ. do Minho, Braga, Portugal, Dec. 2002. http://www.norg.uminho.pt/aivaz/.
19. M.A. Wolfe. Numerical Methods for Unconstrained Optimization, an Introduction. Van Nostrand Reinhold Company, 1978.

Solving Mathematical Programs with Complementarity Constraints with Nonlinear Solvers

Helena Sofia Rodrigues[1] and M. Teresa T. Monteiro[2]

[1] Computational Mathematic Master Student, University of Minho, Portugal
helena.rodrigues@ipb.pt
[2] Production and Systems Department,University of Minho, Portugal
tm@dps.uminho.pt

Summary. MPCC can be solved with specific MPCC codes or in its nonlinear equivalent formulation (NLP) using NLP solvers. Two NLP solvers - NPSOL and the line search filter SQP - are used to solve a collection of test problems in AMPL. Both are based on SQP (Sequential Quadratic Programming) philosophy but the second one uses a line search filter scheme.

1 Introduction

There has been an enormous amount of interest in developing algorithms to solve Mathematical Programs with Complementarity Constraints (MPCC). A series of problems that arise from realistic applications in engineering and economics is the main reason for such interest.

In recent years, the advances in computer technology have increased the capability of a modeler to solve large scale problems with complementarity constraints. On the one hand, the appearance of modelling systems allows to directly express complementarity conditions as part of their syntax and to pass on the complementarity model to the solver. On the other, the ability of a modeler to generate realistic large scale models enables the solvers to be tested on much larger and more difficult classes of models, producing in this way new enhancements and improvements in the solver.

However, solving MPCC is a harder task because it can be shown that constraint qualifications typically assumed to prove convergence of standard NLP algorithms fail for MPCC. As a result, applying specific MPCC solvers is problematic. To circumvent these problems, various reformulations of MPCC have been proposed. One of these approaches involves the possibility of solving MPCC by transforming it to a well-behaved nonlinear program. This endeavor is important because it allows to extend the body of analytical and computational expertise of nonlinear programming to this new class of problems.

In this work it is analyzed the possibility of solving MPCC in its NLP reformulation using certain SQP algorithms. Two NLP solvers based on SQP philosophy are used to solve a set of problems - NPSOL and the line search filter SQP. The last one is a promising recent algorithm developed by our research group still in improvement phase. Another goal of this study is testing this new algorithm for MPCC problems in their NLP reformulation. This study is included in a MSc project.

The organization of this paper is as follows. Section 2 introduces the Mathematical Programs with Complementarity Constraints. The next section defines an equivalent nonlinear program. In Section 4, the NLP solvers (NPSOL and line search filter SQP) are presented as well as their main characteristics. Numerical results obtained with a collection of AMPL test problems are presented in Section 5. Finally, the main conclusions are shown.

2 Mathematical Programs with Complementarity Constraints

A Mathematical Program with Complementarity Constraints is an optimization problem with equality and inequality constraints. In fact, it is a nonlinear optimization problem where the constraints have the same form as the first-order optimality conditions for a constrained optimization problem.

A Mathematical Program with Complementarity Constraint (MPCC) is defined as:

$$\begin{aligned} \min\ & f(z) \\ \text{s.t.}\ & c_E(z) = 0 \\ & c_I(z) \geq 0 \\ & 0 \geq z_1 \perp z_2 \leq 0 \end{aligned} \tag{1}$$

where $z = (z_0, z_1, z_2)$, $z_0 \in \mathbb{R}^n$ is the control variable and $z_1, z_2 \in \mathbb{R}^p$ are the state variables; c_E and c_I are the sets of equality and inequality constraints, respectively. The presence of complementarity constraints is the most prominent feature of a MPCC that distinguishes it from a standard nonlinear optimization problem.

To solve this kind of problems one might be tempted to use a standard nonlinear programming algorithm. Unfortunately, the feasible set of MPCC is ill-posed since the constraints qualifications which are commonly assumed to prove convergence of standard nonlinear programming algorithms do not hold at any feasible point of the complementarity constraints [16].

A number of special purpose algorithms have been developed for MPCCs, such as branch-and-bound, implicit nonsmooth approaches, piecewise sequential programming [4, 14]. Nevertheless, most of the algorithms for solving MPCC need strong assumptions to ensure convergence. Hence, the research on the development of effective algorithms remains vigourous.

This kind of problems gained lot of popularity in the last decade, because the concept of complementarity is synonymous of equilibrium and many real applications can be modelled by MPCC.

In Economics, complementarity is used to express Walsarian and Nash equilibrium, spatial price equilibria, invariant capital stock, game-theoretic models and Stackelberg lender-follower games, used in oligopolistic market analysis [3, 5]. More complex general equilibrium models are used for various aspects of policy design and analysis, including carbon abatement and trade form. Other applications use game theory, where new examples are becoming popular due to deregularization of electricity markets.

Engineering applications of MPCC include contact and structural mechanics, structural design, obstacle and free boundary problems, elastohydrodynamic lubrification and traffic equilibrium. Recently, MPCC has been used in optimal control problems for multiple robot systems [9, 15]. In optimization, this kind of problems involves the formulation of the Karush-Kuhn-Tucker conditions.

3 Nonlinear Programs

An extensive theory of first and second order optimality conditions for MPCC has been developed. However, the numerical analysis of large-scale MPCCs is still an area of investigation. Some recent papers have suggested reformulating the MPCC problem as a standard NLP. The idea behind this approach is to take advantage of certain NLP algorithms features in order to obtain rapid local convergence.

Notice that (1) can be written in the equivalent NLP form:

$$\begin{aligned} \min\ & f(z) \\ \text{s.t.}\ & c_E(z) = 0 \\ & c_I(z) \geq 0 \\ & z_1 \geq 0 \\ & z_2 \geq 0 \\ & z_1^T z_2 \leq 0 \end{aligned} \tag{2}$$

For the success of NLP solvers, Leyffer suggested to replace the usual complementarity condition $z_1^T z_2 = 0$ by the relaxed equivalent condition $z_1^T z_2 \leq 0$. Without this relaxation, several methods cannot converge quadratically near a strongly stationary point.

Unfortunately, the complementarity constraint implies that the KKT conditions are rarely satisfied by MPCC since it can be shown that there always exists a nonlinear abnormal multiplier [17]. Boundedness of the set of KKT multiplier vectors is equivalent to the Mangasarian Fromovitz constraint qualification condition arising in nonlinear programming.

Recall that, for a point z^* and active set $A(z^*) = E \cup \{i \in I \mid c_i(z) = 0\}$, Mangasarian Fromovitz Constraint Qualification (MFCQ) holds if there exists

a vector $w \in \mathbb{R}^n$ such that:

$$\nabla c_i(z^*)^T w > 0, \text{ for all } i \in A(z^*) \cap I$$
$$\nabla c_i(z^*)^T w = 0, \text{ for all } i \in E$$
$$\nabla c_i(z^*), i \in E \text{ are linearly independent}$$

In MPCC formulation all the feasible points are nonregular in the sense that they do not satisfy MFCQ, which is the usual condition for global convergence of a NLP algorithm. Nonregularity implies that the multiplier set is unbounded, that the normal vectors to active constraints are linearly dependent, and that the linearization of the NLP formulation can be inconsistent, arbitrarily close to a stationary point - all arguments against the use of the NLP technique for solving MPCCs.

Recent investigation brings good news: studies concluded that new well-established nonlinear programming solvers with minor modifications present exciting computational results.

4 NLP Solvers

Upon the success of SQP methods for nonlinear programming, the SQP approach has been extended to solve MPCC as well. In this work, it is presented two NLP solvers using SQP algorithms - NPSOL and the line search filter SQP.

4.1 NPSOL

NPSOL was created by Gill, Murray, Saunders and Wright [10]. NPSOL is a Fortran Package designed to solve the nonlinear programming problem: the minimization of a smooth nonlinear function subject to a set of constraints on the variables. The functions should be smooth but not necessarily convex. NPSOL employs a SQP algorithm and is specially effective for nonlinear problems whose functions and gradients are expensive to evaluate. The inner QP subproblem is solved by a LSSOL subroutine. An augmented Lagrangian merit function using a line search scheme promotes convergence from arbitrary starting points. The Hessian matrix of the Lagrangian function is updated with a BFGS quasi-Newton approximation.

4.2 Line Search Filter SQP

The line search filter SQP is a new algorithm for solving NLP problems, developed by Antunes and Monteiro [2] and still in improvement phase. It is based on a SQP algorithm with a filter scheme whose goal is to avoid the need of a merit function. This function requires difficult decisions in order to choose

the penalty parameters and handle other difficulties like nondifferentiability. We now proceed to briefly explain the filter scheme. For simplicity, consider the NLP problem written in the form:

$$\begin{cases} \min & f(x) \\ \text{s.t.} & c(x) \leq 0 \end{cases} \tag{3}$$

where $c : \mathbb{R}^n \to \mathbb{R}^m$ is a function incorporating all the constraints, possibly simple bound constraints.

Problem (1) can be reinterpreted as a problem which consists in minimizing simultaneously the objective function $f(x)$ and the term

$$h(c(x)) := \sum_{j=1}^{m} max\{0, c_j(x)\}$$

representing the sum of the constraints violation. A filter is a list of pairs $(f^{(i)}, h^{(i)})$ such that any pair dominates any other (dominance concept from multicriteria optimization [13]). A pair $(f^{(i)}, h^{(i)})$ obtained on iteration i is said to dominate another pair $(f^{(j)}, h^{(j)})$ if and only if both $f^{(i)} \leq f^{(j)}$ and $h^{(i)} \leq h^{(j)}$.

The line search filter SQP is based on a Fletcher and Leyffer idea [6] presented in 2000. While these authors used a trust region (TR) approach, the line search filter SQP uses a line search strategy to promote global convergence. The inner QP subproblem is solved using LSSOL subroutine from the NPSOL. For more details see [1, 2].

4.3 NPSOL *vs* Line Search Filter SQP

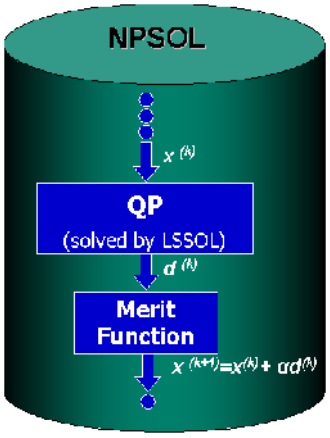


Fig. 1. Scheme of NPSOL

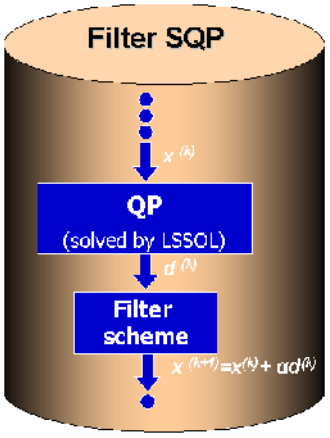


Fig. 2. Scheme of the line search filter SQP

As seen in Figures 1 and 2 for both codes the inner QP subproblem of the SQP algorithm is solved by LSSOL which is a subroutine of the NPSOL. The line search strategy is used also by the two solvers to promote global convergence. The only difference between these codes is the process to obtain the scalar α from the line search - NPSOL uses a merit function and the line search filter SQP consults the filter.

5 Numerical Examples

5.1 MacMPEC

In order to perform some computational experiments using these solvers, a library was used: MacMPEC [11], that is a collection of MPCC models written in the AMPL language [8]. It is a recent library compiled by Sven Leyffer that contains an extensive collection of MPCC problems. Not all the problems in MacMPEC are included in this study. Due to memory limit, it wasn't possible to solve large problems. The numerical tests were done on a Pentium IV 2600Mhz processor with 512Mb Ram in a WindowsXP operating system. More details about the problems and solvers results can be found in Appendix.

5.2 Numerical Results

For all the problems the usual complementarity condition in MPCC formulation (1) was replaced by the equivalent nonlinear condition with relaxation (1). All starting points are standard and fixed by default of AMPL. For both solvers, the stop criterium tolerance was $\epsilon = 1.0E - 06$ or 1000 iterations.

For some problems, the solvers couldn't confirm optimality, because the iteration limit was reached - Table 1 shows, for each solver, the number of the problems where this happened.

Table 1. Failures of NLP solvers

Solver	iter. limit
NPSOL	4
Line search filter SQP	10

Note that the tested problems set contains some problems that are known not to have strongly stationary limit points. For instance, ex9.2.2, and scholtes4 have solutions which are not strongly stationary. Problem gauvin has a global minimum at a point where the lower-level problems fails a constraint violation, so the formulation as MPCC is not appropriate [12]. Both SQP codes are very robust solving MPCC problems in their NLP formulation.

Figure 3 shows the comparison of the CPU time, in seconds. The NPSOL is significantly faster than the line search filter SQP but note that the last one is still in improvement phase.

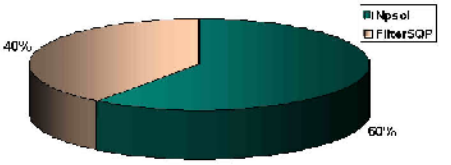


Fig. 3. Percentage of problems with lower CPU time

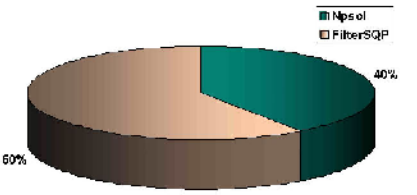


Fig. 4. Percentage of problems with fewer iterations

Figures 4 and 5 show the comparison in terms of iterations. With respect to the number of iterations the line search filter SQP takes advantage when compared with NPSOL - it needs less number of iterations in 60% of problems. It presents also fewer number of iterations used in a general way to solve problems.

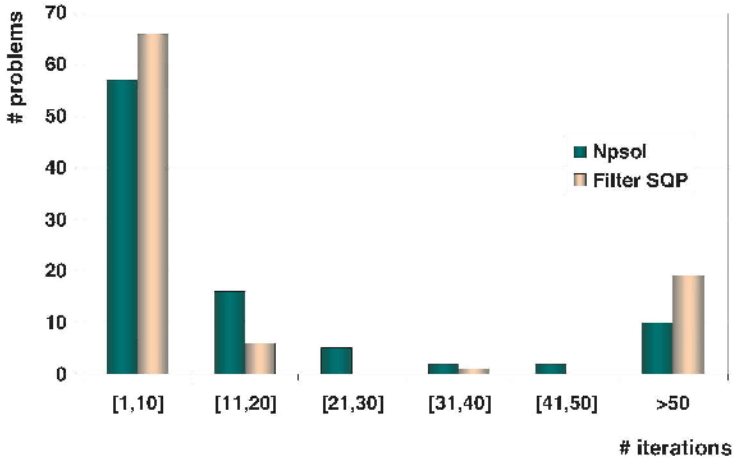


Fig. 5. Number of iterations

Figures 6 and 7 report the ranking of the number of function and gradient evaluations - in 60% of problems the line search filter SQP shows a best behaviour with respect to function and gradient evaluations.

6 Conclusions

A set of MPCC problems were reformulated as NLP problems using the relaxed complementarity condition. Two NLP solvers were tested with these problems and the results confirm their surprising robustness. Using NLP solvers based on SQP algorithms provide an ability to solve a great number of complex problems. The line search filter SQP presents better results than NPSOL with respect to the number of iterations and to the number of function and gradient evaluations. In terms of the CPU time, NPSOL is faster but recall that it is a commercial software already optimized, whereas the line search filter SQP is still in a development phase. The performance of the line search filter SQP, when compared with the NPSOL, is encouraging and this new code should continue to be improved in the future. The research on this area is very important for the modelers, hence it can be a technique for answering important economic and engineering questions.

Appendix - Detailed Numerical Results

Problem	CPU time		Number of iterations		Function Evaluations		Gradient Evaluations	
	Npsol	Filter SQP	Npsol	Filter SQP	Npsol	Filter SQP	Npsol	Filter SQP
bar-truss-3	0.031	0.015	7	9	25	10	25	24
bard1	0	0.031	6	3	8	4	8	4
bard2	0	0	5	2	6	3	6	3
bard3	1	0	1	2	2	3	2	3
bard1m	0	0	6	3	8	4	8	4
bard3m	0	0.015	1	52	2	212	2	56
bilevel1	0	0	3	2	7	3	7	3
bilevel2	0.015	0	5	4	8	5	8	5
bilevel3	0	0.031	13	93	13	673	13	94
bilin	0	0.031	1	104	2	685	2	105
dempe	0	0.015	50	2	173	22	173	3
design-cent-1	0	0.234	29	1000	61	19806	61	1001
design-cent-2	0	0	11	1	14	4	14	2
design-cent-4	0	0	6	2	7	3	7	3
desilva	0	0	2	19	4	27	4	20
df1	0	0	2	2	4	3	4	3
ex9.1.1	0	0	3	1	4	2	4	2
ex9.1.2	0	0.015	2	1	3	2	3	2
ex9.1.3	0	0.015	5	1	6	2	6	2
ex9.1.4	0	0	1	1	5	2	5	2
ex9.1.5	0	0	1	1	2	2	2	2
ex9.1.6	0	0	1	1	1	2	1	2
ex9.1.7	0	0	1	1	2	2	2	2
ex9.1.8	0	0	1	1	4	2	4	2
ex9.1.9	0.015	0	2	1	3	2	3	2
ex9.1.10	0	0	1	1	4	2	4	2
ex9.2.1	0	0	1	1	1	2	1	2
ex9.2.2	0	0.047	25	1	39	2	39	2
ex9.2.3	0	0	1	1	4	2	4	2
ex9.2.4	0	0	5	1	7	2	7	2
ex9.2.5	0.015	0	5	1	9	2	9	2
ex9.2.6	0	0	1	1	1	2	1	2
ex9.2.7	0	0	1	1	1	2	1	2
ex9.2.8	0	0	1	1	1	2	1	2
ex9.2.9	0	0	1	1	1	2	1	2
flp2	0	0	5	2	7	3	7	3
flp4-1	0.05	0.093	1	4	2	19	2	5
flp4-2	0.016	0.156	1	3	2	4	2	4
flp4-3	0.056	1.313	1	5	2	26	2	6
gauvin	0	0	3	8	6	9	6	9
gnash10	0.015	0	10	10	13	17	13	11
gnash11	0	0	10	6	13	7	13	7
gnash12	0	0	9	6	11	7	11	7
gnash13	0	0	10	6	13	7	13	7

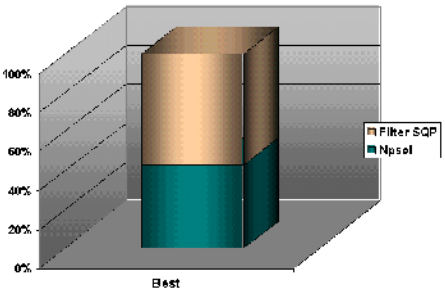


Fig. 6. Ranking of function evaluations

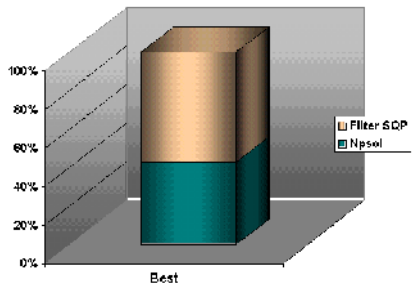


Fig. 7. Ranking of gradient evaluations

Problem	CPU time		Number of iterations		Function Evaluations		Gradient Evaluations	
	Npsol	Filter SQP	Npsol	Filter SQP	Npsol	Filter SQP	Npsol	Filter SQP
gnash14	0	0	10	7	13	8	13	8
gnash15	0	0	10	12	16	16	16	13
gnash16	0	0	10	8	13	12	13	9
gnash17	0	0	10	6	13	7	13	7
gnash18	0	0	14	9	21	14	21	10
gnash19	0	0	10	14	13	15	13	15
hakonsen	0	0.064	54	1	143	3	143	2
hs044-i	0	0.015	9	9	16	11	16	10
incid-set1-8	0.593	0.578	18	5	25	6	25	6
incid-set1c-8	0.562	0.547	14	5	21	6	21	6
incid-set2-8	23.156	151.297	1000	880	2948	16643	2948	881
incid-set2c-8	22.375	107.094	1000	362	2953	1539	2953	363
jr1	0	0	2	2	4	3	4	3
jr2	0	0	9	2	14	3	14	3
kth1	0	0	1	1	2	2	2	2
kth2	0	0	3	3	5	4	5	4
kth3	0	0	5	1	8	2	8	2
liswet1-050	0.14	0.546	6	2	9	3	9	3
nash1	0	0	12	2	16	3	16	3
outrata31	0	0.015	88	14	272	17	272	15
outrata32	0	0	16	8	19	16	19	9
outrata33	0	0	88	12	270	13	270	13
outrata34	0	0	62	36	168	39	168	37
portfl-i-1	0.078	0.171	16	6	17	25	17	7
portfl-i-2	0.046	0.281	12	8	13	38	13	9
portfl-i-3	0.109	0.156	18	6	21	24	21	7
portfl-i-4	0.062	0.328	12	13	14	48	14	14
portfl-i-6	0.062	0.218	13	8	14	22	14	9
qpec-100-1	0.0424	107.747	30	1000	31	8918	31	1001
qpec-100-2	0.609	127.062	24	1000	26	8163	26	1001
qpec-100-3	0.984	219.782	43	1000	44	17355	44	1001
qpec-100-4	0.89	24.125	33	109	34	367	34	110
qpec1	0.031	0.015	63	3	102	4	102	4
qpec2	0.015	0.828	2	1000	4	19830	4	1001
ralph1	0	0.62	31	1000	48	1001	48	1001
ralph2	0	0	20	2	21	3	21	3
ralphmod	3.375	35.985	103	143	148	1451	148	144
scholtes1	0	0	6	4	11	5	11	5
scholtes2	0	0	14	4	34	5	34	5
scholtes3	0	0.046	20	100	32	20001	32	1001
scholtes4	0	0.859	28	1000	41	1680	41	1001
scholtes5	0	0	3	3	4	4	4	4
sl1	0	0.047	4	2	14	3	14	3
stackelberg1	0	0	7	5	8	6	8	6
tap-09	0.062	39.344	6	1000	8	19967	8	1001
tap-15	71.641	330.953	1000	162	3008	3225	3008	163
water-net	0.062	7.828	20	1000	30	19841	30	1001
water-FL	39.953	239.812	1000	1000	2882	19701	2882	1001

References

1. A.S. Antunes and M.T. Monteiro. A SQP-filter algorithm with line search in nonlinear programming. Preprint, 2004.

2. A.S. Antunes and M.T. Monteiro. A filter algorithm and other NLP solvers: performance comparative analysis. Preprint, 2004.
3. S.C. Billups and K.G. Murty. Complementarity problems. JCAM inveted paper, 2000.
4. F. Facchinei, H. Jiang and L. Qi. A smoothing method for mathematical programs with equilibrium constraints. Tech. Rep. AMR 96/15, Univ. of New South Wales, 1996.
5. M.C. Ferris and C. Kanzow. Complementarity and related problems: a survey. In: P.M. Pardalos PM and M.G.C. Resende(eds), Handbook of Applied Optimization. Oxford Univ. Press, New York, 514–530, 2002.
6. R. Fletcher and S. Leyffer. Nonlinear Programming without a penalty function. Math. Programming 91:239–270, 2002.
7. R. Fletcher, S. Leyffer, and P.L. Toint. On the global convergence of a filter-SQP algorithm. SIAM J. Optim., 13(1): 44-59, 2002
8. R. Fourer, D.M. Gay, and B.W. Kernighan. AMPL: A Modelling Language for Mathematical Programming. Duxburg Press, 1993.
9. R. Fourer, M.C. Ferris, and G.M. Gay. Expressing complementary problems and communicating them to solvers. SIAM J. Optim., 9: 991-1009, 1999.
10. P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright. User's guide for NPSOL 5.0: a fortran package for nonlinear programming. Tech. Rep. SOL 86-1 , 1998.
11. S. Leyffer. MacMPEC, webpage: www.mcs.anl.gov/~leyffer/MacMPEC/, 2000.
12. S. Leyffer. Complementarity constraints as nonlinear equations: theory and numerical experience. Tech. Rep. NA/209, Univ. of Dundee, 2002.
13. K.M. Miettinen. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, 1999.
14. Z.Q. Luo, J.S. Pang, and D. Ralph. Mathematical Programs with Equilibrium Constraints, Cambridge University Press, 1996.
15. J. Peng, M. Anitescu, and S. Akella. Optimal control of multiple robot systems with friction using MPCC. Tech. Rep. W-31-109-ENG-38, 2003.
16. H. Pieper. Algorithms for mathematical programs with equilibrium constraints with applications to deregulated electricity markets. Dissertation, Stanford University, 2001.
17. J.J. Ye. Optimality conditions for optimization problems with complementarity constraints. SIAM J. Optim., 9(2):374-387, 1999.

A Filter Algorithm and Other NLP Solvers: Performance Comparative Analysis

António Sanches Antunes[1] and M. Teresa T. Monteiro[2]

[1] University of Minho, Portugal. asanches@ipg.pt
[2] University of Minho, Portugal. tm@dps.uminho.pt

Summary. A new algorithm based on filter SQP with line search to solve nonlinear constrained optimization problems is presented. The filter replaces the merit function avoiding the penalty parameter estimation. This new concept works like an oracle estimating the trial approximation of the iterative SQP algorithm. A collection of AMPL test problems is solved by this new code as well as NPSOL and LOQO solvers. A comparative analysis is made - the filter SQP with line search presents good performance.

1 Introduction

In Fletcher and Leyffer [6] a new technique for globalizing methods for Nonlinear Programming (NLP) is presented. This concept is referred as a NLP filter and it is motivated by the aim of dispensing with the penalty function, avoiding related difficulties like nondifferentiability or the penalty parameter choice. Numerical experiments with this new technique in a sequential quadratic programming (SQP) trust region algorithm are reported in [6] and seem very promising.

In this work the same filter SQP idea is used to promote global convergence but in a line search context instead of the trust region approach.

This paper is divided in 4 sections. The next section presents the mathematical formulation of the NLP problem to solve. Section 3 introduces the concept of a filter and shows how it can be used in a line search based SQP algorithm. An algorithmic refinement that is needed to ensure the robustness of the basic algorithm is presented and the termination criterion is described. The flowchart of the algorithm is also presented. Finally, Section 4 presents numerical results obtained with a collection of AMPL test problems as well as the corresponding comparison with NPSOL and LOQO results and the main conclusions.

2 NLP Problem

The purpose of this work is the development of an algorithm for finding a local solution of an NLP problem of the following form

$$\begin{cases} \min & f(x) \\ \text{s.t.} & lb_x \leq x \leq ub_x \\ & lb_c \leq c(x) \leq ub_c \end{cases} \tag{1}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a nonlinear objective function, $x \in \mathbb{R}^n$, and lb_x, $ub_x \in \mathbb{R}^n$ are the lower and the upper bounds of the variable x, respectively, and $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a set of m general constraints whose lower and upper bounds are lb_c and ub_c, respectively. The set of all constraints defines the feasible region. Note that equality constraints are included in the above formulation by setting the lower bound equal to the upper bound for the relevant constraint. Likewise it is possible to include one-sided constraints by setting the lower bound to $-\infty$ or the upper bound to ∞, depending on which bound is required.

There are several methods to solve problem (1): Sequential Quadratic Programming, reduced-gradient methods, sequential or exact penalty techniques, interior-point methods and more recently the filter method.

3 NLP Filter

3.1 State of Art

Since its first appearance in a 1997 paper by Fletcher and Leyffer [6], the filter technique has been mostly applied to SLP (sequential linear programming) and SQP type methods [6, 5, 7]. Global convergence to first-order critical points was proved for SLP by Fletcher, Leyffer and Toint [5] in 1998 and for SQP by Fletcher, Gould, Leyffer, Toint and Wächter [8] in 1999. The filter idea has proven to be very successful numerically in the SLP/SQP framework [4]. In the context of composite SQP for equality constrained optimization, Ulbrich and Ulbrich [12] have also proposed, based on filter idea, a nonmonotone trust region algorithm. Recently, Audet and Dennis [1] presented a pattern search filter method for derivative-free nonlinear programming. More recently, the filter approach has been adapted to interior-point methods in a number of ways. Benson, Shanno and Vanderbei [2] proposed several heuristics based on the idea of filter methods, for which improved efficiency is reported compared to their previous merit function approach. Ulbrich, Ulbrich and Vicente [13] consider a trust region filter method that bases the acceptance of trial steps on the norm of the optimality conditions. Wächter and Biegler [14] presented a primal-dual interior-point algorithm with a filter line search method for nonlinear programming. Gonzaga, Karas and Vanti in [11] presented an algorithm based on filter method with inexact restoration strategy (IR) and proved its global convergence to stationary points.

3.2 Filter

Penalty functions combine the two competing aims in NLP - minimization of objective function and satisfaction of the constraints - into a single minimization problem. In filter strategies these are seen as separate objectives. Conceptually, these two can be written as

$$\min f(x) \quad \text{and} \quad \min h(c(x))$$

where

$$h(c(x)) := \left\|c^+(x)\right\|_1 := \sum_{j=1}^{m} c_j^+(x)$$

with $c_j^+ = max(0, c_j)$ and the constraints defined in (1) are handled with the form $c(x) \leq 0$.

The filter is a list of pairs $(f^{(l)}, h^{(l)})$ where $f^{(l)} = f(x^{(l)})$ and $h^{(l)}=h(c(x^{(l)}))$ such that no pair dominates any other. A pair $(f^{(k)}, h^{(k)})$ obtained on iteration k is said to dominate another pair $(f^{(l)}, h^{(l)})$ if and only if $h^{(k)} \leq h^{(l)}$ and $f^{(k)} \leq f^{(l)}$, indicating that the point $x^{(k)}$ is at least as good as $x^{(l)}$ for both measures. The filter can be represented graphically in the (f, h) plane as illustrated in Figure 1.

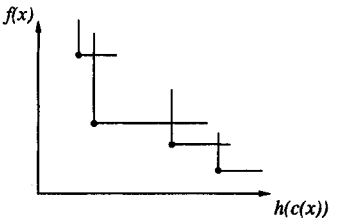


Fig. 1. Example of a filter

The key idea is to use the filter as a criterion for accepting or rejecting a step - it works like an oracle defining a tabu region, *ie*, a points prohibited region. A new step is accepted, which means it doesn't belong to the tabu region, if it reduces either the objective function or the constraints violation function.

3.3 Globalization Techniques

The term "globalization technique" is used to distinguish the method used for selecting the new estimate of the solution from the method for computing the search direction. In almost all algorithms, the formula for the search direction is derived from the Taylor series, which is a "local" approximation to the

function. The method for choosing the new estimate of the solution is designed to guarantee global convergence, defined as convergence from any starting point.

If the underlying optimization method produces good search directions, then the globalization technique will act merely as a safety net protecting against the occasional bad step. For a method that produces less effective search directions, the globalization technique can be a major contributor to the practical success of the method. The two major types of globalization techniques are the trust region method, used by Fletcher and Leyffer in [6] with a filter scheme and the line search method used in this work.

3.4 Algorithm

At iteration k, a quadratic approximation to (1) in $x^{(k)}$ is performed by minimizing the quadratic approximation of the objective function f subject to linear approximation of the constraints. The corresponding QP subproblem is

$$\begin{cases} \min & \frac{1}{2}\, {d^{(k)}}^T W^{(k)} d^{(k)} + {d^{(k)}}^T g^{(k)} \\ \text{s.t} & lb_x \leq x^{(k)} + d^{(k)} \leq ub_x \\ & lb_c \leq {A^{(k)}}^T d^{(k)} + c^{(k)} \leq ub_c \end{cases} \tag{2}$$

where $W^{(k)} = \nabla^2 L(x^{(k)}, \lambda^{(k)})$ is the Hessian of the Lagrangian function $L(x, \lambda) = f(x) + \lambda^T c(x)$, $g^{(k)} = \nabla f(x^{(k)})$ is the gradient of the objective function and ${A^{(k)}}^T = \nabla c(x^{(k)})^T$ is the Jacobian matrix of general constraints $c(x^{(k)})$. The solution of this QP subproblem, which is also an iterative procedure, is the search direction $d^{(k)}$. The next trial point $x^{(k+1)}$, an approximation to the problem solution, is obtained by

$$x^{(k+1)} = x^{(k)} + \alpha\, d^{(k)} \tag{3}$$

where $\alpha \in \mathbb{R}$ is the steplength. This new point $x^{(k+1)}$ is accepted by the filter if the corresponding pair $(f^{(k+1)}, h^{(k+1)})$ is not dominated by any pair in the filter. Otherwise the point is rejected and α is divided by 2 until the point is accepted or α is smaller than the tolerance level.

The condition that a new point is not dominated by any entry in the filter allows for the possibility of an oscillating sequence of points with accumulation points in the (f, h) space. A standard way to avoid this is to require a sufficient reduction. The aim is to produce an envelope below the filter that prevents points arbitrarily close to the filter from being accepted. This idea is illustrated in Figure 2 where the envelope is shown by the dashed line.

A new iterate $x^{(k+1)}$ is said to be acceptable to the filter if its (f, h) pair satisfies either

$$h < \beta h^{(l)} \qquad \text{or} \qquad f < f^{(l)} - \gamma h \tag{4}$$

for all pairs $(f^{(l)}, h^{(l)})$ in the filter, where β and γ are preset parameters such that $0 < \gamma < \beta < 1$, with β close to 1 and γ close to zero [7]

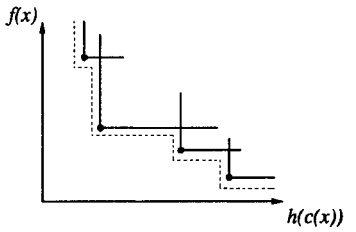


Fig. 2. Envelope created by sufficient reduction conditions

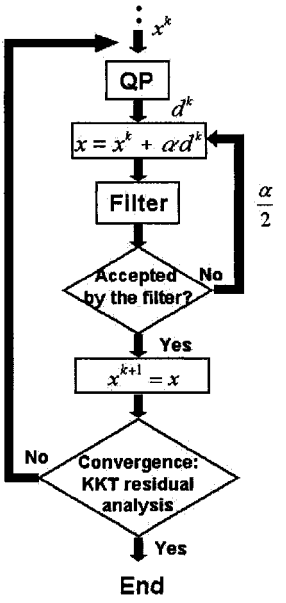


Fig. 3. Filter SQP flowchart

In Fig. 3 we present the corresponding flowchart of the filter SQP algorithm. The algorithm terminates when a Kuhn-Tucker point x^* is found. The solver computes the normalized Kuhn-Tucker residual defined by

$$r = \frac{\|g^* + \nu^* + A^*\lambda^*\|_2}{\max\{\mu_{\max}, 1.0\}}$$

where $g^* = \nabla f(x^*)$, $A^* = \nabla c(x^*)$, ν and λ are the Lagrange multipliers of the variables and constraints, respectively, and the maximum length is calculated

by

$$\mu_{\max} = \max_i \{\|g^*\|_2, |\nu_i|, \|a_i^*\|_2 |\lambda_i^*|\}.$$

The solver terminates if the normalized Kuhn-Tucker residual r is less than ϵ where ϵ is a user provided tolerance level.

ALGORITHM: Filter SQP line search
Given $x^{(0)}$, $\lambda^{(0)}$, set $k = 0$
REPEAT
  Solve QP (26) for a step $d^{(k)}$, set $\alpha = 1$ and set $Accept = FALSE$
  REPEAT
    set $x^{(k+1)} = x^{(k)} + \alpha\, d^{(k)}$
    IF $(f^{(k+1)}, h^{(k+1)})$ is acceptable to the filter **THEN**
      Accept $x^{(k+1)}$ and add $(f^{(k+1)}, h^{(k+1)})$ to the filter
      Remove points dominated by $(f^{(k+1)}, h^{(k+1)})$ from the filter
      set $Accept = TRUE$
    ELSE
      Reject $x^{(k+1)}$
      Set $\alpha = \alpha/2$
    ENDIF
  UNTIL $Accept = TRUE$ or $\alpha \leq TolAlpha$
  set $k = k + 1$
UNTIL Convergence

4 Computational Experiments and Conclusions

This section presents some initial numerical experiments performed on a Pentium III with 128MB RAM. The filter algorithm was implemented in C Language for Windows operating system.

The QP subproblem (26) is solved using the LSSOL routine from NPSOL solver [10]. NPSOL is a set of Fortran subroutines based on SQP algorithm with an augmented Lagrangian merit function with Quasi-Newton approximation to the Hessian of the Lagrangian. The comparison between NPSOL and the new algorithm based on filter method is a fair comparison because both packages use the same philosophy (SQP with line search strategy) and the same subroutine to solve the QP subproblem (LSSOL). The search direction d calculated by the QP subroutine is used to update the next approximation to the solution of the problem using a line search globalization method. The only difference is that NPSOL obtains the scalar α in (3) using an augmented Lagrangian merit function and our algorithm consults an oracle - the filter - to verify the acceptance of the new point $x^{(k+1)}$.

We have also used LOQO, an infeasible primal-dual interior-point algorithm, and compared its results with those of this new method, based on a SQP approach.

Table 1. Numerical results for Filter SQP, NPSOL and LOQO codes

			Filter SQP			NPSOL			LOQO		
Problem	n	m	# f	# g	Iter	# f	# g	Iter	# f	# g	Iter
hs001	2	1	34	25	24	21	21	18	68	68	33
hs003	2	1	2	2	1	8	8	3	21	21	11
hs004	2	2	3	3	2	2	2	1	15	15	8
hs005	2	2	19	13	12	9	9	7	19	19	10
hs007	2	1	9	9	8	15	15	11	27	27	14
hs008	2	2	3	3	2	9	9	6	17	17	9
hs010	2	1	12	12	11	17	17	13	29	29	15
hs011	2	1	6	6	5	12	12	8	25	25	13
hs012	2	1	7	7	6	9	9	8	19	19	10
hs014	2	2	6	6	5	8	8	6	21	21	11
hs015	2	3	5	5	4	5	5	3	61	61	31
hs016	2	4	6	6	5	5	5	4	35	35	18
hs017	2	4	11	10	9	15	15	12	58	58	29
hs018	2	4	8	8	7	23	23	13	29	29	15
hs019	2	4	6	6	5	8	8	6	33	33	17
hs020	2	4	7	7	6	5	5	4	47	47	24
hs021	2	3	2	2	1	4	4	2	23	23	12
hs022	2	2	3	3	2	7	7	5	17	17	9
hs023	2	5	8	8	7	7	7	6	35	35	18
hs024	2	3	3	3	2	6	6	2	25	25	13
hs026	3	1	16	16	15	56	65	44	29	29	15
hs027	3	1	10	9	8	28	28	20	33	33	17
hs028	3	1	2	2	1	6	6	4	17	17	9
hs030	3	4	3	3	2	4	4	2	17	17	9
hs031	3	4	7	7	6	12	12	8	33	33	17
hs032	3	2	3	3	2	3	3	2	46	46	23
hs033	3	3	6	6	5	11	11	5	21	21	11
hs034	3	5	8	8	7	8	8	7	27	27	14
hs035	3	1	2	2	1	7	7	5	19	19	10
hs036	3	4	3	3	2	2	2	1	35	35	16
hs038	4	2	68	39	38	33	33	26	99	99	44
hs039	4	2	122	46	45	15	15	12	29	29	15
hs040	4	3	8	8	7	9	9	6	17	17	9
hs042	4	2	5	5	4	10	10	6	17	17	9
hs043	4	3	7	7	6	17	17	11	21	21	11
hs044	4	6	10	7	6	6	6	4	31	31	16
hs046	5	2	23	19	18	53	53	50	44	44	22
hs047	5	3	26	21	20	57	57	34	44	44	21
hs048	5	2	2	2	1	9	9	5	19	19	10
hs049	5	2	17	17	16	43	43	40	47	47	24
hs050	5	3	10	10	9	19	19	15	31	31	16
hs051	5	3	2	2	1	7	7	4	15	15	8
hs052	5	3	2	2	1	6	6	3	15	15	8
hs053	5	3	2	2	1	7	7	4	21	21	11
hs054	6	1	2	2	1	7	7	5	23	23	12

The code is interfaced to the modelling language AMPL [9] and a set of a hundred AMPL problems is tested to compare the new algorithm with NPSOL and LOQO. After eliminating the problems for which the solvers obtained different local minima, 88 problems remained. The percentage of fails were almost the same for the three codes, for that reason the robustness is similar.

The tolerances used in the algorithm were set to $\gamma = 1.0E-05$, $\beta = 1-\gamma$, $\epsilon = 1.0E-06$ and $TolAlfa = 1.0E-06$.

Tables 1 and 2 report the numerical results of the three codes tested - Filter SQP, NPSOL and LOQO. The tables show the problem name, its dimension (n number of variables and m number of constraints), $\#f$ and $\#g$ are the number of function and gradient evaluations, respectively, and *iter* is the iterations count. Table 3 presents the cumulative results for the three solvers.

Without further examination the results of the filter SQP seem very encouraging, but measuring and comparing software is a very difficult task. Dolan and Moré in [3] present a tool for the evaluation and performance of optimization codes. The performance profile for a solver is the (cumulative)

Table 2. Numerical results for Filter SQP, NPSOL and LOQO codes (cont)

Problem	n	m	Filter SQP # f	# g	Iter	NPSOL # f	# g	Iter	LOQO # f	# g	Iter
hs057	2	3	25	6	5	21	21	19	31	31	16
hs060	3	1	6	6	5	11	11	8	17	17	9
hs062	3	1	7	6	5	16	16	9	25	25	13
hs066	3	5	10	9	8	8	8	7	29	29	15
hs073	4	3	3	3	2	5	5	4	39	39	19
hs076	4	3	2	2	1	8	8	7	21	21	11
hs079	5	3	9	8	7	12	12	10	17	17	9
hs080	5	3	7	7	6	10	10	8	17	17	9
hs083	5	3	5	5	4	7	7	5	25	25	13
hs084	5	3	6	5	4	3	3	2	64	64	21
hs085	5	48	22	22	21	18	18	17	61	61	29
hs086	5	1	4	4	3	7	7	5	25	25	13
hs087	6	4	6	6	5	15	15	12	47	47	24
hs095	6	4	5	3	2	2	2	1	32	32	16
hs097	6	4	7	7	6	6	6	3	36	36	18
hs100	7	4	24	8	7	29	29	14	22	22	11
hs104	8	6	25	24	23	20	20	18	27	27	14
hs105	8	8	17	13	12	69	69	50	35	35	18
hs112	10	3	12	12	11	36	36	16	35	35	17
hs113	10	8	14	7	6	18	18	13	31	31	16
hs116	13	28	10	9	8	20	20	10	194	194	74
hs118	15	8	2	2	1	30	30	14	29	29	15
hs119	16	1	8	8	7	16	16	13	57	57	29
bqpgabim	50	4	2	2	1	98	98	51	27	27	14
braess	5	3	2	2	1	4	4	2	23	23	12
chemeq	126	16	20	20	19	112	112	61	151	151	53
deconvc	52	1	22	17	16	132	132	84	63	63	32
eigminb	101	2	6	6	5	6	6	5	19	19	10
expquad	120	2	2	2	1	3	3	2	19	19	10
hager1	20	2	2	2	1	10	10	7	19	19	10
hydrothermal	58	11	5	5	4	11	11	10	113	113	57
integreq	52	2	4	4	3	11	11	7	15	15	8
liarwhd	36	1	20	20	19	27	27	26	55	55	27
liswet10	103	1	2	2	1	3	3	2	37	37	19
minsurf	64	4	13	5	4	27	27	15	23	23	17
nondquar	100	1	19	19	18	317	317	310	45	45	23
optcntrl	32	5	4	4	3	5	5	4	67	67	34
polak1	2	2	8	8	7	15	15	13	27	27	14
polak3	12	10	9	9	8	29	29	23	47	47	24
polak6	5	4	11	10	9	41	41	18	53	53	27
tf12	3	101	2	2	1	32	32	20	29	29	15
weapon	100	2	10	10	9	64	64	54	45	45	23
zigzag	64	11	8	8	7	65	65	37	85	85	43

Table 3. Cumulative results

	#f	#g	iter
Filter SQP	966	764	673
NPSOL	2078	2078	1514
LOQO	3252	3252	1592

distribution function for a performance metric. Performance profiles provide a means of visualizing the expected performance difference among solvers, while avoiding arbitrary parameter choices and need to discard solver failures from the performance. This comparison is very interesting when the test set has a large number of problems. So, for a solver s, it is plotted

$$\log_2\left(\frac{\#\text{iter}(s,p)}{\text{bestiter}(p)}\right), \ \forall\, p \in \text{problem set}, \tag{5}$$

where $\#\text{iter}(s,p)$ is the number of iterations that solver s tooks on problem p and $\text{bestiter}(p)$ is the smallest number of iteration any solver tooks. Graphi-

cally it is interpreted as the probability distribution that a given solver is at worse x times slower than the best.

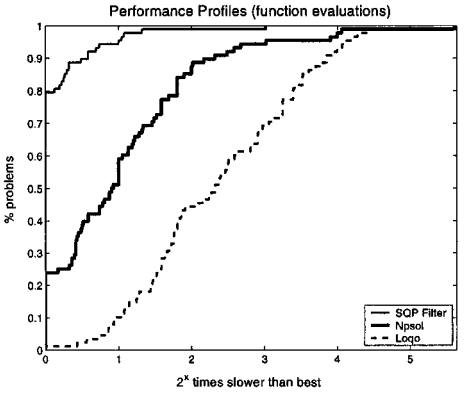


Fig. 4. Function evaluations performance profiles

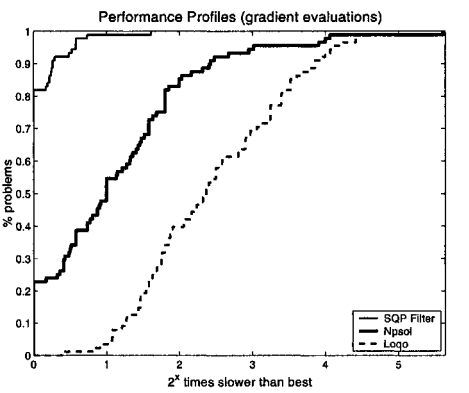


Fig. 5. Gradient evaluations performance profiles

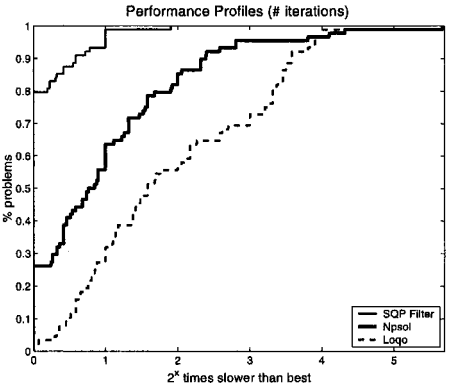


Fig. 6. Number of iterations performance profiles

This tool is used to compare Filter SQP, NPSOL and LOQO. The performance metrics are the function evaluations, gradient evaluations and iteration counts. The graphics of performance profiles are presented in Figures 4, 5 and 6 and a *log* scale is used.

An easy interpretation of these graphics is that for any given measure a solver is the best when its graphic is tending faster to 1. From these figures it is clear that Filter SQP has the highest probability of being the optimal

solver for all measures. This comparison permits an estimation of the behavior of the filter method in a SQP algorithm since this method only replaces the merit function scheme based on an augmented Lagrangian function at the SQP algorithm of the NPSOL.

The main conclusion of this work is that the initial computational results are very much encouraging. The next phase is to study some details related to the algorithm convergence and to test larger dimension problems. This new algorithm - a simple code based on an easy idea - compares favourably with NPSOL and LOQO for all performance measures and presents similar robustness.

References

1. C. Audet and J.E. Dennis. A pattern search filter method for nonlinear programming without derivatives. Tech. Rep. 00-09, Dep. of Computational and Applied Mathematics, Rice University, Houston, 2000.
2. H.Y. Benson, F.F. Shanno, and R.J. Vanderbei. Interior-point methods for nonconvex nonlinear programming: filter methods and merit functions. Computational Optimization and Applications 23(2):257–272, 2002.
3. E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles, Math. Programming, Ser. A 91: 201–213, 2002.
4. R. Fletcher and S. Leyffer. User manual for filter SQP. Numerical Analysis Report NA/181, Univ. of Dundee, 1998.
5. R. Fletcher R, S. Leyffer, and P.L. Toint. On the global convergence of an SLP-filter algorithm. Numerical Analysis Report NA/183, Univ. of Dundee, 1999.
6. R. Fletcher and S. Leyffer. Nonlinear Programming without a penalty function. Math. Programming 91:239–270, 2002.
7. R. Fletcher, S. Leyffer, and P.L. Toint. On the global convergence of a filter-SQP algorithm. SIAM J. Optim., 13(1):44–59, 2002.
8. R. Fletcher R, N.I.M. Gould, S. Leyffer, P.L. Toint PL, and A. Wächter. Global convergence of trust-region SQP-filter algorithms for general nonlinear programming. SIAM J. Optim., 13(3): 635–659, 2002.
9. D.M. Gay. Hooking your solver to AMPL. Tech. Rep. 97-4-06, Computing Sciences Research Center, Bell Laboratories, Murray Hill, 1997.
10. P.E. Gill, W. Murray, M.A. Saunders, and M.H. Wright MH. User's guide for NPSOL 5.0: a fortran package for nonlinear programming. Tech. Rep. , 1998.
11. C.C. Gonzaga CC, E. Karas, and M. Vanti. A globally convergent filter method for nonlinear progamming. SIAM J. Optim., 14(3):646–669, 2003.
12. M. Ulbrich and S. Ulbrich. Nonmonotone trust region methods for nonlinear equality constrained optimization without a penalty function. Math. Programming 95:103–135, 2003.
13. M. Ulbrich, S. Ulbrich, and L.N. Vicente. A globally convergent primal-dual interior-point filter method for nonlinear programming. Math. Programming 100(2):379–410, 2004.
14. A. Wächter and L.T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Tech. Rep. RC 23149, IBM T. J. Watson Research Center, Yorktown - USA, 2004.

How Wastewater Processes can be Optimized Using LOQO

I. A. C. P. Espírito-Santo[1], Edite M. G. P. Fernandes[1], M. M. Araújo[1], and E. C. Ferreira[2]

[1] Systems and Production Department, Minho University, Braga, Portugal.
{iapinho;emgpf;mmaraujo}@dps.uminho.pt
[2] Centre of Biological Engineering, Minho University, Braga, Portugal.
ecferreira@deb.uminho.pt

Summary. This paper describes the optimization of an activated sludge system which comprises an aeration tank and a secondary settler. This system is by far the most widely used biological process in wastewater treatment plants. The optimization process is represented as a smooth nonlinear problem with highly nonlinear equality constraints, some linear equality constraints, one inequality constraint and simple bounds, in which the objective is to minimize the total cost associated with the installation and operation of the biological process. We use the software LOQO to solve the problem. Several computational results show that the quality of the effluent, especially in terms of carbonaceous matter, influences directly the cost and the main contribution to the total cost is the air flow, due to the acquirement of the electromechanical equipment and spent energy.

1 Introduction

The optimization of natural processes is becoming nowadays more and more important. In the case of wastewater treatment plants (WWTPs), besides the simulation, which is quite common, see for example [9], it seems crucial to reduce, as much as possible, the costs associated with the design and operation of such plants, because they are so high that can threaten the very survival of many industries.

Besides the densely populated and industrial regions, it is also very important to treat the domestic effluents in small country regions. In particular, there is a small region in the north of Portugal, Trás-os-Montes, that produces high quality wines and has significant effluent variations in terms of amount of pollution and flow, during the vintage season.

A typical WWTP is schematically represented in Figure 1. The first three unit processes define the primary treatment which is a physical process and aims to eliminate the gross solids and grease, so avoiding the blocking up of the secondary treatment. Although the dimensioning of such units is usually

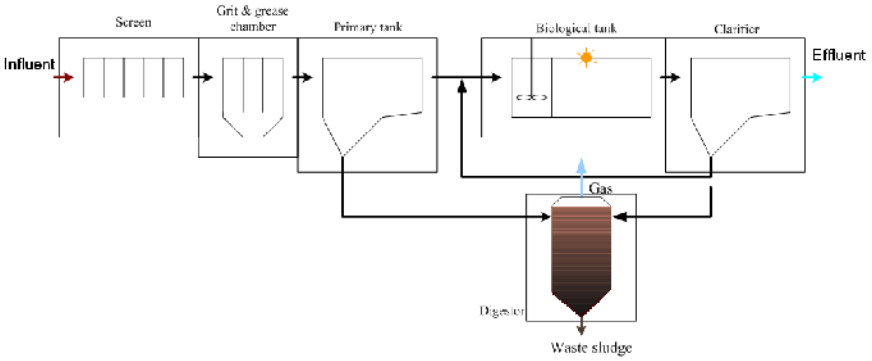


Fig. 1. Schematic representation of a typical WWTP (Adapted from [10]).

empirical and based on the wastewater to be treated, its cost does not depend too much on the characteristics of the wastewater. This is the reason why this process is not included in the optimization procedure.

The next two units define the secondary treatment of the wastewater. It is the most important treatment in the plant because it eliminates the soluble pollutants. This is a biological process which, in the case herein studied, comprises an aeration tank and a clarifier and aims to separate the biological sludge from the treated water. There are other biological treatments but this is, by far, the most widely used.

Finally, the last unit is used to treat the biological sludge that is wasted by the secondary settler.

There are many other possible WWTP layouts, but most of them have the above described treatments. When the wastewater is very polluted and the secondary treatment does not provide the demanded quality, a tertiary treatment, usually a chemical process, can be included.

This paper is part of an ongoing research project in which we are engaged to optimize the design and the operation of the whole plant in terms of minimum total cost (investment and operation costs). The work herein presented focus solely on the secondary treatment, in particular on an activated sludge system that is represented in Figure 2. This system consists of an aeration tank and a secondary settler. The influent enters the aeration tank where the biological reactions take place, in order to remove the dissolved carbonaceous matter and nitrogen. The sludge that leaves this tank enters in the secondary settler where suspended solids are removed. After this treatment, the treated final effluent leaves the settling tank and the thickened sludge is recycled to the aeration tank and part of it is wasted.

For the case presented in this paper, the predominant costs related to the installation of a plant are concerned with the civil construction of the tanks and the acquirement of electromechanical equipment. The predominant operation costs are due to the power required to the aeration of the activated sludge. The mathematical models used to describe the aeration tank and the settling tank are the ASM1 model [5] and the ATV model [2], respectively.

This paper is organized as follows. In Section 2, we present the developed mathematical formulation of the problem. Section 3 contains details of the implementation of the resulting optimization problem and Sections 4 and 5 contain a brief discussion of the results and the conclusions.

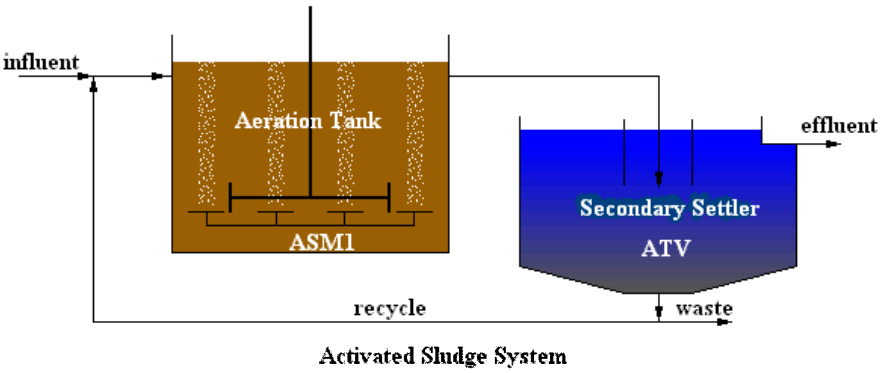


Fig. 2. Schematic representation of the activated sludge system (Adapted from the GPS-X simulator [6]).

2 The Mathematical Model

The mathematical model can be subdivided in seven types of equations, as it will be described. The system under study consists of an aeration tank, where the biological reactions take place, and a secondary settler for the sedimentation of the sludge and clarification of the effluent.

To describe the aeration tank we chose the activated sludge model n.1, described by Henze *et al.* [5], which considers both the elimination of the carbonaceous matter and the removal of the nitrogen compounds. This model is widely accepted by the scientific community, as it produces good predictive values by simulations. This means that all state variables keep their biological interpretation. The tank is considered a completely stirred tank reactor (CSTR) in steady state.

For the settling tank the ATV design procedure [2] is used, which is a very simple model but describes the settling process very well, besides considering also peak flow events.

2.1 Mass Balances around the Aeration Tank

The first step in this unit process is to do mass balances around it, using the Peterson matrix of the ASM1 model [5].

The generic equation for a mass balance around a certain system is

$$\text{In - Out + Reaction = Accumulation.}$$

In mathematical language, for a CSTR

$$\frac{Q}{V_a}(\xi_{in} - \xi) + r_\xi = \frac{d\xi}{dt},$$

where Q is the flow that enters the tank, V_a is the aeration tank volume, ξ and ξ_{in} are the concentrations of the component around which the mass balances are being made inside the reactor and on entry, respectively. It is convenient to refer that in a CSTR the concentration of a compound is the same at any point inside the reactor and at the effluent of that reactor. The reaction term for the compound in question, r_ξ, is obtained by the sum of the product of the stoichiometric coefficients, $\nu_{\xi j}$, with the expression of the process reaction rate, ρ_j, of the ASM1 Peterson matrix [5]

$$r_\xi = \sum_j \nu_{\xi j}\rho_j.$$

In steady state, the accumulation term given by $\frac{d\xi}{dt}$ is zero, because the concentration is constant in time. A WWTP in labor for a sufficiently long period of time without significant variations can be considered at steady state. As our purpose is to make cost predictions in a long term basis it is reasonable to do so.

The ASM1 model involves 8 processes incorporating 13 different components. The mass balances for the inert materials, S_I and X_I, are not considered because they are transport-only components. All the symbols used in these formulae and throughout the paper are listed in the Appendix - Notation. The processes rates are the following:

Aerobic growth of heterotrophs, ρ_1

$$\rho_1 = \mu_H \left(\frac{S_S}{K_S+S_S}\right)\left(\frac{S_O}{K_{OH}+S_O}\right) X_{BH};$$

Anoxic growth of heterotrophs, ρ_2

$$\rho_2 = \mu_H \left(\frac{S_S}{K_S+S_S}\right)\left(\frac{K_{OH}}{K_{OH}+S_O}\right)\left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\eta_g X_{BH};$$

Aerobic growth of autotrophs, ρ_3

$$\rho_3 = \mu_A \left(\frac{S_{NH}}{K_{NH}+S_{NH}}\right)\left(\frac{S_O}{K_{OA}+S_O}\right) X_{BA};$$

Decay of heterotrophs, ρ_4

$$\rho_4 = b_H X_{BH};$$

Decay of autotrophs, ρ_5

$$\rho_5 = b_A X_{BA};$$

Ammonification of soluble organic nitrogen, ρ_6

$$\rho_6 = k_a S_{ND} X_{BH};$$

Hydrolysis of entrapped organics, ρ_7

$$\rho_7 = k_h \frac{\frac{X_S}{X_{BH}}}{K_X+\frac{X_S}{X_{BH}}}\left[\left(\frac{S_O}{K_{OH}+S_O}\right)+\eta_h\left(\frac{K_{OH}}{K_{OH}+S_O}\right)\left(\frac{S_{NO}}{K_{NO}+S_{NO}}\right)\right] X_{BH};$$

Hydrolysis of entrapped organic nitrogen, ρ_8

$$\rho_8 = \rho_7 \frac{X_{ND}}{X_S}.$$

The unit adopted for concentration is $g\ COD/m^3$ and the equations obtained from the ASM1 model with mass balances are as follows:

Soluble substrate (S_S)

$$-\frac{1}{Y_H}\rho_1 - \frac{1}{Y_H}\rho_2 + \rho_7 + \frac{Q}{V_a}(S_{S_{in}} - S_S) = 0; \tag{1}$$

Slowly biodegradable substrate (X_S)

$$(1-f_P)\rho_4 + (1-f_P)\rho_5 - \rho_7 + \frac{Q}{V_a}(X_{S_{in}} - X_S) = 0; \tag{2}$$

Heterotrophic active biomass (X_{BH})

$$\rho_1 + \rho_2 - \rho_4 + \frac{Q}{V_a}(X_{BH_{in}} - X_{BH}) = 0; \tag{3}$$

Autotrophic active biomass (X_{BA})

$$\rho_3 - \rho_5 + \frac{Q}{V_a}(X_{BA_{in}} - X_{BA}) = 0; \tag{4}$$

Particulate products arising from biomass decay (X_P)

$$f_P\rho_4 + f_P\rho_5 + \frac{Q}{V_a}(X_{P_{in}} - X_P) = 0; \tag{5}$$

Nitrate and nitrite nitrogen (S_{NO})

$$-\frac{1-Y_H}{2.86Y_H}\rho_2 + \frac{1}{Y_A}\rho_3 + \frac{Q}{V_a}(S_{NO_{in}} - S_{NO}) = 0; \tag{6}$$

$NH_4^+ + NH_3$ nitrogen (S_{NH})

$$-i_{X_B}\rho_1 - i_{X_B}\rho_2 - \left(i_{X_B} + \frac{1}{Y_A}\right)\rho_3 + \rho_6 + \frac{Q}{V_a}(S_{NH_{in}} - S_{NH}) = 0; \quad (7)$$

Soluble biodegradable organic nitrogen (S_{ND})

$$-\rho_6 + \rho_8 + \frac{Q}{V_a}(S_{ND_{in}} - S_{ND}) = 0; \quad (8)$$

Particulate biodegradable organic nitrogen (X_{ND})

$$(i_{X_B} - f_P i_{X_P})\rho_4 + (i_{X_B} - f_P i_{X_P})\rho_5 - \rho_8 + \frac{Q}{V_a}(X_{ND_{in}} - X_{ND}) = 0; \quad (9)$$

Alkalinity (S_{alk})

$$-\frac{i_{X_B}}{14}\rho_1 + \left(\frac{1 - Y_H}{14 \times 2.86 Y_H} - \frac{i_{X_B}}{14}\right)\rho_2 - \left(\frac{i_{X_B}}{14} + \frac{1}{7Y_A}\right)\rho_3 + \frac{1}{14}\rho_6 + \frac{Q}{V_a}(S_{alk_{in}} - S_{alk}) = 0; \quad (10)$$

Oxygen (S_O)

$$K_L a\,(S_{O_{sat}} - S_O) - \frac{1 - Y_H}{Y_H}\rho_1 - \frac{4.57 - Y_A}{Y_A}\rho_3 + \frac{Q}{V_a}(S_{O_{in}} - S_O) = 0. \quad (11)$$

For oxygen mass transfer, the aeration by diffusion is considered:

$$K_L a = \frac{\alpha\, G_S\, \eta\, P_{O_2}\, 1333.3}{V_a S_{O_{sat}}}\theta^{(T-20)} \quad (12)$$

where

$$S_{O_{sat}} = \frac{1777.8\beta\rho P_{O_2}}{HenryO_2},$$

$$\rho = 999.96(2.29 \times 10^{-2}T) - (5.44 \times 10^{-3}T^2), \quad HenryO_2 = 708\ T + 25700.$$

2.2 Composite Variables

In a real system, some state variables are, most of the time, not available from direct measurements. Thus, readily measured composite variables are used instead. They are defined as follows.

Particulate COD

$$X = X_I + X_S + X_{BH} + X_{BA} + X_P; \quad (13)$$

Soluble COD

$$S = S_I + S_S; \tag{14}$$

Total COD

$$COD = X + S; \tag{15}$$

Volatile suspended solids

$$VSS = \frac{X}{icv}; \tag{16}$$

Total suspended solids

$$TSS = VSS + ISS; \tag{17}$$

Biochemical oxygen demand

$$BOD = f_{BOD}\,(S_S + X_S + X_{BH} + X_{BA})\,; \tag{18}$$

Total nitrogen of Kjeldahl

$$TKN = S_{NH} + S_{ND} + X_{ND} + i_{X_B}\,(X_{BH} + X_{BA}) + i_{X_P}\,(X_P + X_I)\,; \tag{19}$$

Total nitrogen

$$N = TKN + S_{NO}. \tag{20}$$

2.3 Quality Constraints

Quality constraints are usually derived from environmental law restrictions. The most used are related with limits in the chemical oxygen demand (COD), total nitrogen (N), and total solids (TSS) at the effluent. In mathematical terms, these constraints are defined as:

$$COD_{ef} \leq COD_{law} \tag{21}$$

$$N_{ef} \leq N_{law} \tag{22}$$

$$TSS_{ef} \leq TSS_{law}. \tag{23}$$

2.4 Constraints of the Secondary Settler

Traditionally the secondary settler is underestimated when compared with the aeration tank. However, it plays a crucial role in the activated sludge system.

When the wastewater leaves the aeration tank, where the biological treatment took place, the treated water should be separated from the biological sludge, otherwise, the COD would be higher than it is at the entry of the system. The most common way of achieving this purpose is by sedimentation in tanks.

A good settler tank has to accomplish three different functions. As a thickener, it aims to produce a continuous underflow of thickened sludge to return to the aeration tank; as a clarifier, it produces a good quality final effluent

and as a storage tank it allows the conservation of the sludge in peak flow events. None of these functions could fail. If that happens the effluent will be of poor quality and the overall behavior of the system can be compromised.

The behavior of a settling tank depends on its design and operation, namely the hydraulic features, as the flow rate, the physical features, as inlet and sludge collection arrangements, site conditions, as temperature and wind, and sludge characteristics. The factors that most influence the size of the tank are the wastewater flow and the characteristics of the sludge. As the former is known, the optimization of the sedimentation area and depth must rely on the sludge characteristics, which in turn are related with the performance of the aeration tank. So, the operation of the biological reactor influences directly the performance of the settling tank and for that reason, one should never be considered without the other.

The ATV design procedure contemplates the peak wet weather flow (PWWF) events, in which the sludge mass transferred from the biological reactor is $\Delta X V_a$, where ΔX is the change in the sludge concentration within the aeration tank. A reduction of 30% on the sludge concentration for a PWWF event is considered. A higher reduction of the sludge concentration into the biological reactor may compromise the entire process.

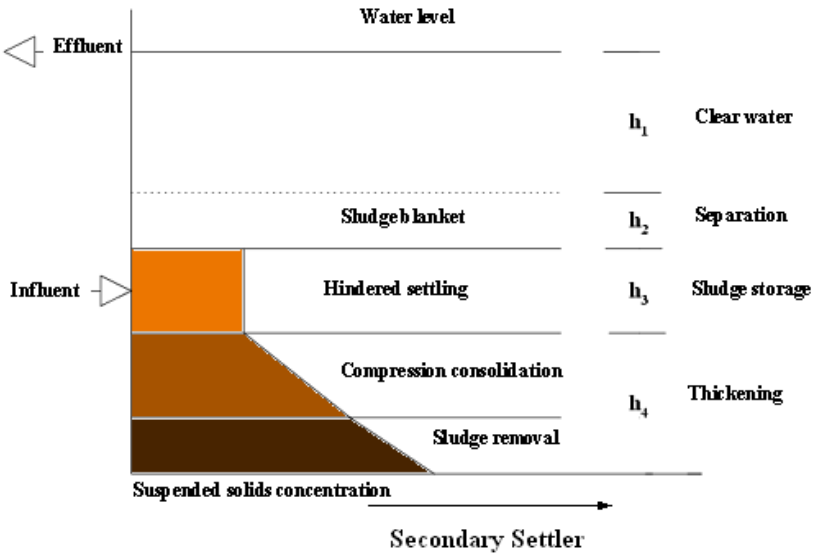


Fig. 3. Typical solids concentration-depth profile adopted by the ATV design procedure (Adapted from [2]).

A way of turning around this problem is to allocate a certain depth (h_3 from Figure 3) to support the fluctuation of solids during these events. Thus, the sludge storage depth depends on the mass that needs to be stored during a PWWF and is given by

$$h_3 = \Delta X V_a \frac{DSVI}{480 A_s}. \tag{24}$$

When this zone is considered, a reduction in the sedimentation area is allowed.

The transferred sludge causes the biological sludge concentration in the reactor at PWWF to decline, which allows a higher overflow rate and therefore a smaller surface area. However, the greater the decrease in reactor concentration is, the greater is the mass of sludge to be stored in the settler tank, so the deeper the tank needs to be. The ATV procedure allows a trade-off between surface area and depth and one may select the area/depth combination that suites the particular site under consideration. As the work herein presented aims to reduce costs, both the surface area and depth are considered as variables and the combination goals for the minimum cost.

The compaction zone, h_4, where the sludge is thickened in order to achieve the convenient concentration to return to the biological reactor, depends only on the characteristics of the sludge, and is given by

$$h_4 = X_p \frac{DSVI}{1000} \tag{25}$$

where X_p is the sludge concentration in the biological reactor during a PWWF event.

The clear water zone, h_1, and the separation zone, h_2, are set empirically, in our case to $0.5m$. The depth of the settling tank, h, is the sum of these four zones, and should not be less than $2m$.

The sedimentation area is still related to the peak flow, Q_p, by the expression

$$\frac{Q_p}{A_s} \leq 2400 \left(X_p DSVI\right)^{-1.34} \quad (m/h). \tag{26}$$

2.5 Flow and Mass Balances around the System

The system behavior, in terms of concentration and flows, may be predicted by balances. In order to achieve a consistent system, these balances must be done around the entire system and not only around each unitary process. They were done to the suspended matter, dissolved matter and flows.

In the case of the suspended matter, the mass balances concern the organic (X) and inorganic (X_{II}) solids :

$$(1+r)\, Q_{inf} X_{in} = Q_{inf} X_{inf} + (1+r)\, Q_{inf} X - \frac{V_a X}{SRT X_r} \left(X_r - X_{ef}\right) - Q_{inf} X_{ef} \tag{27}$$

$$Q_{inf}0.2TSS_{inf} = \frac{V_a X_{II}}{SRTX_r}\left(X_{II_r} - X_{II_{ef}}\right) + Q_{inf}X_{II_{ef}}. \tag{28}$$

The balances of the dissolved matter are done for each one of the dissolved components:

$$(1+r)\,Q_{inf}S_{S_{in}} = Q_{inf}S_{S_{inf}} + rQ_{inf}S_{S_r} \tag{29}$$

$$(1+r)\,Q_{inf}S_{O_{in}} = Q_{inf}S_{O_{inf}} + rQ_{inf}S_{O_r} \tag{30}$$

$$(1+r)\,Q_{inf}S_{NO_{in}} = Q_{inf}S_{NO_{inf}} + rQ_{inf}S_{NO_r} \tag{31}$$

$$(1+r)\,Q_{inf}S_{NH_{in}} = Q_{inf}S_{NH_{inf}} + rQ_{inf}S_{NH_r} \tag{32}$$

$$(1+r)\,Q_{inf}S_{ND_{in}} = Q_{inf}S_{ND_{inf}} + rQ_{inf}S_{ND_r} \tag{33}$$

$$(1+r)\,Q_{inf}S_{alk_{in}} = Q_{inf}S_{alk_{inf}} + rQ_{inf}S_{alk_r}. \tag{34}$$

Besides the mass balances, flow balances are also necessary:

$$Q = Q_{inf} + Q_r \tag{35}$$

$$Q = Q_{ef} + Q_r + Q_w. \tag{36}$$

2.6 Simple Bounds

All variables must be nonnegative, although more restricted bounds are imposed to some of them due to operational consistencies, namely:

$$\begin{array}{cc} 0 \le K_L a \le 300 & 0.05 \le HRT \le 2 \\ 800 \le TSS \le 6000 & 0.5 \le r \le 2 \\ 2500 \le TSS_r \le 10000 & 6 \le S_{alk} \le 8 \\ 6 \le S_{alk_{in}} \le 8 & S_O \ge 2 \end{array} \tag{37}$$

2.7 System Variables Definition

To complete the model, some definitions are added:

Sludge retention time

$$SRT = \frac{V_a X}{Q_w X_r}; \tag{38}$$

Hydraulic retention time

$$HRT = \frac{V_a}{Q}; \tag{39}$$

Recycle rate

$$r = \frac{Q_r}{Q_{inf}}; \tag{40}$$

$$r = \frac{TSS}{TSS_{r_{max}} - TSS}; \tag{41}$$

Recycle rate in a PWWF event

$$r_p = \frac{0.7TSS}{TSS_{max_p} - 0.7TSS}; \tag{42}$$

Recycle flow rate during a PWWF event

$$Q_{r_p} = r_p Q_p; \tag{43}$$

Maximum overflow rate

$$\frac{Q_p}{A_s} \leq 2. \tag{44}$$

A fixed value for the relation between volatile and total suspended solids was considered

$$\frac{VSS}{TSS} = 0.7. \tag{45}$$

2.8 The Objective Cost Function

The cost function represents the total cost and includes both investment and operation costs. In this paper, for the sake of simplicity, no pumps were considered, which means that all the flows in the system move by the effect of gravity.

The operation cost is usually on annual basis, so it has to be updated to a present value with the updating term Γ:

$$\Gamma = \sum_{j=1}^{n} \frac{1}{(1+i)^j} = \frac{1-(1+i)^{-n}}{i}, \tag{46}$$

where i is the discount rate and n is the life span of the WWTP. We use $i = 0.05$ and $n = 20$ years. The total cost is given by the sum of the investment (IC) and operation (OC) costs:

$$TC = IC + OC.$$

To obtain a cost function based on portuguese real data, a study was carried out with a WWTP building company. The basic structure of the model is $C = aZ^b$ [7], where a and b are the parameters to be estimated and Z is the characteristic of the unit process that most influences the cost. This model is nonlinear in the parameters, but it can be easily linearized. The obtained linear model is

$$\ln C = \ln a + b \ln Z$$

and the parameters $\ln a$ and b are estimated by the least squares technique.

The real data collected from the portuguese company are presented in Tables 1 and 2. The investment cost function obtained for the aeration tank is

Table 1. Real data obtained for the aeration tank

V_a	G_S	$HP(KW.h)$	$Cost(euro)$	
(m^3)	$(m^3/day\ STP)$	(annual basis)	Civil construction	Electromechanical
400	437	27720	87158	8416
600	597	129600	159796	10029
1000	943	95040	189952	12199
1300	1283	162000	366146	17056

Table 2. Real data obtained for the secondary settler

A_s	h	V_s	$Cost$
(m^2)	(m)	(m^3)	(euro)
51.8	3.55	184	39355
71.4	3.9	278	61348
100	4.55	455	103741

$$IC_a = 148.6V_a^{1.07} + 7737G_S^{0.62}.$$

The collected data come from a set of WWTPs in design, thus operation data are not available. However, from the company experience, the maintenance expenses for the civil construction are around 1% of the investment costs during the first 10 years and around 2% otherwise. For the electromechanical components, the maintenance expenses are negligible, but all the materials are usually replaced after 10 years. The energy cost is directly related with the air flow. The second value in Table 1 was not used in our estimation because in this plant the planned air flow was in excess for some reason. The power cost (P_c) in Portugal is 0.08 $euro/KW.h$. With this information and with the updating term Γ (46), the operation cost of the aeration tank is then

$$OC_a = \left[0.01\Gamma + 0.02\Gamma(1+i)^{-10}\right]\left(148.6V_a^{1.07}\right) + (1+i)^{-10}7737G_S^{0.62} + 115.1\Gamma P_c G_S.$$

The term $(1+i)^{-10}$ is used to bring to present a future value, in this case, 10 years from now.

The settling tank was considered to operate only by gravity. Thus, with the data from Table 2 and adopting the same mathematical procedure, the correspondent investment cost function is

$$IC_s = 955.5A_s^{0.97}$$

and for the operation cost function, that only concerns the maintenance for the civil construction,

$$OC_s = \left[0.01\Gamma + 0.02\Gamma\,(1+i)^{-10}\right]\left(148.6(A_s\,h)^{1.07}\right).$$

The total cost function is given by the sum of all the functions previously presented:

$$TC = 174.2V_a^{1.07} + 12487G_S^{0.62} + 114.8G_S + 955.5A_s^{0.97} + 41.3\left[A_s\,(1+h_3+h_4)\right]^{1.07} \tag{47}$$

3 Numerical Resolution

The problem of the optimal design and operation of the activated sludge system consists of finding the volume of the aeration tank, the air flow needed for the aeration tank, the sedimentation area, the secondary settler depth, the recycle rate, the effluent flow and concentration of total suspended solids, carbonaceous matter and total nitrogen in the treated water, to name a few, in such a way that, verifying the aeration tank balances (1-12) and system balances (27-36), satisfy the composite variables constraints (13-20), the secondary settler constraints (24-26), the system variables constraints (38-45), the quality constraints (21-23) and the bounds (37), and minimize the cost function (47). This problem has 57 parameters, 82 variables and 64 constraints, where 28 are nonlinear equalities, 1 is a nonlinear inequality and 35 are linear equalities. Seventy one variables are bounded below and eleven are bounded below and above. Table 3 lists all the variables involved, as well as the initial values supplied to the solver.

In Tables 4 to 7, we summarize the data parameter chosen to obtain the results of Section 4. These values are as the default values presented in the GPS-X simulator [6], and are usually found in the real activated sludge based plants for domestic effluents. The remaining parameters are defined as functions of the already listed parameters.

This problem was coded in the AMPL [3] and the results were obtained running the solver LOQO V6.06 (http://www.princeton.edu/~rvdb/loqo/), an infeasible primal-dual interior point method [8]. In the stopping rule, we considered primal and dual infeasibilities $\leq 10^{-5}$ and 2 digits of agreement between the primal and dual objective functions. All the other adjustable LOQO parameters were left as default.

4 Discussion

Several experiments were done for different values of the required COD at the effluent, considering 15 g/m^3 as the N concentration limit and 35 g/m^3 as the TSS concentration limit, at the effluent. The COD limits vary from 45, that is the lowest value with which convergence was obtained, till 140. The value imposed by the portuguese environmental law is 125. The problem

Table 3. Variables of the problem and their initial values

Variable	init. val.	Variable	init. val.	Variable	init. val.	Variable	init. val.
Q	4000	$X_{BH_{in}}$	0	$X_{ND_{in}}$	0	VSS_r	0
Q_w	100	X_{BH}	350	X_{ND}	20	VSS_{ef}	0
Q_r	2000	X_{BH_r}	711	X_{ND_r}	0	TSS_{in}	0
Q_{ef}	1900	$X_{BH_{ef}}$	0	$X_{ND_{ef}}$	0	TSS	1800
Q_{r_p}	0	$X_{S_{in}}$	0	K_La	100	TSS_r	5000
r	1	X_S	350	G_S	10000	TSS_{ef}	10
V_a	1000	X_{S_r}	807	X_{in}	0	X_{II}	0
A_s	1000	$X_{S_{ef}}$	0	X	1000	$X_{II_{ef}}$	0
h_3	0	$X_{BA_{in}}$	0	X_r	4440	X_{II_r}	0
h_4	0	X_{BA}	10^{-6}	X_{ef}	0	BOD_{in}	0
r_p	0	X_{BA_r}	2×10^{-6}	S_{in}	0	BOD	500
X_I	727	$X_{BA_{ef}}$	0	S	50	BOD_r	0
X_{I_r}	950	$S_{NH_{in}}$	0	$S_{alk_{in}}$	0	BOD_{ef}	0
$X_{I_{ef}}$	0	S_{NH}	7.5	S_{alk}	7	TKN_{in}	0
$S_{S_{in}}$	0	$X_{P_{in}}$	0	COD_{in}	0	TKN	106
S_S	10	X_P	90	COD	1600	TKN_r	0
S_O	2	X_{P_r}	175	COD_r	0	TKN_{ef}	0
$S_{O_{in}}$	0	$X_{P_{ef}}$	0	COD_{ef}	0	N_{in}	0
$S_{NO_{in}}$	0	$S_{ND_{in}}$	0	VSS_{in}	0	N_r	0
S_{NO}	10^{-6}	S_{ND}	0.5	VSS	1050	N_{ef}	0
HRT	3.5	N	106				

Table 4. Stoichiometric parameters

Parameter	Value	Parameter	Value
Y_A	0.24	i_{X_B}	0.086
Y_H	0.666	i_{X_P}	0.06
f_P	0.08		

does not converge for smaller values of COD_{law} because there is a minimum under which it is not possible to treat the effluent more, due to the inert

Table 5. Kinetic parameters

Parameter	Value	Parameter	Value
μ_H	6	k_h	3
K_S	20	K_X	0.03
K_{OH}	0.2	μ_A	0.8
K_{NO}	0.5	K_{NH}	1
b_H	0.62	b_A	0.04
η_g	0.8	K_{OA}	0.4
η_h	0.4	k_a	0.08

Table 6. Operational parameters

Parameter	Value	Parameter	Value
T	20	α	0.8
P_{O_2}	0.21	η	0.07
$DSVI$	150	Q_p	150
SRT	20	β	0.95
θ	1.024		

Table 7. Characteristics of the influent to the system

Parameter	Value	Parameter	Value
Q_{inf}	2000	$X_{I,inf}$	73.65
$S_{I,inf}$	30	$X_{S,inf}$	123
$X_{BH,inf}$	0	$S_{NH,inf}$	11.7
$X_{BA,inf}$	0	$S_{ND,inf}$	0.63
$X_{P,inf}$	0	$X_{ND,inf}$	1.251
$S_{O,inf}$	0	$X_{II,inf}$	59.6
$S_{NO,inf}$	0	X_{inf}	196.7
$S_{alk,inf}$	7	S_{inf}	82.73
$S_{S,inf}$	52.73		

contribution (S_I), which is not possible to eliminate with a biological process, from the total COD.

The solution values of the most important variables, such as the total suspended solids and total nitrogen at the effluent, the aeration tank volume, the sedimentation area and depth of the secondary settler, the demanded air flow, the recycle rate, the effluent flow and the total cost are reported in Table 8, for different values of the imposed COD limit (COD_{law}). The number of iterations used by LOQO to converge to the solution, according to the previously defined stopping rule, is shown in the last column of the table.

For an easier interpretation of the results, two graphics were constructed. Figure 4 maps the total cost and the value of the quality index (QI) [1], which defines the amount of pollution at the final effluent, as function of the imposed COD limit. Figure 5 compares the contributions of investment/operation costs and aeration tank/secondary settler costs.

As it can be observed from Table 8 and Figure 4, the total cost decreases and the quality of the effluent deteriorates as the imposed COD at the effluent increases. In terms of total cost, the reduction is more pronounced for COD limits between 45 and 85. For $COD_{law} = 85$ and over the observed cost reduction is very small. For example, when $COD_{law} = 85$ the attained minimum cost is 1.4 millions of euros whereas for $COD_{law} = 125$ the minimum cost decreases to 1.3 millions of euros. This is due to the operational limits and from a certain point on the project cost can no longer decrease.

Table 8. Results for some of the variables for $N_{law} = 15$ and $TSS_{law} = 35$

COD_{law} (g/m^3)	TSS_{ef} (g/m^3)	N_{ef} (g/m^3)	V_a m^3	A_s m^2	h (m)	G_S $(m^3/d\ STP)$	r	Q_{ef} (m^3/d)	*Total cost* 10^6 euros	LOQO iterations
45	0.40	5.9	1346	337	3.5	14521	1.8	1933	7.2	86
50	0.16	9.0	1483	332	3.7	8735	1.8	1929	5.3	72
55	1.8	10.2	1525	330	3.7	6320	1.8	1931	4.4	75
60	1.8	11.0	1498	341	3.7	4796	1.9	1931	3.8	75
65	0.93	11.7	1462	351	3.6	3629	2.0	1931	3.2	89
70	0.50	12.2	1454	351	3.6	2641	2.0	1930	2.8	92
75	0.32	12.4	1444	351	3.6	1755	2.0	1932	2.3	100
80	0.15	12.7	1431	351	3.6	933	2.0	1929	1.8	113
85	1.6	13.1	1408	351	3.6	503	2.0	1930	1.4	84
90	4.9	13.2	1373	351	3.5	503	2.0	1930	1.4	71
95	8.3	13.3	1338	351	3.5	503	2.0	1935	1.4	52
100	11.7	13.6	1303	351	3.4	503	2.0	1932	1.4	54
105	15.1	13.9	1268	351	3.4	503	2.0	1933	1.4	55
110	18.4	14.1	1233	351	3.3	503	2.0	1934	1.4	61
115	21.8	14.2	1197	351	3.3	503	2.0	1942	1.3	67
120	25.2	14.3	1162	351	3.2	503	2.0	1941	1.3	59
125	28.7	14.4	1136	347	3.2	503	2.0	1941	1.3	110
130	32.3	14.4	1117	340	3.1	503	1.9	1946	1.3	209
135	35.0	14.5	1099	334	3.1	503	1.8	1948	1.3	266
140	35.0	14.5	1099	334	3.1	503	1.8	1946	1.3	271

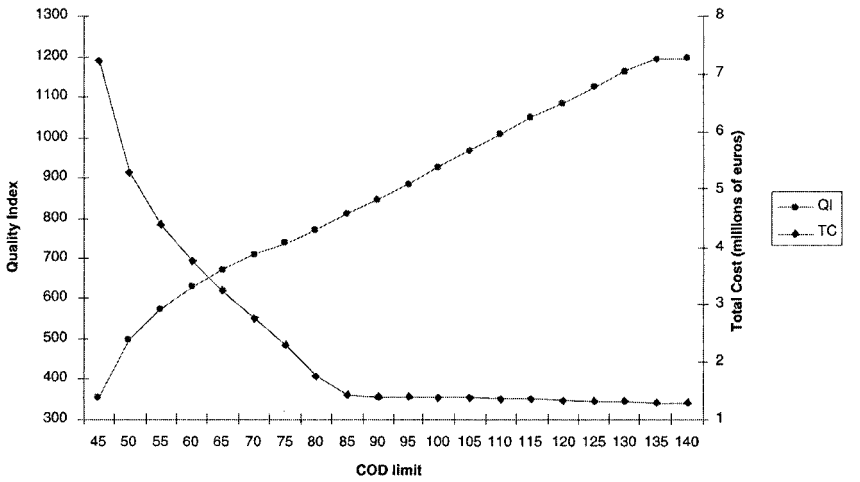


Fig. 4. Total cost (TC) and quality index (QI) *versus* COD limit at the effluent.

Regarding the quality index, it deteriorates more stressfully until a COD limit of 60. For higher values the deterioration of the effluent grows almost linearly.

Another important observation is that although the COD limit is always achieved, TSS only reaches its limit when the COD limit is 135 and over, and the N limit is never attained. Nevertheless, as the imposed COD limit increases, the TSS and N at the effluent get larger. This means that for the considered interval, the carbonaceous matter dominates the process, being the parameter that determines the cost.

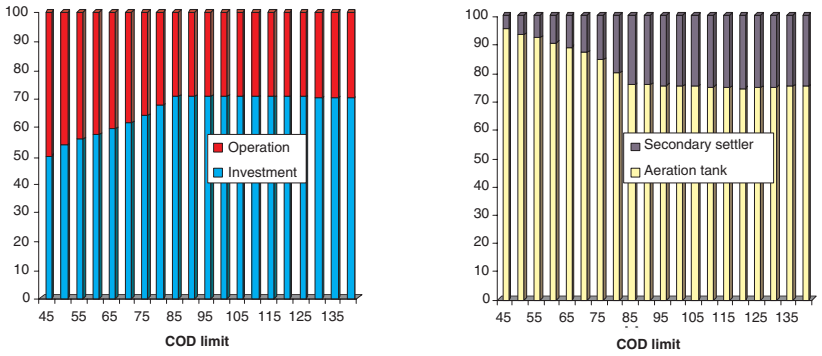


Fig. 5. Contributions of investment/operation costs and aeration tank/secondary settler costs.

For COD limit values between 90 and 120, the sedimentation area is maintained at 351 m^2 while the depth decreases because, as mentioned in Subsection 2.4, this model allows a trade-off between the sedimentation area and the settler depth. However, the five cases displayed with $A_s = 351$ and $h = 3.6$ are acceptable because the TSS concentrations are kept very small. For higher imposed COD limits, both sedimentation area and depth decrease.

The aeration tank volume and the air flow generally decrease as the imposed COD limit increases. From the dramatic reduction of the air flow verified for the COD limit interval [45,85], we may conclude that the air flow is the main contribution to the total cost.

As to the recycle rate, it reaches 2 (its maximum value) at a $COD_{law} = 65$, and this happens for two reasons. First, to get the minimum of TSS at the effluent, the sludge goes to the underflow. Second, to achieve the demanded quality it is necessary to keep as much biomass as possible inside the aeration tank.

Now, looking into the costs with more detail, from Figure 5 we can see that the investment costs overlap the operation costs, for a life span of 20 years. However, for a more demanding system this difference is smaller. For example, with a $COD_{law} = 45$, the operation costs equal the investment costs. Regarding the two involved unit processes, we may conclude that the cost associated with the aeration tank is much higher than that of the secondary settler, in particular for COD_{law} values smaller than 85. In both cases, the relations between costs are maintained almost constant for COD_{law} values of 85 and higher.

5 Conclusions

In this paper we present the mathematical formulation of a nonlinear problem that considers the optimal design and operation, in terms of minimum installation and operation cost, of an activated sludge system in WWTP's, based on portuguese real data and effluent quality law limits. The optimization process was carried out running LOQO.

From our numerical experience we may conclude that the quality of the effluent influences directly the cost of the treatment plant project, especially in terms of carbonaceous matter (COD). With the environmental law invigorating in Portugal, only the COD limit influences the cost of the design. With the compelling observance of portuguese environmental law, neither TSS nor N reach their legal limits, 35 and 15 respectively, when the $COD_{ef} = 125$.

The main contribution to the total cost is the air flow because it influences the design in the project (electromechanical material) and during the operation (consumption of energy by the aeration).

Our next task is to add an anoxic preliminary treatment for phosphorous removal in series with the aeration tank, that may be described by the model ASM2d [4], and to reoptimize the process.

Acknowledgements. The authors acknowledge the company Factor Ambiente (Braga, Portugal) for the data provided. We are also grateful to the referee comments that greatly improved the final version of the paper.

Appendix - Notation

b_A=decay coefficient for autotrophic biomass, day^{-1}
b_H=decay coefficient for heterotrophic biomass, day^{-1}
BOD=biochemical oxygen demand, $g\ O_2/m^3$
BOD_U=ultimate BOD, $g\ O_2/m^3$
COD=chemical oxygen demand, $g\ COD/m^3$
$DSVI$=Diluted sludge volume index
f_{BOD}=BOD/BOD_U ratio
f_P=fraction of biomass leading to particulate products
G_S=air flow rate, m^3/day at STP
$HenryO_2$=Henry constant
HRT=hydraulic retention time, day
i=discount rate
icv=$X : VSS$ ratio, $g\ COD/g\ VSS$
i_{X_B}=nitrogen content of active biomass, $g\ N/g\ COD$
i_{X_P}=nitrogen content of endogenous/inert biomass, $g\ N/g\ COD$
IC=investment cost, 2003 $euros$
ISS=inorganic suspended solids, g/m^3
k_a=ammonification rate, $m^3/g\ COD/day$
k_h=maximum specific hydrolysis rate, day^{-1}

K_La=overal mass transfer coefficient, day^{-1}

K_{NH}=ammonia half-saturation coefficient for autotrophic biomass growth, $g\ N/m^3$

K_{NO}=nitrate half saturation coefficient for denitrifying heterotrophic biomass, $g\ N/m^3$

K_{OA}=oxygen half-saturation coefficient for autotrophs growth,$g\ O_2/m^3$

K_{OH}=oxygen half-saturation coefficient for heterotrophs growth, $g\ O_2/m^3$

K_S=readily biodegradable substrate half-saturation coefficient for heterotrophic biomass, $g\ COD/m^3$

K_X=half-saturation coefficient for hydrolysis of slowly biodegradable substrate, $g\ COD/g\ COD$

n=life span of the treatment plant, $years$

N=total nitrogen, $g\ N/m^3$

OC=operation costs, 2003$euros$

P_{O_2}=partial pressure of oxygen uncorrected

PWWF=Peak wet weather flow

Q=flow, m^3/day

QI=quality index, Kg of $pollution/day$

r=recycle rate

S=soluble COD, $g\ COD/m^3$

S_{alk}=alkalinity, molar units

S_I=soluble inert organic matter, $g\ COD/m^3$

S_{ND}=soluble biodegradable organic nitrogen, $g\ N/m^3$

S_{NH}=free and ionized ammonia, $g\ N/m^3$

S_{NO}=nitrate and nitrite nitrogen, $g\ N/m^3$

S_O=dissolved oxygen, $g\ (-COD)/m^3$

$S_{O_{sat}}$=saturated oxygen concentration, g/m^3

S_S=readily biodegradable soluble substrate, $g\ COD/m^3$

SRT=sludge retention time, day

STP=standard temperature and pressure

TC=total costs, 2003$euros$

V_a=aeration tank volume, m^3

VSS=volatile Suspended Solids, g/m^3

T=temperature, C

TKN=total nitrogen of Kjeldahl, $g\ N/m^3$

TSS=total Suspended Solids, g/m^3

X=particulate COD, $g\ COD/m^3$

X_{BA}=active autotrophic biomass, $g\ COD/m^3$

X_{BH}=active heterotrophic biomass, $g\ COD/m^3$

X_I=particulate inert organic matter, $g\ COD/m^3$

X_{II}=inert inorganic suspended solids, g/m^3

X_{ND}=particulate biodegradable organic nitrogen, $g\ N/m^3$

X_P= particulate products arising from biomass decay, $g\ COD/m^3$

X_S=slowly biodegradable substrate, $g\ COD/m^3$

Y_A=yield for autotrophic biomass, $g\ COD/g\ N$

Y_H=yield for heterotrophic biomass, $g\ COD/g\ COD$
α=wastewater/clean water coefficient
β=salts and ions correction factor
η=standard oxygen transfer efficiency
η_g=correction factor for μ_H under anoxic conditions
η_h=correction factor for hydrolysis under anoxic conditions
μ_A=maximum specific growth rate for autotrophic biomass, day^{-1}
μ_H=maximum specific growth rate for heterotrophic biomass, day^{-1}
ρ=density of water, Kg/m^3
θ=temperature correction factor

subscripts
a=aeration tank
ef=effluent
in=entery of the aeration tank
inf=influent
p=during a PWWF event
r=recycle
s=settling tank
w=sludge waste
no index=inside the aeration tank=exit of the aeration tank

References

1. J.B. Copp (ed). The Cost Simulation Benchmark - Description and Simulator Manual. Office for Official Publications of the European Comunities, 2002.
2. G. A. Ekama, J. L. Barnard, F. W. Günthert, P. Krebs, J. A. McCrquodale, D. S. Parker, and E. J. Wahlberg. Secondary settling tanks: Theory, modelling, design and operation. Tech. Rep. 6, Int. Association on Water Quality, 1997.
3. R. Fourer, D. M. Gay, and B. Kernighan. A modeling language for mathematical programing. Management Science, 36(5):519–554, 1990.
4. M. Henze, W. Gujer, T. Mino, T. Matsuo, M. C. Wentzel, G. V. R. Marais, and M. C. M. Van Loosdrecht. Activated sludge model no. 2d, ASM2d. Water Science and Technology, 39(1):165–182, 1999.
5. M. Henze, C. P. L. Grady Jr, W. Gujer, G. V. R. Marais, and T. Matsuo. Activated sludge model no. Tech. Rep. 1, IAWPRC Task Group on Mathematical Modelling for design and operation of biological wastewater treatment, 1986.
6. Hydromantis, Inc., Canada.GPS-X V4.1, 2002. http://www.hydromantis.com /software02.html.
7. D. Tyteca. Sensivity analysis of the optimal design of a municipal wastewater treatment plant. In D. Dubois (ed.), Progress in Ecological Engineering and Management by Mathemativa Modelling, Proc. Second Int. Conf. on the State-of-the-Art in Ecological Modelling, pp 743–766, Liège, 1981. Éditions Cebedoc.
8. R. J. Vanderbei and D. F. Shanno. An interior-point algorithm for nonconvex nonlinear programming. Computational Optimization and Aplications, 13:231–252, 1999.

9. P. A. Vanrolleghem and S. Gillot. Robustness and economic measures as control benchmark performance criteria. Water Science and Technology, 45(4-5):117–126, 2002.
10. P. A. Vanrolleghem, U. Jeppsson, J. Carstensen, B. Carlsson, and G. Olsson. Integration of watewater treatment plant design and operation - a systematic approach using cost functions. Water Science and Technology, 36(3-4):159–171, 1996.